Proceedings of the
31st International Workshop on Statistical Modelling

Volume I

July 4–8, 2016
Rennes, France

Jean-François Dupuy, Julie Josse
(editors)

Proceedings of the 31st International Workshop on Statistical Modelling
volume I,
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)

Preface

Dear Participants,

We welcome you to the 31st International Workshop on Statistical Modelling (IWSM) in Rennes, France. Since the inaugural workshop held in Innsbruck (Austria) in 1986, IWSM visited several European countries and travelled to USA and Australia. But this is only the second time that IWSM visits France, after Toulouse edition (1990). We are particularly happy and honoured to host this conference in Rennes, a city which has a long tradition of research in Statistics (Rennes hosted the development of the French School of *analyse des données*). Nowadays, Rennes has become one main place for Statistics in France, with several high level training programs provided by Rennes universities and Grandes Ecoles. Statistics is also one of the priority themes of the Centre Henri Lebesgue (a research center for mathematics founded in 2012), whose aim is to promote research and graduate studies in mathematics in Western France. IWSM2016 is the closing conference of a "Statistics thematic semester" supported by the Centre Henri Lebesgue. This semester hosted height international events devoted to Statistics and its applications.

IWSM'2016 will perpetuate the tradition of a convivial event where stimulating discussions, exchange of ideas and interactions between participants are strongly encouraged through some well-established features of the workshop: no parallel session, a particular emphasis placed on interdisciplinarity and the participation of young researchers, and an attractive social program.

The high standards of the conference and the quality of all presentations are ensured by the scientific committee, who made a great work in reviewing all submitted abstracts. The scientific committee also invited renowned experts to give plenary talks and we are very glad that Francesca Chiaromonte, Stefan Lang, Jean-Michel Marin, Adrian Raftery and Sujit Sahu accepted the invitation to give a presentation. We are also glad that Søren Højsgaard accepted our invitation to give a one day course, preceding the workshop.

IWSM also constitutes a wonderful opportunity for students to exchange with senior scientists. For this reason, and following IWSM tradition, student participation has been strongly encouraged. Three students will receive awards for the best student paper, best oral presentation and best poster. Furthermore, two student travel grants have been kindly provided by the Statistical Modelling Society. Seven additional student grants were generously provided by the Centre Henri Lebesgue.

Finally, we thank all authors for their contributions to this edition of IWSM and for their careful work in preparing their manuscripts for the proceedings volumes. We wish you a pleaseant stay in Rennes, a stimulating conference and hopefully…some nice weather on Brittany!

On behalf of the local organizing committee
Jean-François Dupuy
Rennes, May 2016

Contents

Part I – Invited Papers

Part II – Contributed Papers

Part I – Invited Papers

Functional Data Analysis at the boundary of "Omics"

Marzia Cremona[1], Rebeca Campos-Sanchez[2], Alessia Pini[3], Simone Vantini[3], Kateryna Makova[1], Francesca Chiaromonte[1,4]

[1] The Pennsylvania State University, USA
[2] University of Costa Rica, Costa Rica
[3] Politecnico di Milano, Italy
[4] Sant Anna School of Advanced Studies, Italy

E-mail for correspondence: `fxc11@psu.edu`

Abstract: In this talk, we will describe two collaborative projects in which Functional Data Analysis techniques have been successfully applied to large "Omics" data sets. In the first, we considered a collection of thousands of endogenous retrovirus sequences detected in the human and mouse genomes, and quantitated a large number of genomic landscape features around their integration sites and in control regions. Using a recently proposed Interval Testing Procedure (ITP; Pini and Vantini, 2016) and Functional Logistic Regression, we were able to gain important insights on the effects of such features on the integration and fixation of endogenous retroviruses (Campos-Sanchez *et al.*, 2016). In the second project, we developed an algorithm for probabilistic k-means clustering with alignment to perform Functional Motif Discovery across a set of curves. We are using this algorithm to explore the high-resolution profiles of different mutation rates in regions of the human genome identified in Kuruppumullage *et al.* (2013), and expect it to have broad applicability to other "Omics" studies.

Keywords: Functional Data Analysis; "Omics" data; Interval Testing Procedure; Functional Motif Discovery.

1 Functional Data Analysis in "Omics" research

Functional Data Analysis has been instrumental to advances in many scientific domains. In the last years, it has increasingly been applied to "Omics" research. One example are functional linear models used to screen variants (e.g., SNPs; single nucleotide polymorphisms) genome-wide, and possibly

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).

accounting for a number of covariates, to identify effects on a complex phenotype quantitated as a response curve (e.g., Reimherr and Nicolae, 2014). This is an important extension of classical GWAS (genome-wide association studies), where the phenotypes are expressed as case/control binary variables, and QTL (quantitative trait loci) analyses, where the phenotypes are expressed as continuous variables. Another example are analyses of the shapes of peaks produced by ChiP-seq experiments, which indicate the putative binding locations of proteins interacting with the genome under certain conditions or in certain tissues/cell types. Clustering of peak shapes can be used to identify meaningful groups or types of binding sites genome-wide (e.g., Cremona *et al.*, 2015).

2 Investigating fixation and integration preferences of endogenous retroviruses with ITP and Functional Logistic Regression

Recently, we used Functional Data Analysis techniques to investigate features of the genomic landscape that may affect the integration and fixation of endogenous retroviruses (ERVs), based on their profiles around ERVs' integration sites (Campos-Sanchez *et al.*, 2016). ERVs are the remnants of retroviral infections in the germ line. They occupy a large portion of many mammalian genomes ($\sim$8% and $\sim$10% of the human and mouse genomes, respectively) and distribute unevenly along them – contributing to shape genomic structure, evolution and function. In our study, we considered a large collection of the most recently active ERVs in the human and mouse genomes, comprising 826 fixed and 1,065 in vitro HERV-Ks in human, and 1,624 fixed and 242 polymorphic ETns, as well as 3,964 fixed and 1,986 polymorphic IAPs, in mouse. We quantitated over 40 human and mouse genomic landscape features (e.g., non-B DNA structure, recombination rates, and histone modifications) at 1kb resolution in the ± 32kb flanking regions of these ERVs and in control regions, and analyzed the resulting profiles with a powerful functional hypothesis test, the Interval Testing Procedure (ITP; Pini and Vantini, 2016) – which we generalized for our study – as well as Functional Logistic Regression. These analyses allowed us to identify genomic scales and locations where various features display their influence, and to understand how they work in concert to provide signals essential for integration and fixation of ERVs.

Importantly, contrasting ERVs of different evolutionary ages (young in vitro and polymorphic ERVs, older fixed ERVs) we were able to disentangle integration vs. fixation preferences and to gain important insights on the mechanisms underlying the uneven distribution of ERVs along the genome. We found that ERVs integrate preferentially in late-replicating, AT-rich regions with abundant microsatellites, mirror repeats, and repressive histone marks. We also found that ERVs fixate preferentially in regions depleted of genes and evolutionarily conserved elements, and with low recombination rates – likely reflecting the fact that purifying selection and ectopic recombination act to remove ERVs from the genome. Interestingly, in addition to a negative effect on fixation of high recombination rates in both human and mouse genomes, we found a positive association between recombination hotspots and ERVs fixation in human, and one between hotspots and ERVs integration in mouse.

3 Exploring patterns in mutation rates profiles with Functional Motif Discovery

On a different front, motivated by an attempt to explore the high-resolution profiles of different types of mutation rates (point substitutions, small insertions and small deletions) along some special regions of the human genome identified in (Kuruppumullage *et al.*, 2013), we are developing an approach for Functional Motif Discovery – i.e. for finding shapes that recur within a given set of curves. This exercise connects Functional Data Analysis with the notion of motif discovery, typically on sequences of categorical symbols, which is ubiquitous in bioinformatics. Unlike other approaches proposed in the literature (e.g., Chiu *et al.*, 2003; Castro and Azevedo, 2010), which discretize the domains of continuous signals and then exploit traditional motif discovery algorithms, ours represents the data as curves – allowing us to leverage the full statistical arsenal of Functional Data Analysis, from smoothing, to exploiting the information in derivatives, to providing rigorous significance assessments. Moreover, our approach does not require the length of the unknown motifs to be fixed, can handle gaps in the data, and is applicable to multidimensional curves.

At the heart of our Functional Motif Discovery is an algorithm that, given a minimum length c and a target number of motifs k, performs probabilistic k-mean clustering with alignment to identify recurrent shapes across the input curves. The algorithm works in a way very similar to a local alignment algorithm in bioinformatics (e.g., Altschul *et al.*, 1990; Kent, 2002); it starts by locating very high similarity “seeds” (short, almost identical portions of the curves) and attempts to extend these seeds on either side until a running similarity score gets too low and extension is terminated. Notably, this local curve alignment differs from the global alignment of a set of curves that is typically used for registration in Functional Data Analysis. Iterating the algorithm on multiple initializations of the k-means and different choices of c and k generates a set of candidate recurrent shapes that is then post-processed to screen out weaker motifs and locate instances of the stronger ones along the curves (the latter is akin to a motif search, as opposed to motif discovery). In addition to the motivating study of high-resolution mutation rates profiles, we expect our approach to have a very broad range of applications in “Omics” research.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403-10.

Campos-Sanchez, R., Cremona, M.A., Pini, A., Chiaromonte, F., and Makova K.D. (2016). Integration and fixation preferences of human and mouse endogenous retroviruses uncovered with Functional Data Analysis. To appear in *PLOS Computational Biology.*

Castro, N., and Azevedo, P. J. (2010). Multiresolution motif discovery in time series. *Proceeding of the 10th SIAM International Conference on Data Mining*, 665-676.

Chiu, B., Keogh, E., and Lonardi S. (2003). Probabilistic discovery of time series motifs. *Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining*, 493-498.

Cremona, M.A., Sangalli, L.M., Vantini, S., Dellino, G.I., Pelicci, P.G., Secchi, P., and Riva, L. (2015). Peak shape clustering reveals biological insights. *BMC Bioinformatics*, **16**(1), 349.

Kent, W.J. (2002). BLAT – the BLAST-like alignment tool. *Genome Research*, **12**(4), 656-64.

Kuruppumullage Don, P., Ananda, G., Chiaromonte, F., and Makova, K.D. (2013). Segmenting the human genome based on states of neutral genetic divergence. *Proceedings of the National Academy of Sciences*, **110**(36), 14699-14704.

Pini, A., and Vantini, S. (2016). The Interval Testing Procedure: a general framework for inference in Functional Data Analysis. To appear in *Biometrics.*

Reimherr, M., and Nicolae, D. (2014). A functional data analysis approach for genetic association studies. *The Annals of Applied Statistics*, **8**(1), 406-429.

Bayesian distributional structured additive regression

Nadja Klein[1], Thomas Kneib[1], Stefan Lang[2], Alexander Razen[2]

[1] Department of Statistics, University of Göttingen, Germany
[2] Department of Statistics, University of Innsbruck, Austria

E-mail for correspondence: `stefan.lang@uibk.ac.at`

Abstract: In this talk, we discuss a generic Bayesian framework for inference in multilevel distributional regression models in which each parameter of a potentially complex response distribution and not only the mean is related to a multilevel structured additive predictor. The latter is composed additively of a variety of different functional effect types such as nonlinear effects, spatial effects, random coefficients, interaction surfaces or other (possibly non-standard) basis function representations. Particular emphasis is given on a specific form of multiplicative random effects that scale a particular nonlinear curve while their overall shape is preserved.
Inference is based on efficient Markov chain Monte Carlo simulation techniques where a generic procedure makes use of distribution-specific iteratively weighted least squares approximations to the full conditionals.
The importance and flexibility of Bayesian structured additive distributional regression to estimate all parameters as functions of explanatory variables and therefore to obtain more realistic models, is exemplified in a detailed case study on modelling house prices in Austria and Germany.

Keywords: GAMLSS; hedonic regression models, iteratively weighted least squares proposals, MCMC; multilevel models

1 Introduction

Classical regression models within the exponential family framework, such as generalised linear models or generalised additive models (GAMs, e.g. Fahrmeir et al., 2013), focus exclusively on relating the mean of a response variable to covariates but neglect the potential dependence of higher order moments or other features of the response distribution on covariates.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016.

As a consequence, the advantage of obtaining covariate effects that are straightforward to estimate and easy to interpret is at least partly offset by the likely misspecification of the model that may render inferential conclusions invalid. A completely distribution-free alternative to mean regression is provided by quantile or expectile regression where the assumptions on the error term are generalised such that the regression predictor is related to a local feature of the response distribution, indexed by a pre-specified asymmetry parameter (the quantile or expectile level), see Koenker (2005) and Fahrmeir et al. (2013) for recent overviews. Both approaches have the distinct advantage that basically no assumptions on the specific type of the response distribution or homogeneity of certain parameters such as the variance are required. However, this flexibility also comes at a price since properties of the determined estimates are more difficult to obtain, the flexibility of the predictor specification is somewhat limited and estimates for a set of asymmetries may cross leading to incoherent distributions for the response. Moreover, model choice and model comparison tend to be difficult since the models only relate to local properties of the response. Finally, if prior knowledge on specific aspects of the response distribution is available, quantile and expectile regression may be less efficient and are also less appropriate for discrete distributions or mixed discrete continuous distributions.

As a consequence, it is of considerable interest to derive models that are in between the simplistic framework of exponential family mean regression and distribution-free approaches. Such an approach is given by the class of generalised additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005) in which all parameters of a potentially complex response distribution are related to additive regression predictors in the spirit of GAMs. In this talk, we build upon GAMLSS to develop a generic Bayesian treatment of distributional regression relying on Markov chain Monte Carlo simulation algorithms. To construct suitable proposal densities, we follow the idea of iteratively weighted least squares proposals (Gamermann, 1997, Brezger and Lang, 2006) and construct local quadratic approximations to the full conditionals.

The full potential of distributional regression is only exploited when the regression predictor is also broadened beyond the scope of simple linear or additive specifications. We will consider structured additive predictors (Fahrmeir et al., 2013, Brezger and Lang, 2006) where each predictor is determined as an additive combination of various types of functional effects, such as nonlinear effects of continuous covariates, seasonal effects of time trends, spatial effects, random intercepts and slopes, varying coefficient terms or interaction surfaces. All of these approaches can be represented in terms of possibly non-standard basis functions in combination with a multivariate Gaussian prior to enforce desired properties of the estimates, such as smoothness or sparsity.

In a second step, we generalise the approach to a hierarchical or multilevel version of regression models with structured additive predictors where the regression coefficients of a particular nonlinear term may obey another regression model with structured additive predictor. In that sense, the model is then composed of a hierarchy of complex structured additive regression models. The proposed model may be regarded as an extended version of a multilevel model with nonlinear covariate terms in every level of the hierarchy.

In many applications, the data consists of a number of clusters that can be regarded as submarkets of one larger market. In general, it is unlikely that a covariate's effect on a parameter of the response distribution – be it its mean or another parameter – is homogeneous across these submarkets. In real estate data, for example, a frequently observed phenomenon is that the price effects of covariates vary from one spatial unit to another. However, completely different functional forms in each cluster are not common. In order to deal with this challenge, we suggest the use of cluster specific random scaling factors. In doing so, one still assumes homogeneity for the functional form of the response function but allows for heterogeneity with respect to its scaling. Lang et al. (2015) and Weber et al. (2016) successfully have applied this approach to store sales models, thereby considerably improving the predictive validity of the models. Razen and Lang (2016b) provide a more systematic approach for random scaling factors in the context of distributional regression.

2 Structured Additive Distributional Regression

We assume that observations on a scalar response variable $y_1, \dots, y_n$ as well as covariate information $\boldsymbol{\nu}_i$, $i = 1, \dots, n$, have been collected for n individuals. The conditional distribution of observation y_i given the covariate information $\boldsymbol{\nu}_i$ is assumed to be from a pre-specified class of K-parametric distributions $f_i(y_i|\vartheta_{i1}, \dots, \vartheta_{iK})$ indexed by the (in general covariate-dependent) parameters $\vartheta_{i1}, \dots, \vartheta_{iK}$. Note that f_i is considered a general density, i.e. we use the same notation for continuous responses, discrete responses and also mixed discrete-continuous responses. Each parameter ϑ_{ik} is linked to a semiparametric regression predictor η_{ik} formed of the covariates via a suitable (one-to-one) response function such that $\vartheta_{ik} = h_k(\eta_{ik})$ and $\eta_{ik} = h_k^{-1}(\vartheta_{ik})$. The response function is usually chosen to ensure appropriate restrictions on the parameter space such as the exponential function $\vartheta_{ik} = \exp(\eta_{ik})$, to ensure positivity, the logit link $\vartheta_{ik} = \exp(\eta_{ik})/(1 + \exp(\eta_{ik}))$ for parameters representing probabilities or the identity function if the parameter space is unrestricted.

In most applications, linear regression models are too restrictive to capture the underlying true, complex structure of real life problems. We therefore consider structured additive distributional regression models, a generic framework in which each of the K model parameters $\boldsymbol{\vartheta}_k = (\vartheta_{1k}, \dots, \vartheta_{nk})'$,

$k = 1, \ldots, K$ is related to a semiparametric predictor with the general form

$$\eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\boldsymbol{\nu}_i) + \ldots + f_{J_k}^{\vartheta_k}(\boldsymbol{\nu}_i)$$

where β_0 represents the overall level of the predictor and the functions $f_j(\boldsymbol{\nu}_i)$, $j = 1, \ldots, J_k$, relate to different covariate effects required in the applications. Note that of course each parameter vector $\boldsymbol{\vartheta}_k$ may depend on different covariates and especially a different number of effects J_k. To simplify notation, we suppress this possibility and also drop the parameter index in the following.

In structured additive regression, each function f_j is approximated by a linear combination of D_j appropriate basis functions, i.e.

$$f_j(\boldsymbol{\nu}_i) = \sum_{d_j=1}^{D_j} \beta_{j,d_j} B_{j,d_j}(\boldsymbol{\nu}_i)$$

such that in matrix notation we can write $\boldsymbol{f}_j = (f_j(\boldsymbol{\nu}_1), \ldots, f_j(\boldsymbol{\nu}_n))' = \mathbf{Z_j}\boldsymbol{\beta}_\mathbf{j}$ where $\mathbf{Z_j[i, d_j] = B_{j,d_j}}(\boldsymbol{\nu}_\mathbf{i})$ is a design matrix and $\boldsymbol{\beta}_j$ is the vector of coefficients to be estimated. Specific examples can be found in Fahrmeir et al. (2013).

For regularisation reasons it is common to add a penalty term $pen(\boldsymbol{f}_j) = pen(\boldsymbol{\beta}_j) = \boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j$ that controls specific smoothness or sparseness properties. The Bayesian equivalent to this frequentist formulation is to put multivariate Gaussian priors

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \left(\frac{1}{\tau_j^2}\right)^{\frac{rank(\boldsymbol{K}_j)}{2}} \exp\left(-\frac{1}{2\tau_j^2}\,\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right) \qquad (1)$$

on the regression coefficients $\boldsymbol{\beta}_j$ with prior precision matrix $\boldsymbol{K}_j$ which corresponds to the penalty matrix in a frequentist formulation. The hyperparameters τ_j^2 are assigned inverse gamma hyperpriors $\tau_j^2 \sim IG(a_j, b_j)$ (with $a_j = b_j = 0.001$ as a default option) in order to obtain a data-driven amount of smoothness.

3 Multilevel data

As outlined in the introduction, in many applications the data is clustered. Real estate data, for example, typically is clustered in spatial units (e.g. districts, counties, etc.).

We therefore propose a hierarchical or multilevel version of distributional STAR models, see Lang et al. (2014). That is the regression coefficients $\boldsymbol{\beta}_j$ of a term f_j may themselves obey a regression model with structured additive predictor

$$\boldsymbol{\beta}_j = \boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j = \boldsymbol{Z}_{j1}\boldsymbol{\beta}_{j1} + \ldots + \boldsymbol{Z}_{jq_j}\boldsymbol{\beta}_{jq_j} + \boldsymbol{\varepsilon}_j, \qquad (2)$$

where the terms $\boldsymbol{Z}_{j1}\boldsymbol{\beta}_{j1}, \dots, \boldsymbol{Z}_{jq_j}\boldsymbol{\beta}_{jq_j}$ correspond to additional nonlinear functions $f_{j1}, \dots, f_{jq_j}$ and $\boldsymbol{\varepsilon}_j \sim N(\boldsymbol{0}, \tau_j^2 \boldsymbol{I})$ is a vector of i.i.d. Gaussian random effects. Here, we restrict ourselves to i.i.d. Gaussian random effects although more sophisticated structures like the Bayesian LASSO, Dirichlet process mixtures or spike and slab priors can be implemented in a straightforward way. Moreover, a third level or even higher levels in the hierarchy are possible by assuming that the second level regression parameters $\boldsymbol{\beta}_{jl}$, $l = 1, \dots, q_j$, obey again a STAR model. In that sense, the model is composed of a hierarchy of complex structured additive regression models.
The typical application of the proposed models are multilevel data where a hierarchy of units or clusters grouped at different levels is given. Here, we will analyze real estate data to model house prizes in Austria.
The hierarchical structure of the Austrian political-administrative units suggests the use of the following four level predictor for each parameter being modelled, see also Razen et al. (2016a):

$$
\begin{array}{lll}
\text{level-1:}\ \boldsymbol{\eta} & = & \boldsymbol{f}_1(\mathit{area}) + \boldsymbol{f}_2(\mathit{area\_plot}) + \boldsymbol{f}_3(\mathit{age}) + \boldsymbol{f}_4(\mathit{time\_ind}) + \\
 & & \boldsymbol{f}_5(\mathit{municipality}) + \boldsymbol{X}\boldsymbol{\gamma} \\
 & = & \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_2\boldsymbol{\beta}_2 + \boldsymbol{Z}_3\boldsymbol{\beta}_3 + \boldsymbol{Z}_4\boldsymbol{\beta}_4 + \boldsymbol{Z}_5\boldsymbol{\beta}_5 + \boldsymbol{X}\boldsymbol{\gamma} \\
\text{level-2:}\ \boldsymbol{\beta}_5 & = & \boldsymbol{f}_{5,1}(\mathit{pp\_ind}) + \boldsymbol{f}_{5,2}(\mathit{ln\_educ}) + \boldsymbol{f}_{5,3}(\mathit{age\_ind}) + \\
 & & \boldsymbol{f}_{5,4}(\mathit{comm}) + \boldsymbol{f}_{5,5}(\mathit{ln\_dens}) + \boldsymbol{f}_{5,6}(\mathit{district}) + \boldsymbol{\varepsilon}_5 \\
 & = & \boldsymbol{Z}_{5,1}\boldsymbol{\beta}_{5,1} + \boldsymbol{Z}_{5,2}\boldsymbol{\beta}_{5,2} + \boldsymbol{Z}_{5,3}\boldsymbol{\beta}_{5,3} + \boldsymbol{Z}_{5,4}\boldsymbol{\beta}_{5,4} + \\
 & & \boldsymbol{Z}_{5,5}\boldsymbol{\beta}_{5,5} + \boldsymbol{Z}_{5,6}\boldsymbol{\beta}_{5,6} + \boldsymbol{\varepsilon}_5 \\
\text{level-3:}\ \boldsymbol{\beta}_{5,6} & = & \boldsymbol{f}_{5,6,1}(\mathit{wko\_ind}) + \boldsymbol{f}^{mrf}_{5,6,2}(\mathit{district}) + \boldsymbol{f}_{5,6,3}(\mathit{county}) + \boldsymbol{\varepsilon}_{5,6} \\
 & = & \boldsymbol{Z}_{5,6,1}\boldsymbol{\beta}_{5,6,1} + \boldsymbol{Z}_{5,6,2}\boldsymbol{\beta}_{5,6,2} + \boldsymbol{Z}_{5,6,3}\boldsymbol{\beta}_{5,6,3} + \boldsymbol{\varepsilon}_{5,6} \\
\text{level-4:}\ \boldsymbol{\beta}_{5,6,3} & = & \boldsymbol{1}\gamma_0 + \boldsymbol{\varepsilon}_{5,6,3}.
\end{array}
\tag{3}
$$

The categorical covariates on level-1, describing the quality and equipment of the house, are encoded as dummy variables and are subsumed in the design matrix $\boldsymbol{X}$ with estimated parameters $\boldsymbol{\gamma}$. The possibly nonlinear functions $\boldsymbol{f}_1, \boldsymbol{f}_2, \dots$ are modeled by Bayesian P-splines (Eilers and Marx, 1996, and Lang and Brezger 2004).
The level-1 equation contains an uncorrelated random municipal effect $\boldsymbol{f}_5(\mathit{municipality})$, controlling for unordered spatial heterogeneity. This municipal specific heterogeneity is modeled through the level-2 equation and is further decomposed into a district and finally into a county level effect (levels 3 and 4). Furthermore, district specific spatial heterogeneity is modeled through a correlated spatial effect $\boldsymbol{f}^{mrf}_{5,6,2}(\mathit{district})$ in the level-3 equation by Markov random fields, denoted by the superscript "*mrf*", see Fahrmeir et al. (2013) for details regarding MRFs.
Details regarding Bayesian inference regarding multilevel STAR models are given in the talk, see also Lang et al. (2014), Klein et al. (2014) and Klein et al. (2015).

4 Multiplicative random effects

Usually, there is no economic reason to assume homogeneous covariate effects across spatial units in real estate data. In contrast, different consumer price sensitivities originating from varying levels of income, diverse value of land or different ways of construction suggest spatial heterogeneity in price response. Indeed, it is reasonable to assume the effects to have the same functional form but to vary with respect to the scaling of the function. Thus, in order to account for this kind of heterogeneity, we allow for cluster specific random scaling factors for some or all of the nonlinear functions f_j. This leads to predictors of the form

$$\eta_i = (1 + \alpha_{1c_i})\, f_1(\boldsymbol{\nu}_i) + \ldots + (1 + \alpha_{Jc_i})\, f_J(\boldsymbol{\nu}_i), \tag{4}$$

$i = 1, \ldots n$, where $c_i \in \{1, \ldots, C\}$ is the cluster index of the respective observation and the α_{jc_i}, $j = 1, \ldots, J$, are normally distributed random effects with mean 0 and variance ψ_j, i.e.

$$\alpha_{jc_i} \sim \mathcal{N}\left(0, \psi_j\right), \quad c_i = 1, \ldots, C.$$

A positive random effect $\alpha_{jc} > 0$ leads to a scaling up of the function f_j indicating an increased price sensitivity while a negative random effect $\alpha_{jc} < 0$ refers to weaker price sensitivity.
A priori, the parameters are not identifiable since there is an arbitrary multiplicative constant for the functions f_j. Therefore we assume

$$\sum_{c=1}^{C} \alpha_c = 0.$$

Details regarding Bayesian inference are given in Weber et al. (2016) and Razen et al. (2016b).

References

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, **50**, 967–991.

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

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013). *Regression. Models, Methods and Applications.* Springer Verlag.

Gamerman, D. (1997). Efficient Sampling from the Posterior Distribution in Generalized Linear Models. *Statistics and Computing*, **7**, 57–68.

Klein, N. and Kneib, T. and Lang, S. (2014). Bayesian Generalized Additive Models for Location, Scale and Shape for Zero-Inflated and Overdispersed Count Data. *Journal of the American Statistical Association*, **10**, 405–419.

Klein, N. and Kneib, T. and Lang, S. and Sohn, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *The Annals of Applied Statistics*, **9**, 1024–1052.

Koenker, R. (2005). *Quantile Regression.* Cambrigde University Press.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K. and Kneib, T. (2014). Multilevel Structured Additive Regression. *Statistics and Computing*, **24**, 223–238.

Lang, S., Steiner, W., Weber, A. and Wechselberger, P. (2015): Accommodating Heterogeneity and Functional Flexibility in Store Sales Models: A Bayesian Semiparametric Approach. *European Journal of Operational Research*, **246**, 232–241.

Razen, A., Brunauer, W., Klein, N., Kneib, T., Lang, S. and Umlauf, N. (2016a): Statistical Risk Analysis for Real Estate Collateral Valuation using Bayesian Distributional and Quantile Regression. *Working Papers in Economics and Statistics*, University of Innsbruck, 2014-12.

Razen, A. and Lang, S. (2016b): Random scaling factors in Bayesian distributional regression models with an application to real estate data. Working paper, University of Innsbruck.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society C*, **54**, 507–554.

Weber, A., Steiner, W. and Lang, S. (2016): A Comparison of Semiparametric and Heterogeneous Store Sales Models for Optimal Category Pricing. Revised for *OR Spectrum.*

Approximate Bayesian Computation using Random Forests

Jean-Michel Marin[1]

[1] Institut Montpelliérain Alexander Grothendieck

E-mail for correspondence: jean-michel.marin@umontpellier.fr

Abstract: Approximate Bayesian Computation (ABC) has grown into a standard methodology to handle Bayesian inference in models associated with intractable likelihood functions. In a first part, we will show how our ABC Random Forests (RF) methodology can be used to select a model in a Bayesian context. We modify the way Bayesian model selection is both understood and operated, in that we rephrase the inferential goal as a classification problem, first predicting the model that best fits the data with RF and postponing the approximation of the posterior probability of the selected model for a second stage also relying on RF. Compared with earlier implementations of ABC model choice, the ABC RF approach offers several potential improvements:

(i) it often has a larger discriminative power among the competing models,

(ii) it is more robust against the number and choice of statistics summarizing the data,

(iii) the computing effort is drastically reduced (with a gain in computation efficiency of at least 50) and

(iv) it includes an approximation of the posterior probability of the selected model.

In a second part, we will consider parameter estimation questions. We advocate the derivation of a random forest for each component of the parameter vector, a tool from which an approximation to the marginal posterior distribution can be derived. Correlations between parameter components are handled by separate random forests. We will show that this technology offers significant gains in terms of robustness to the choice of the summary statistics and of computing time, when compared with the standard ABC solutions.
In the last part, we will cover some population genetics applications.

Keywords: Approximate Bayesian Computation; Random Forests; Population genetics

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).

Bayesian Population Projections with Migration Uncertainty

Adrian Raftery[1]

[1] University of Washington, Seattle, WA, USA

E-mail for correspondence: raftery@u.washington.edu

Abstract: The United Nations recently issued official probabilistic population projections for all countries for the first time, using a Bayesian hierarchical modeling framework developed by our group at the University of Washington. These take account of uncertainty about future fertility and mortality, but not international migration. We propose a Bayesian hierarchical autoregressive model for obtaining joint probabilistic projections of migration rates for all countries, broken down by age and sex. Joint trajectories for all countries are constrained to satisfy the requirement of zero global net migration. We evaluate our model using out-of-sample validation and compare point projections to the projected migration rates from a persistence model similar to the UN's current method for projecting migration, and also to a state of the art gravity model. We also resolve an apparently paradoxical discrepancy between growth trends in the proportion of the world population migrating and the average absolute migration rate across countries. This is joint work with Jonathan Azose and Hana Ševčíková. It is based on the following articles, both of which are Open Access:

- http://link.springer.com/article/10.1007/s13524-015-0415-0
- http://www.pnas.org/content/early/2016/05/18/1606119113.abstract

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016.

Efficient parameterisations of Gaussian process based models for Bayesian computation using MCMC

Sujit K. Sahu[1]

[1] University of Southampton, UK

E-mail for correspondence: S.K.Sahu@soton.ac.uk

Abstract: MCMC algorithms for Bayesian computation for Gaussian process based models under default parameterisations are slow to converge due to the presence of spatial and other induced dependence structures. The main focus of this paper is to study the effect of the assumed spatial correlation structure on the convergence properties of the Gibbs sampler under the default non-centered parameterisation (NCP) and a rival centered parameterisation (CP), for the mean structure of a general multi-process Gaussian spatial model. Our investigation finds answers to many pertinent, but as yet unanswered, questions on the choice between the two. Assuming the covariance parameters to be known, we compare the exact rates of convergence of the two by varying: the strength of the spatial correlation, the level of covariance tapering, the scale of the spatially varying covariates, the number of data points, the number and the structure of block updating of the spatial effects and the amount of smoothness assumed in a Matérn covariance function. We also study the effect of introducing differing levels of geometric anisotropy in the spatial model. The case of unknown variance parameters is investigated by using well-known MCMC convergence diagnostics. A simulation study and a real data example on modelling air pollution levels in London are used for illustrations. A generic pattern emerges that the CP is preferable in the presence of more spatial correlation or more information obtained through, for example, additional data points or by increased covariate variability.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016.

Part II – Contributed Papers

A ratio regression approach to estimate the size of the Salmonella infected flock data using validation information

Carla Azevedo[1], Dankmar Böhning[1], Mark Arnold[2]

[1] University of Southampton, United Kingdom
[2] Animal and Plant Health Agency, United Kingdom

E-mail for correspondence: `cfa1e14@soton.ac.uk`

Abstract: Capture-Recapture methods are used to estimate the size of a target population of interest when it cannot be completely observed. In capture-recapture studies just the positive counts of repeated identifications are observed and we might be able to predict the number of unobserved identifications. Sometimes, additional information on the unobserved units is available through a validation sample. In this paper, we will use the ratio plot to explore the pattern of the count distribution and a ratio regression approach allowing for the heterogeneity naturally present in the data. The guiding principle of the ratio regression approach is considering ratios of neighbouring count probabilities which can be estimated by ratios of the observed frequencies (Böhning *et al.*, 2016). After fitting an appropriated regression model the hidden zero-identifications are derived projecting the model backwards. Simulation studies were conducted to evaluate the performance of the suggested approach.

Keywords: capture-recapture; zero-truncated model; ratio plot; ratio regression.

1 Introduction and Background

The purpose of this framework is to determine the size N of an elusive target population. Let us assume that the members of the population are identified at m observational occasions where m is considered fixed in this work. For each member i the count of identifications X_i returns a count in$0, 1, ..., m$ and i takes values from 1 to N. It is assumed that X_i is available if unit i has been identified for at least one occasion. We have then that X_i is observed and let $X_1, ..., X_n$ denote the observed counts with n representing the total number of recorded individuals. We assume w.l.o.g. that $X_{n+1} =$

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).

... $= X_N = 0$. Let f_x be the frequency of units with count $X = x$. The population can be described by a probability density function $p_x(\theta)$ which denotes the probability of exactly x identifications for a generic unit where $p_x \geq 0$ and $\sum_{x=0}^{\infty} p_x = 1$. Situations of heterogeneity in the population can be detected by means of the ratio plot which works like a diagnostic device for the presence of a particular distribution (Böhning *et al.*, 2016). We can then extend this theory to a regression approach which will consider the ratios of the observed frequencies and fit a proper model to the data. Finally, we use the model to derive an estimate for f_0. It is possible to incorporate the information coming from the validation sample into the modelling and decrease the bias in the estimation process.

2 Case study - Salmonella data

The following data was provided by the Animal and Plant Health Agency, UK. Human salmonellosis is a major public health concern in Europe and the most common source of infection is thought to be through the consumption of contaminated eggs (Arnold, 2014).To assess the current prevalence of infected commercial egg-laying flocks, a European Union wide baseline survey of *Salmonella* infection was carried out between October 2004 and September 2005. The results of that survey were used as a basis for setting flock prevalence reduction targets for *Salmonella* national control programmes in each member state of the EU. As part of the baseline survey in the UK, a randomized sample of 454 commercial layer flock holdings was tested for *Salmonella*. It is important to achieve effective control on the infection at farm level and monitoring *Salmonella* strains. Hence, it is crucial that infected flocks are detected so that measures can be taken to avoid consumption of *Salmonella* contaminated eggs by the public (Arnold, 2014). In order to be able to monitor the progress of control measures for *Salmonella*, it is important to be able to obtain an accurate estimate of the initial prevalence at the time of the EU baseline survey. Therefore, it is important to adjust the under-count of disease occurrence appropriately. The main goal of the present study is to determine the number of undetected cases, i.e. the number of farms which had *Salmonella* infected poultry but for which result in the survey was negative. 53 holdings tested positive for *Salmonella* in one or more samples of the survey using a EU baseline survey method which consists of a total of 7 tests, so each farm could have 0,1,...,7 positives as table 1 shows.

TABLE 1. Positive sample of salmonella data.

x	0	1	2	3	4	5	6	7
f_x	?	17	9	5	6	5	5	6

The same method were conducted in 21 of the farms which established the validation sample as shown in table 2.

TABLE 2. Validation sample of salmonella data.

x	0	1	2	3	4	5	6	7
g_y	3	1	3	2	3	3	4	2

The validation sample also allows to observe infected holdings with all repeated tests negative.

3 Ratio regression

We are interested in estimating the size N of an elusive target population. Note that f_0 is the frequency of units that were not captured any time causing a reduction on the observable available sample. To model $p_x = p_x(\theta)$ we need to find an estimate $\hat{\theta}$ for θ so that $\hat{p}_x = p_x(\hat{\theta})$. Since we are dealing with a fixed number of sampling occasions $m = 7$, we can think about a binomial distribution to model the data and estimate θ using the EM algorithm. However, we are working with simple homogeneous models and, in fact, have not considered the benefits of having a validation sample. Consequently, we proceed with methodology which allows heterogeneity. Let us now consider the ratios:

$$\frac{p_{x+1}}{p_x} = \frac{\binom{m}{x+1}\theta^{x+1}(1-\theta)^{m-x-1}}{\binom{m}{x}\theta^{x}(1-\theta)^{m-x}} = \frac{m-x}{x+1}\,\frac{\theta}{1-\theta} \tag{1}$$

using the binomial distribution. If we reparametrize these ratios we achieve:

$$R_x = \underbrace{\frac{x+1}{m-x}}_{a_x}\frac{p_{x+1}}{p_x} = a_x\frac{p_{x+1}}{p_x} = \frac{\theta}{1-\theta} \tag{2}$$

where we used the coefficients $a_x = \frac{x+1}{m-x}$. For the binomial distribution, the ratio R_x is constant with x and we can estimate R_x by $r_x = a_x\frac{f_{x+1}/N}{f_x/N} = a_x\frac{f_{x+1}}{f_x}$ where f_x is the observed frequency of count x. We proceed with the ratio plot ($x \mapsto r_x$) which works as a diagnostic device for this distribution. In fact, we can expect the graph to show an horizontal line pattern in the case of a binomial (Böhning *et al.*, 2016).

However, there is no evidence of a horizontal line in figure 1 (left panel), independent of whether we consider the validation or the positive sample.

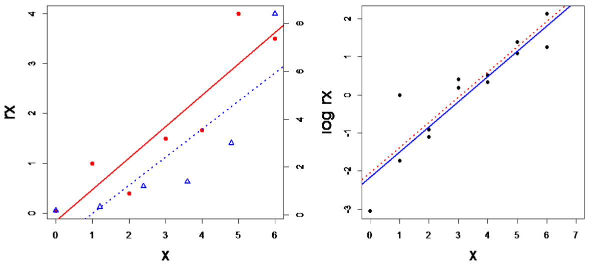


FIGURE 1. Left panel: ratio plot for the validation sample (solid points) and for the positive sample (empty triangles) with respective regression lines, continuous for the validation and dashed for the positive sample; Right panel: regression lines of log ratio on x, continuous for positive sample, dashed for validation sample, the estimated regression model is $-2.04 + 0.66x - 0.12S$.

Instead, we observe a positive line pattern. The ratio plot suggests a regression model taking advantage of the straight line pattern to determine an estimate of f_0 (Böhning *et al.*, 2016). Namely, as $log(r_x) = \alpha + \beta x + \epsilon_x$, an estimate of f_0 can be found using $log\left(a_0 \frac{f_1}{f_0}\right) = \hat{\alpha} + \hat{\beta} \times 0$, or, $\hat{f}_0 = a_0 f_1 \exp(-\hat{\alpha})$.

4 Ratio regression using validation information

This approach can be extended to incorporate the information coming from the validation sample into the ratio regression model. Considering our data, it can be done as follows:

$$log(r_x) = \alpha + \beta x + \delta S + \epsilon_x \tag{3}$$

where S represents a dummy variable taking the value of 1 if x is from the positive sample and 0 otherwise. With this approach we allow a regression line for the two samples having the same slope but different intercepts as figure 1 (right panel) shows. The resulting estimate $\hat{f}_0 = f_1 \exp(-\hat{\alpha} - \hat{\delta})$ is 25 undetected farms. Here f_1 is the frequency of ones from the positive sample. Note that if $\delta = 0$ both lines become identical and we allow for a single straight line regression model. The use of a validation sample increases the efficiency of our estimation as well as it guarantees that our model provides a reasonable final estimative (Böhning, 2016). We can also consider a model with interaction between the variable S and count x. However, in the case of interaction, the model becomes identical to fitting two separate lines and the benefit of the validation sample diminishes. A zero-inflated model was also considered as it appears we have a large quantity of zeros in addition to those predicted by the non-inflated models. We conducted simulations based on these models and the results show evidence that using the validation sample not only decreases the bias in our estimation, but also leads to more accuracy in the estimation of the population size.

5 Application to the case study

The three models (single line, parallel lines, separate lines) were applied to the salmonella data and the results are presented in table 3. Note that $n = 53$ for the positive sample and the coefficients a_x obey the binomial distribution in our analysis.

TABLE 3. Estimates of the population size N; RR denotes the ratio regression approach, PI denotes the prediction interval for the estimate and S the variable indicating type of sample ($S = 1$: positive sample). The model equation for the ratio regression model using just the positive sample is $-2.47 + 0.75x$; for model 1 (single line) we have $-2.30 + 0.70x$; for model 2 (parallel lines) is $-2.21 + 0.70x - 0.12S$ and for model 3 (separate lines) is $-1.85 + 0.60x - 0.63S + 0.15(S \times x)$; column 6 refers to the p-value of the last coefficient of the respective model.

Application	$\hat{f}_0$	**PI for** f_0	$\hat{N}$	**PI for** N	**p-value**
RR Positive	29	(1.01,56.63)	82	(54.02,109.64)	0.000
Model 1	24	(3.65,44.90)	77	(56.65,97.90)	0.000
Model 2	25	(1.49,48.35)	78	(54.49,101.35)	0.660
Model 3	29	(5.98,51.68)	82	(58.98,104.68)	0.316

We obtained 29 undetected farms using just the positive sample. Model 3 provides exactly the same results as we expect. The interaction term is not significant in model 3. The simple regression model (model 1) and the parallel lines model (model 2) produce a very similar result. Model 1 indicates 24 undetected farms while model 2 suggests 25 undetected farms. Table 3 includes the estimates for the coefficients of each model as well as prediction intervals for each estimate. As model 2 has a non-significant term for S, we conclude that model 1 is most suitable in our case and the estimate for f_0 is 24 with the shortest prediction interval.

6 The inflated model

The previous modelling does not allow for any zero-inflation. Zero-inflation would lead to a first ratio potentially much lower than the others. To account for zero-inflation, at least in an approximate way, we suggest the model $log(R_x) = \alpha + \beta x + \delta S + \lambda x^2$ estimated as $log(r_x) = -2.47 + 0.94x - 0.13S - 0.04x^2$. This model will allow a bend in the upper straight line corresponding to the positive sample and at the same time taking advantage of the validation sample. A total of 33 undetected farms were obtained employing this model as table 4 shows. In other words, a population size of 86 farms. Here, the question arises, if this kind of approach performs well on our data. As it turns out, the quadratic term is not significant. In fact, the best model here is the single line model (details not reported here). We conducted simulations that show that the estimation of N using the inflated model, with the validation sample incorporated, produces substantially better results in terms of precision along with an enormous reduction in the bias.

TABLE 4. Estimate of the population size N for the zero-inflated model according to the model equation $log(r_x) = -2.47+0.94x-0.13S-0.04x^2$; column 6 refers to the p-value of the last coefficient of the model; PI denotes the prediction interval for the estimate and S denotes the variable S.

Application	$\hat{f}_0$	**PI for** f_0	$\hat{N}$	**PI for** N	**p-value**
Inflated model	33	(-8.36,73.63)	86	(44.64,126.63)	0.437

7 Conclusion

The ratio regression approach was discussed and it could be seen how the ratio regression approach for the positive sample could be extended to include information from the validation sample, the untruncated sample including also zero counts which are not observed in conventional capture-recapture settings. Including validation samples will reduce bias and increase efficiency. The identical model might be used for positive and validation sample, or a partly congruent model such as the parallel lines model, or two separate models such as the separate lines model. In the latter case, there is no gain in efficiency. Using the ratio regression approach there are numerous ways in selecting the right model. We have focused here on the Wald-statistic selecting significant coefficients. Other ways would be the likelihood ratio statistic or model selection criteria such as AIC (Böhning *et al.*, 2016). We see the most important aspect of the use of validation information in the fact that more trust can be developed in the model for the unobserved part.

References

Arnold, M.E., Martelli, F., McLaren, I., Davies, R.H. (2014). *Estimation of the rate of eggs contamination from salmonella infected chickens.* Zoonoses and Public Health 61:18-27.

Böhning, D., Rocchetti, I., Alfó, M., Hollings, H. (2016). *A flexible ratio regression approach for zero truncated capture-recapture counts.* Biometrics DOI: 10.1111/biom.12485.

Böhning, D. (2016). *Ratio plot and ratio regression with application to social and medical sciences. Statistical Sciences.* Statistical Science arXiv:math.PR/0000000 (to appear).

Modelling of Varying Dispersion in Cumulative Regression Models

Moritz Berger[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: moritz.berger@stat.uni-muenchen.de

Abstract: In ordinal regression models a potential variation of dispersion is often ignored. However, the omission of present dispersion effects might affect the accuracy of estimates. A cumulative model that accounts for varying dispersion is proposed. It can be embedded into the framework of multivariate generalized linear models. The model allows to use well-established computational tools and yields an easy interpretation of location and dispersion effects.

Keywords: Proportional Odds Model; Dispersion Modelling; Ordinal Responses

1 Cumulative Models for Location and Dispersion

Let $Y_i \in \{1, \ldots, k\}$ denote the response and $\boldsymbol{x}_i$ a vector of explanatory variables. With $\pi_i(r) = P(Y_i \leq r|\boldsymbol{x}_i)$ the basic form of the simple cumulative regression model is given by

$$\pi_i(r) = F(\theta_r + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}), \quad r = 1, \ldots, k-1, \tag{1}$$

where $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \ldots, \beta_p)$ and $F(\cdot)$ is a cumulative distribution function, see, for example, the seminal paper of McCullagh (1980) or Agresti (2010). By choosing $F(\cdot)$ as the logistic distribution function one obtains the most widely used *proportional odds model*. An attractive feature of the model is the easy interpretation of parameters. If the fit of this simple model is unsatisfactory one often uses an extended version of the model with category-specific parameters. That means the predictor is replaced by $\theta_r + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_{\boldsymbol{r}}$. The corresponding *partial proportional odds model* was, inter alia, investigated by Brant (1990) and Peterson and Harrell (1990). However, a lack-of-fit can also be caused by an insufficient modelling of dispersion effects.

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).

A cumulative type model that accounts for dispersion effects, also called *location-scale model*, was introduced by McCullagh (1980) and is given by

$$\pi_i(r) = F\left(\frac{\gamma_{0r} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}{\tau_{\boldsymbol{x}_i}}\right), \quad r = 1, \ldots, k-1, \tag{2}$$

where an additional scale parameter $\tau_{\boldsymbol{x}_i}$ is included. Of course, one has to find appropriate ways to link the dispersion parameter to the covariates $\boldsymbol{x}_i$. For example, one can use $\tau_{\boldsymbol{x}_i}$ =exp($\boldsymbol{x}_i^{\mathrm{T}}\gamma$). Model (2), is highly non-linear and not a member of the class of generalized linear models. Here we propose a model that models dispersion by including specific effects in the linear predictor of model (1), to obtain a model that can be estimated within the generalized linear model framework.

Let us consider the case of a symmetric response, for example with categories of agreement as *strongly disagree*, *moderately disagree*,...,*moderately agree*, *strongly agree* and an even number of response categories k. Then $m = \lfloor k/2 \rfloor$ splits the response categories into equally sized sets. With an additional vector of explanatory variables $\boldsymbol{z}_i$, which can be the same or different from $\boldsymbol{x}_i$, the thresholds θ_r in the proposed extended cumulative model are determined by

$$\begin{aligned}
\theta_r &= \beta_{0r} - s_r\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha}, \quad r = 1, \ldots, m-1\\
\theta_m &= \beta_{0m}\\
\theta_r &= \beta_{0r} + s_r\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha}, \quad r = m+1, \ldots, k-1.
\end{aligned}$$

While the middle threshold θ_m remains fixed, the lower and upper thresholds are shifted by $\delta = s_r\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha}$, where s_r are scale values that reflect the distance between categories r and m. If δ is positive the intervals defined by the thresholds are widened, indicating weaker dispersion, if δ is negative the intervals are shrunk, indicating stronger dispersion. If one simply chooses $s_1 = \ldots = s_{k-1} = 1$, all the thresholds are shifted away from the middle by the value $\delta = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha}$. A more attractive choice is obtained by shifting of the thresholds proportional to the distance from the middle threshold. Then one uses $s_r = m - r$ for $r = 1, \ldots, m$ and $s_r = r - m$ for $r = 1, \ldots, k-1$ to obtain the model

$$\begin{aligned}
\pi_i(r) &= F(\beta_{0r} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - (m-r)\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha}), \quad r = 1, \ldots, m\\
\pi_i(r) &= F(\beta_{0r} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + (r-m)\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha}), \quad r = m+1, \ldots, k-1.
\end{aligned} \tag{3}$$

The effect is that the intervals for all categories are widened or shrunk by the value $\delta = \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha}$. It is easily derived that for $\delta \to \infty$ one obtains $P(Y = m|\boldsymbol{x}_i, \boldsymbol{z}_i) + P(Y = m+1|\boldsymbol{x}_i, \boldsymbol{z}_i) \to 1$, which means a tendency towards the middle categories and therefore weak dispersion. For an odd number of response categories k there is a neutral middle category $m = \lfloor k/2 \rfloor + 1$. In this case the widening of intervals by the fixed value δ yields the model

$$\begin{aligned}
\pi_i(r) &= F(\beta_{0r} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - [(m-r-1) + 1/2]\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha}), \quad r = 1, \ldots, m-1\\
\pi_i(r) &= F(\beta_{0r} + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + [(r-m) + 1/2]\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\alpha}), \quad r = m, \ldots, k-1.
\end{aligned} \tag{4}$$

Since models (3) and (4) are defined by a shifting of thresholds they are called *location-shift models* with location parameters $\boldsymbol{\beta}$ and dispersion parameters $\boldsymbol{\alpha}$.

2 Interpretation of Parameters

For simplicity we consider an even number of response categories k and one-dimensional x and z. If x and z are distinct it is easily derived that in model (3) and (4) the proportional odds assumption still holds for x. With $\gamma(r|x_i, z_i) = P(Y \leq r|x_i, z_i)/P(Y > r|x_i, z_i)$, denoting the odds for categories smaller or equal r, the location parameter β is given by

$$e^{\beta} = \frac{\gamma(r|x_i + 1, z_i)}{\gamma(r|x_i, z_i)}.$$

That means, if x_i increases by one unit the cumulative odds for each category change by the factor e^{β}. For the dispersion parameter α one obtains

$$\begin{aligned} e^{-(m-r)\alpha} &= \frac{\gamma(r|x_i, z_i + 1)}{\gamma(r|x_i, z_i)}, \quad r \in \{1, \dots, m\} \\ e^{(r-m)\alpha} &= \frac{\gamma(r|x_i, z_i + 1)}{\gamma(r|x_i, z_i)}, \quad r \in \{m+1, \dots, k-1\}. \end{aligned}$$

Thus, if z_i increases by one unit the cumulative odds for categories $r < m$ change by factor $e^{-(m-r)\alpha}$ and for categories $r > m$ by the factor $e^{(r-m)\alpha}$. For $\alpha > 0$ the probabilities for extreme categories get smaller, which means a stronger concentration in the middle. If $x = z$ the interpretation of parameters is similar. For the cumulative odds one obtains

$$\frac{\gamma(r|x_i + 1)}{\gamma(r|x_i)} = \begin{cases} e^{\beta} e^{-(m-r)\alpha}, & r \in \{1, \dots, m\} \\ e^{\beta} e^{(r-m)\alpha}, & r \in \{m+1, \dots, k-1\}. \end{cases}$$

Thus e^{β} represents the odds ratio for categories smaller or equal m if x_i increases by one unit. This basic preference is modified by factor $e^{-(m-r)\alpha}$ for categories $r < m$ and by factor $e^{(r-m)\alpha}$ for categories $r > m$.
For an odd number of response categories the interpretations are the same, but they hold for different response categories.

3 Computation of Parameters and Implementation

The strength of the proposed location-shift models is that they can be embedded within the framework of multivariate generalized linear models (GLMs), which allows to use known asymptotic results and goodness-of-fit tests for this class of models. The models have the form $g(\boldsymbol{\pi}_i) = \boldsymbol{X_i \delta}$, where $\boldsymbol{\pi}_i^{\mathrm{T}} = (\pi_{i1}, \dots, \pi_{i,k-1})$ is the vector of response probabilities with

TABLE 1. Quality of right eye vision in men and women.

	Vision Quality			
	Highest (1)	2	3	Lowest (4)
Men	1053	782	893	514
Women	1976	2256	2456	789

TABLE 2. Parameter estimates and standard errors for quality of eye vision data.

Covariate	**Proportional Odds Model**			**Location-Shift Model**		
	estimate	std error	z value	estimate	std error	z value
Intercept1	-0.905	0.034	-26.613	-0.721	0.037	-19.397
Intercept2	0.293	0.033	8.911	0.236	0.033	7.104
Intercept3	2.005	0.039	50.398	1.710	0.045	37.563
gender (female)						
location	-0.038	0.038	-1.003	0.042	0.038	1.109
dispersion				0.353	0.031	11.348

components $\pi_{ir} = P(Y_i = r|\boldsymbol{x}_i, \boldsymbol{z}_i)$, $\boldsymbol{X_i}$ is a design matrix constructed from the predictors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, $\boldsymbol{\delta}$ is the total parameter vector and $g(\cdot)$ is a vector-valued link function. For details of the representation as a multivariate GLM see Tutz (2012). Estimates can be obtained by using the function `vglm()` implemented in the R-package `VGAM` (Yee, 2010), which allows to compute so-called vector generalized linear models. By appropriate specification of the design matrix that includes the $\boldsymbol{z}$-variables in the specific form the proposed location-shift model with dispersion effects can easily be obtained.

4 Application

For illustration we consider an example that has already been used by McCullagh (1980). Table 1 gives Stuart's (1953) quality of eye vision data for men and women. From the data it is obvious that women are more concentrated in the middle categories while men have relatively high proportions in the extreme categories. The estimated coefficients and corresponding standard errors of the simple proportional odds model and the proposed location-shift model (3) are shown in Table 2. It is seen that in both models the location effect ($\hat{\beta} = -0.038$ and $\hat{\beta} = 0.042$) is rather weak and not significant at significance level $\alpha = 0.05$. In contrast the dispersion parameter in the location-shift model $\hat{\alpha} = 0.353$ can definitely not be neglected. The deviance of the proportional odds model is 128.39 on 2 df but reduces to 5.896 on 1 df for the model with location and dispersion effect. The estimated factor $e^{-\hat{\alpha}} = 0.70$ decreases the odd for category 1 and $e^{\hat{\alpha}} = 2.01$ increases the odd for categories smaller or equal 3 for females when compared to males.

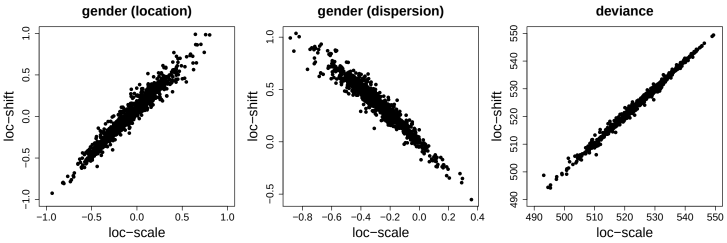


FIGURE 1. Parameter estimates and deviances of model fits for sub-samples of size $n = 200$ from the quality of eye vision data.

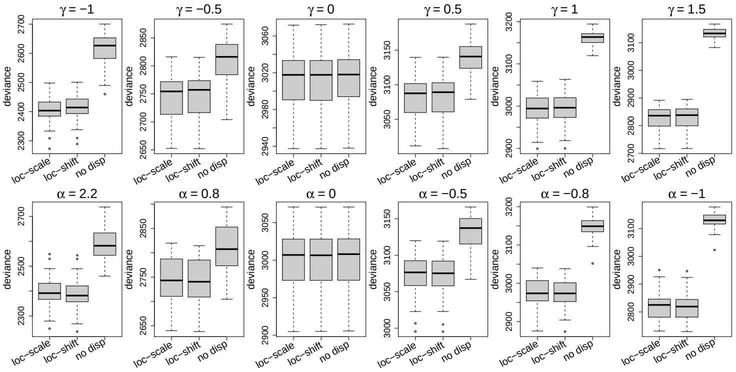


FIGURE 2. Deviances for data generated by the location-scale model (first row) and data generated by the corresponding location-shift model (second row).

5 Comparison of Models

In the proposed location-shift models (3) and (4) the dispersion is modelled by an explicit shifting of the thresholds determined by the parameters $\boldsymbol{\alpha}$. In contrast, in the location-scale model (2) the dispersion is generated by the scale parameter $\tau_{\boldsymbol{x}_i}$ and the effect is multiplicative on the thresholds. Although the two models are not equivalent we found that in applications the differences in terms of goodness-of-fit can be rather small.

We first consider again the eye vision data example (Table 1). Figure 1 shows the estimated location effects, dispersion effects and the deviances of the two models based on 100 sub-samples of size $n = 200$. It is seen that the estimates and deviances show strong correlation. In particular the deviances of the two models are very close.

Secondly, we illustrate the fitting in a small simulation study. We consider two binary covariates with $\boldsymbol{\beta}^\top = (0.5, 0.5)$, $k = 5$ response categories and thresholds $\theta_r \in \{-2, \dots, 2\}$. The first row of Figure 2 shows the resulting deviances for data generated by the location-scale model with varying strength of dispersion (parameter γ) in the first variable. In order to match the strength of dispersion we computed the parameter α of the location-shift model that corresponds to the parameter γ of the location-scale model. Then data were generated by the locations-shift model. It is important to note that the relation between the two parameters is non-linear. The resulting deviances are shown in the second row of Figure 2. It is seen that the deviances of the two models are quite close with just slightly better fits of the data generating model. If, however, the dispersion is ignored and a simple proportional odds model is fitted (no disp), the fit suffers strongly. Further investigations show that the omission of present dispersion effects does not only yield large deviances but in particular reduces the accuracy of estimates of the location effects.

6 Link to the Partial Proportional Odds Model

To yield a better model fit an alternative to include dispersion effects is to introduce category-specific parameters. In the case of three response categories ($k = 3$) and $\boldsymbol{x} = \boldsymbol{z}$ the two predictors of the location-shift model are $\eta_{i1} = \beta_{01} + \boldsymbol{x}_i^\top \boldsymbol{\beta} - \boldsymbol{x}_i^\top \boldsymbol{\alpha}$ and $\eta_{i2} = \beta_{02} + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{x}_i^\top \boldsymbol{\alpha}$. This is equivalent to $\eta_{ir} = \beta_{0r} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_r$, where $\boldsymbol{\beta}_1 = \boldsymbol{\beta} - \boldsymbol{\alpha}$, $\boldsymbol{\beta}_2 = \boldsymbol{\beta} + \boldsymbol{\alpha}$. Therefore, the location-shift model is equivalent to the partial proportional odds model. Nevertheless, there are some benefits when using the location-shift parameterization. With the hypothesis $H_0 : \alpha_j = 0$ it can directly be tested, if the j-th variable has a global effect. The equivalent hypothesis within the partial proportional odds model is $H_0 : \beta_{j1} = \beta_{j2}$, which typically makes refitting of the model under constraints necessary and interpretation of effects less accessible.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data, 2nd Edition.* New York: Wiley.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, **46(4)**, 1171-1178.

McCullagh, P. (1980). Regression models for ordinal data (with Discussion). *Journal of the Royal Statistical Society B*, **42(2)**, 109-127.

Peterson, B. and Harrell, F.E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, **39(2)**, 205-217.

Tutz, G. (2012). *Regression for Categorical Data.* Cambridge: University Press.

Yee, T. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, **32(10)**, 1-34.

Item Focused Trees for the Detection of Differential Item Functioning in Partial Credit Models

Stella Bollmann[1], Moritz Berger[1], Gerhard Tutz[1]

[1] Ludwig Maximilians University Munich, Germany

E-mail for correspondence: stella.bollmann@psy.lmu.de

Abstract: Recently novel recursive partitioning techniques have been proposed that allow for the detection of differential item functioning (DIF) being induced by an arbitrary number of covariates. While several methods are available for the simple binary Rasch model the extension to rating scale items is still in its infancy. In the present paper we propose item focussed trees for the ordinal Partial Credit Model (PCM). The new procedure is compared to a global recursive partitioning approach for DIF detection in PCMs proposed by El-Komboz et al. (2014).

Keywords: Differential Item Functioning; Partial Credit Model; Item focussed Trees.

1 Introduction

For proper measurement, psychometric test models generally assume that test and measurement properties are stable across individuals what is also known as measurement invariance (Millsap, 2012). However, there is a strong possibility that different groups of people react differently on the same test and therefore validity of studies on measurement properties is threatened. Also, test fairness may be violated when test diagnoses lead to different conclusions for different groups of people. When measurement equivalence is violated on the item level it is called item bias or differential item functioning (DIF). DIF is present when one item is significantly more difficult for one group than for the other after controlling for the underlying ability or trait. For an overview of methods for the detection of DIF, see Magis et al. (2010) and Holland and Wainer (1993).
Recently a strategy was proposed that is able to detect DIF in Rasch models generated by multiple covariates. It uses recursive partitioning techniques,

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

often called tree methods. The strong advantage of the method is that no pre-specified subgroups are needed. One has to distinguish between two quite different forms of recursive partitioning method in DIF detection. Strobl et al. (2015) proposed a global recursive partitioning technique for the partial credit model (PCM) called PC trees. Here, the covariate space is recursively partitioned to identify regions of the covariate space in which item parameters differ. Regions are suspected to be relevant if the parameter estimates in the regions differ strongly. Therefore, regions in the covariate space are identified that show different difficulties. A disadvantage of the method is that it remains unknown which of the items show DIF. In order to overcome this problem, Tutz et al. (2015) suggested the *Item focussed tree* (IFT) method for the detection of DIF for single items in the binary Rasch model. Recursive partitioning is used on the item level not on the global level. In contrast to the PC tree it directly identifies items that carry DIF. Since the method is able to flag DIF items it is referred to as item-focussed trees (IFTs). El-Komboz et al. (2014) extended the PC tree approach to models with multiple categories. In the following also the IFT approach is extended to allow for DIF detection in the partial credit model.

2 DIF in Partial Credit Models

In the following we consider I items with ordered categories and P persons. For simplicity we assume that the number of categories K is equal across items.

Let the data be given by the response on a rating scale $Y_{pi} \in \{0, 1, \dots, K\}$, of person p on item i. The partial credit model (PCM), which was proposed by Masters (1982), assumes for the probabilities

$$P(Y_{pi} = r) = \frac{\exp \sum_{l=1}^{r} \theta_p - \delta_{il}}{\sum_{s=0}^{K} \exp(\sum_{l=1}^{s} \theta_p - \delta_{il})} \qquad r = 0, \dots, K$$

where θ_p is the parameter for person p and $(\delta_{il}, \dots, \delta_{iK})$ are the item parameters of item i. For notational convenience, the definition of the model uses implicitly $\sum_{l=1}^{0} \theta_p - \delta_{il} = 0$.

The link to the binary Rasch model becomes obvious if one considers responses in adjacent categories. Given response categories r and $r-1$, the presentation

$$\log \frac{P(Y_{pi} = r)}{P(X_{pi} = r-1)} = \theta_p - \delta_{ir}, \quad r = 1, \dots, K \tag{1}$$

shows that the model is locally a binary Rasch model with person parameter θ_p and item difficulty δ_{ir}.

Therefore, it is possible to embed the partial credit model into the framework of vector generalized linear models (VGLMs; Yee and Wild, 1996).

For the item responses one assumes a multinomial distribution $Y_{pi}|\boldsymbol{x}_p \sim M(1, \boldsymbol{\pi}_{pi})$, where $\boldsymbol{\pi}_{pi}^\top = (\pi_{pi1}, \ldots, \pi_{piK})$ with components $\pi_{pir} = P(Y_{pi} = r|\boldsymbol{x}_p)$. The link function has the form

$$g(\pi_{pir}) = \eta_{pir} = \log\left(\frac{P(Y_{pi} = r)}{P(Y_{pi} = r-1)}\right) = (\mathbf{1}_p^{(P)})^\top \boldsymbol{\theta} - (\mathbf{1}_r^{(k)})^\top \boldsymbol{\delta}_i, \qquad (2)$$

where $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_P)$, $\boldsymbol{\delta}_i^\top = (\delta_{i1}, \ldots, \delta_{iK})$ and $\mathbf{1}_r^{(k)}$ denotes the unit vector of length k with a 1 in component r. To ensure the identifiability of the model one parameter has to be fixed. In the following we set $\theta_P = 0$. By defining the whole parameter vector $\boldsymbol{\beta}^\top = (\boldsymbol{\theta}^\top, \boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_I^\top)$ the PCM can be written in the closed form

$$\eta_{pir} = \boldsymbol{z}_{pir}\boldsymbol{\beta},$$

where $\boldsymbol{z}_{pir}$ is the design vector for person p, item i and threshold r that has to be specified accordingly.

For the implementation of the vector generalized model we make use of the function vglm of the package VGAM (Yee, 2010). One just has to specify the design matrix as described above and estimation can easily be obtained. In addition one can make use of the argument `parallel()` to specify category-specific item parameters. In the following algorithm for item focussed trees this estimation procedure serves as building block in each iteration. All the results presented in this article were obtained by an R program that is available from the authors.

3 Fitting Trees

When growing trees one has to take two decisions in each step. One has to determine the best split due to an optimality criterion and has to decide if the split is relevant or not. In contrast to alternative approaches the trees are not pruned to an adequate size afterwards. By early stopping the size of the trees is directly controlled beforehand.

Let $\mathbf{x}_p^T = (x_{p1}, \ldots, x_{pV})$ denote a person specific covariate vector of length V. To determine the first split one examines for all the items, all the variables and possible split-points the PCM with predictors

$$\eta_{pir} = \theta_p - [\gamma_{ir(1)} I(x_{pv} \leq c_{vj}) + \gamma_{ir(2)} I(x_{pv} > c_{vj})], \quad r = 1, \ldots, k.$$

where $I(\cdot)$ denotes the indicator function with $I(\alpha) = 1$ if α is true and $I(\alpha) = 0$ otherwise.

Here c_{vj} indicates the split point j in variable v. The parameter $\gamma_{ir(l)}$ denotes the threshold parameter of item i and threshold r in the left node $(x_{pv} \leq c_{vj})$ and $\gamma_{ir(r)}$ in the right node $(x_{pv} > c_{vj})$. This parameter may differ across thresholds within one item which means that not all thresholds within one item are shifted by an equal amount.

DIF occurs, if $\gamma_{i(1)} \neq \gamma_{i(2)}$. The corresponding hypothesis $H_0 : \gamma_{i(1)} - \gamma_{i(2)} = \mathbf{0}$ can be tested by a likelihood ratio (LR) test. One simply selects the combination of item, variable and split-point that yields the smallest p-value, which is equivalent to select the model with minimal deviance. In later steps the basic procedure is the same. One performs LR-tests testing the two parameters that are involved in the splitting and selects the combination as the optimal one that yields the smallest p-value.

In order to determine the optimal size of the trees one has to decide in each step if the split should be performed or not. In answering this question one investigates the dependence of the response and the selected variable. For fixed item i and variable v let the maximal value statistic $T_v = max_{c_v} T_{vc_v}$ be defined as the maximum of all the LR test statistics T_{vc_v}, where c_v is from the set of possible split points. Typically the test statistics T_{vc_v} are strongly correlated. The relevance of variable v is judged by the p-value of the distribution of T_v, which is not influenced by the number of split-points, since it has already taken into account, see e.g. Hothorn and Lausen (2003). For the decision on the null hypothesis controlling for a given significance level α a permutation test is used. Thus no distributional assumption has to be made. The test statistic T_v is computed based on a data matrix in which variable v is randomly permuted. The maximal value statistics for a large number of permutations provide a distribution of T_v under the assumption of the null hypothesis that variable v has no effect. The derived p-value is used to make the splitting decision.

Finally one has to address the problem of multiple testing. In DIF detection one typically controls for the type I error, that is, the item-wise significance level. To ensure that the proposed procedure also controls this level a Bonferroni adjustment is applied when multiple variables are available. For fixed item and variable the local significance level for one permutation test is set to α/V, where V is the number of variables. Using this adaption the probability of a false DIF result or the probability of falsely identifying at least one variable as responsible for DIF is controlled by α.

For determining when to stop splitting, we use two different criteria: 1. The minimum sample size of $P = 30$ per node and 2. No (additional) significant permutation test. The significance of model improvement was determined by using a permutation test.

4 Simulation studies

In two simulation studies, the proposed procedure was compared to the alternative PC tree approach. In one simulation, only one binary covariate was simulated that induces DIF of three different strengths. Additionally, the number of items and the number of categories per item were varied. Results of this simulation are shown in Table 1.

Percentages of significant test results are shown. For PC tree these have to

TABLE 1. **Results of Simulation I:** Percentages of significant test results for both procedures

DIF strength	**PC tree**	**PCM-IFT**	
		TPR	**FPR**
	8 Items with 3 categories		
no DIF	0.100	-	0.048
weak	0.180	0.193	0.068
medium	0.660	0.560	0.124
strong	1.000	1.000	0.036
	8 Items with 5 categories		
no DIF	0.040	-	0.060
weak	0.220	0.260	0.064
medium	0.860	0.847	0.076
strong	1.000	1.000	0.080
	20 Items with 3 categories		
no DIF	0.060	-	0.047
weak	0.140	0.233	0.054
medium	0.580	0.780	0.056
strong	1.000	1.000	0.053

be interpreted as false posite rates (FPR) in the *no DIF* condition, as a true positive rates (TPR) in all other conditions.

Results show that the new procedure can compete with the alternative one. For 20 Items (scenario 3), PCM-IFT yields the best results.

5 Application

As an example we consider the norm data from the German version of the personality test NEO-PI R (Ostendorf & Angleitner, 2004). It is designed to measure personality in five domains. An example tree of Item 6 of the sub-facet *Fantasy* of the domain *Opennes to experiences* is shown in Figure 1.

At the terminal nodes, the four threshold parameters for the respective partition are shown in a graphical representation. It can be seen that threshold parameters differ between the three groups that were built by one split for the variable gender and one split for the covariate age.

References

El Komboz, B.A., Zeileis, A. and Strobl, C. (2014). Detecting Differential Item and Step Functioning with Rating Scale and Partial Credit Trees. *Technical Report 152, Department of Statistics LMU.*

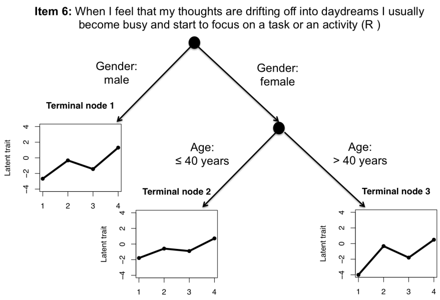


FIGURE 1. Tree for Item 6 of the sub-facet *Fantasy.*

Holland, W. and Wainer, H. (1993). *Differential Item Functioning.* Lawrence Erlbaum Associates.

Hothorn, T. and Lausen, B. (2010). On the exact distribution of maximally selected rank statistics. *Computational Statistics and Data Analysis* **43**, 121 – 137.

Magis, D., Bland, F., Tuerlinckx, F. and Boeck, P. (2010). A general framework and an r package for the detection of dichotomous differential item functioning. *Behavior Research Methods* **42(3)**, 847 – 862.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance.* Routledge.

Ostendorf, F. and Angleitner, A. (2004). NEO-Persoenlichkeitsinvetar nach Costa und McCrae, revidierte Fassung. Goettingen, Hogrefe.

Strobl, C., Kopf, J. and Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, **80(2)**, 289 – 316.

Tutz, G. and Berger, M. (2016). Item Focussed Trees for the Identification of Items in Differential Item Functioning. *Psychometrika*, to appear, DOI: 10.1007/s11336-015-9488-3.

Yee, T. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, **32(10)**, 1 – 34.

Yee, T. and Wild, C. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society B*, 481 – 493.

Inference in a Stochastic SIR Epidemic Model using Bayesian Filtering

Wilfried Bonou[1], Philippe Lambert[1,2]

[1] Faculté des Sciences Sociales, Méthodes Quantitatives en Sciences Sociales, Université de Liège, Liège, Belgium. ,

[2] Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

E-mail for correspondence: w.bonou@ulg.ac.be

Abstract: We consider State Space Models (SSMs) as Discrete Time Markov Chains (DTMC) to describe a stochastic SIR Epidemic dynamic. The unknown static parameters are estimated by combining Sequential Monte Carlo and Markov Chain Monte Carlo algorithms (SMC-within-MCMC) also known as Particle Marginal Metropolis-Hastings (PMMH). The performances of the strategy are evaluated using simulations. The method is illustrated by modeling the spread of a viral infection in a small community.

Keywords: Stochastic SIR Epidemic Model; State Space Models; Sequential Monte Carlo; Particle Marginal Metropolis-Hastings.

1 Introduction and context

The following data provide the evolution of the number of the new infected subjects in a small community ($N = 40$) over a 21-day period during a Common-Cold epidemic on the island of Tristan da Cunha (see Table 1, Shibli et al., 1971).

TABLE 1. Common-cold epidemic data on the island of Tristan da Cunha.

Day	1	2	3	4	5	6	7	8	9	10	11	12	...	21
# *new cases*	1	0	2	4	4	6	4	5	1	3	3	1	...	0

We suggest to use the stochastic SIR model pictured in Figure 1 to describe

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 propagation of the epidemic. It is given by the following set of equations

$$\begin{aligned} S_t &= S_{t-dt} - dI^+_{t-dt}, \\ I_t &= I_{t-dt} + dI^+_{t-dt} - dR^+_{t-dt}, \\ R_t &= R_{t-dt} + dR^+_{t-dt}, \end{aligned} \tag{1}$$

where the capital letters indicate the total number of susceptible (S_t), infectious (I_t) and recovered (R_t) subjects at time t, and dI^+_{t-dt} and dR^+_{t-dt} provide the number of new entries (also known as increments or innovations) in each corresponding state from $(t - dt)$ to t. The time step dt is chosen small enough to ensure that a single person can only experience at most one state transition during $(t - dt, t)$.

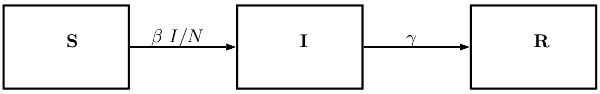


FIGURE 1. Graphical representation of a SIR model.

The innovations are assumed Poisson distributed:

$$\begin{aligned} (dI^+_{t-dt}|S_{t-dt}, I_{t-dt}, \boldsymbol{\theta}) &\sim \text{Pois}\left(\psi S_{t-dt}\frac{I_{t-dt}}{N}dt\right) \\ (dR^+_{t-dt}|I_{t-dt}, \boldsymbol{\theta}) &\sim \text{Pois}(\gamma I_{t-dt}dt), \end{aligned}$$

with $\boldsymbol{\theta} = (\psi, \gamma)$ and $\psi = \beta\pi$ where γ is the recovery rate and β is the average number of contacts per susceptible subject during $(t - dt, t)$ that lead to a disease transmission with probability π when involving an infectious person.

2 A State-space model (SSM) formulation

The aforementioned stochastic SIR model verifies the Markov process properties (Särkkä, 2013) and therefore can be treated as a SSM. It consists of a sequence of conditional probability distributions:

$$(\mathbf{x}_t|\mathbf{x}_{t-dt}, \boldsymbol{\theta}) \sim f_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{t-dt}) \tag{2}$$

$$(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}) \sim g_{\boldsymbol{\theta}}(\mathbf{y}_t|\mathbf{x}_t) \tag{3}$$

where $\mathbf{x}_t \in \mathcal{R}^n$ ($t = 0, dt, 2dt, ..., T$ and $n \in \mathbb{N}^*$) denotes the (hidden) states of the system and $\mathbf{y}_t \in \mathcal{R}^q$ ($t = dt, 2dt, ..., T$ and $q \in \mathbb{N}^*$) is the vector of observed measurements. The function $f_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{t-dt})$ describes the stochastic dynamic in the system and $g_{\boldsymbol{\theta}}(\mathbf{y}_t|\mathbf{x}_t)$ connects the observed data to the latent states. Prior distributions on $\boldsymbol{\theta}$ ($\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \Theta \subseteq \mathcal{R}^d$: $d \in \mathbb{N}^*$) and on the state $\mathbf{x}$ ($\mathbf{x}_0 \sim \mu_{\boldsymbol{\theta}}(\mathbf{x}_0)$) complete the model specification. If (only) the number of newly infected subjects were reported every

dt units of time, then the connection to our epidemic SIR model could be obtained by setting $\mathbf{x}_t = (S_t, I_t, R_t)^T$ and $\mathbf{y}_t = dI_{t-dt}^{+(obs)}$.

Bayesian inference relies on the posterior density for the unknown quantities:

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{x}_{0:T}|\mathbf{y}_{dt:T}) &\propto p_{\boldsymbol{\theta}}(\mathbf{x}_{0:T}, \mathbf{y}_{dt:T})p(\boldsymbol{\theta}) \\ &= \mu_{\boldsymbol{\theta}}(\mathbf{x}_0) \left\{ \prod_{t \geq dt} f_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{t-dt}) g_{\boldsymbol{\theta}}(\mathbf{y}_t|\mathbf{x}_t) \right\} p(\boldsymbol{\theta}). \qquad (4) \end{aligned}$$

There are generally no closed form expressions for $p(\boldsymbol{\theta}, \mathbf{x}_{0:T}|\mathbf{y}_{1:T})$ in the context of non-linear non-Gaussian models. That makes inference difficult for $\boldsymbol{\theta}$ and the hidden states $\mathbf{x}_{0:T}$ (Andrieu et al., 2010). But SMC and MCMC methods can be combined to approximate the posterior densities sequentially, see Section 3.

3 Particle Marginal Metropolis-Hastings (PMMH)

Also named SMC-within-MCMC, PMMH is an *hybridization* of two algorithms: one for the state approximation and the other one for static parameter estimation. To achieve this, we need, first, to compute the so-called *Marginal likelihood* and the weights of particles representing possible state trajectories obtained by Monte Carlo simulation.

Assume that the innovations were observed in an aggregated way over the time interval of length:

$$\triangle = \delta dt \text{ with } \delta \in \mathbb{N}^*.$$

It corresponds to 1-day period in our application. Denote by $\mathbf{y}_{t-\triangle}^t$ the number of new cases reported during $(t-\triangle, t)$. The previous report on the aggregated number of new cases provided information on the state values till time $(t - \triangle - dt)$. Possible trajectories for the state vectors can be generated sequentially using (2) by steps of dt units of time till $(t - dt)$. Conditionally on a state trajectory, the probability to observe $\mathbf{y}_{t-\triangle}^t$ can be calculated to enter the computation of the plausibility of the state updates.

Our proposal works iteratively as follows. Assume that P state trajectories (named *particles*) were generated till time $(t - \triangle - dt)$:

$$\mathbf{x}_{0:t-\triangle-dt}^{(i)} = (\mathbf{x}_0^{(i)}, \mathbf{x}_{dt}^{(i)}, \mathbf{x}_{2dt}^{(i)}, \ldots, \mathbf{x}_{t-\triangle-dt}^{(i)}) : i = 1, \ldots, P,$$

with posterior probability $\mathbf{w}_{t-\triangle-dt}^{(i)}$ given data $(\mathbf{y}_0^{\triangle}, \mathbf{y}_{\triangle}^{2\triangle}, \ldots, \mathbf{y}_{t-2\triangle}^{t-\triangle})$.

For $i = 1, \dots, P$**, do:**

- Update the ith trajectory by sampling new state values sequentially using (2):

$$\begin{aligned}
(\mathbf{x}^i_{t-\triangle}|\mathbf{x}^i_{t-\triangle-dt}, \boldsymbol{\theta}) &\sim p(\mathbf{x}^i_{t-\triangle}|\mathbf{x}^i_{t-\triangle-dt}, \boldsymbol{\theta}) \\
(\mathbf{x}^i_{t-\triangle+dt}|\mathbf{x}^i_{t-\triangle}, \boldsymbol{\theta}) &\sim p(\mathbf{x}^i_{t-\triangle+dt}|\mathbf{x}^i_{t-\triangle}, \boldsymbol{\theta}) \\
&\vdots \\
(\mathbf{x}^i_{t-dt}|\mathbf{x}^i_{t-2dt}, \boldsymbol{\theta}) &\sim p(\mathbf{x}^i_{t-dt}|\mathbf{x}^i_{t-2dt}, \boldsymbol{\theta})
\end{aligned} \tag{5}$$

- Given that $\mathbf{y}^t_{t-\triangle} = \sum_{k=1}^{\delta} \mathbf{y}^{t-\triangle+kdt}_{t-\triangle+(k-1)dt}$ and (3), one can conclude that $(\mathbf{y}^t_{t-\triangle}|\mathbf{x}^{(i)}_{t-\triangle}, \dots, \mathbf{x}^{(i)}_{t-dt}) \sim \text{Pois}(\mu^{t-dt^{(i)}}_{t-\triangle})$ where

$$\mu^{t-dt^{(i)}}_{t-\triangle} = \psi \left(S^{(i)}_{t-\triangle} \frac{I^{(i)}_{t-\triangle}}{N} + \dots + S^{(i)}_{t-dt} \frac{I^{(i)}_{t-dt}}{N} \right) dt. \tag{6}$$

End for.

Therefore, the plausibility of the ith updated trajectory can be computed recursively as follows:

$$\mathbf{w}^i_{t-dt} \propto \mathbf{w}^i_{t-\triangle-dt} \text{ dpois}(\mathbf{y}^t_{t-\triangle} = dI^{t^{(obs)}}_{t-\triangle}, \mu^{t-dt^{(i)}}_{t-\triangle}) \tag{7}$$

Building upon these ideas, we extend the Bootstrap particle filter (SMC) algorithm (Andrieu et al., 2010) to estimate the states of the system model:

***Algorithm 1*:** State estimation in a SIR epidemic model using the Bootstrap Particle Filter (BPF) given $\boldsymbol{\theta}$:

1. A time t = 0, ***initialize*** each *particle* i ($i = 1, \dots P$) by sampling $\mathbf{x}^{(i)}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{x}_0)$, and set the particle weights to $\mathbf{w}^{(i)}_0 = 1/P$.
2. **For** $t = \triangle, \dots, T = \tau\triangle$**, do:**
   (a) Resample the particles $(\mathbf{x}^{(i)}_{0:t-\triangle-dt}, \mathbf{w}^{(i)}_{t-\triangle-dt})$ resulting in equally weighted particles $\{(\tilde{\mathbf{x}}^{(i)}_{0:t-\triangle-dt}, 1/P), i = 1, \dots, P\}$;
   (b) Update the particle trajectories using (5), yielding $\mathbf{x}^{(i)}_{0:t-dt}$;
   (c) Compute the weights $\mathbf{w}^{(i)}_{t-dt}$ using (7), then:

$$\mathbf{w}^{(i)}_{t-dt} \longleftarrow \mathbf{w}^{(i)}_{t-dt} / \sum_{j=1}^{P} \mathbf{w}^{(j)}_{t-dt};$$

(d) Compute the state estimate (or more exactly, the expected state value) as follows: $\hat{\mathbf{x}}^P_{0:t-dt} = \sum_{i=1}^P \mathbf{w}^{(i)}_{t-dt}\mathbf{x}^{(i)}_{0:t-dt}$;

End for.

The logarithm of the marginal likelihood:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{y}_{0:\tau\triangle:\triangle}) &= p_{\boldsymbol{\theta}}(\mathbf{y}^{\triangle}_0, \ldots, \mathbf{y}^{\tau\triangle}_{\tau\triangle-\triangle}) \\ &= p_{\boldsymbol{\theta}}(\mathbf{y}^{\triangle}_0)\prod_{s=2}^{\tau} p_{\boldsymbol{\theta}}(\mathbf{y}^{s\triangle}_{(s-1)\triangle}|\{\mathbf{y}^{k\triangle}_{0:(k-1)\triangle} : k = 1, \ldots, s-1\}) \end{aligned} \tag{8}$$

can be approximated sequentially using the particle weights, yielding:

$$\log p_{\boldsymbol{\theta}}(\mathbf{y}_{0:\tau\triangle:\triangle}) = \sum_{s=1}^{T} \log\left(\frac{1}{P}\sum_{i=1}^{P}\mathbf{w}^{(i)}_s\right). \tag{9}$$

Then, we sequentially estimate the unknown parameter $\boldsymbol{\theta}$ by using the PMMH algorithm (Andrieu et al., 2010):

***Algorithm 2*:** Static parameter estimation in a SIR epidemic model using the PMMH scheme:

1. Initialization ($m = 0$):

   (a) Set an arbitrary $\boldsymbol{\theta}^0$;

   (b) Run the (SMC) Algorithm 1 to sample $\mathbf{x}^0_{0:T-dt} \sim \hat{p}_{\boldsymbol{\theta}^0}(.|\mathbf{y}_{0:T-\triangle:\triangle})$ and let $\hat{p}_{\boldsymbol{\theta}^0}(\mathbf{y}_{0:\tau\triangle:\triangle})$ denote the marginal likelihood estimate (see Algorithm 1) at that iteration;

2. MCMC step in a for loop ($m \geq 1$):

   (a) Sample $\boldsymbol{\theta}^* \sim q(.|\boldsymbol{\theta}^{m-1})$ (a proposal distribution for $\boldsymbol{\theta}$);

   (b) Run Algorithm 1 to sample $\mathbf{x}^*_{0:T-dt}$ from $p_{\boldsymbol{\theta}^*}(.|\mathbf{y}_{0:\tau\triangle:\triangle})$ yielding weighted trajectories for the states. Then compute the marginal likelihood estimate $\hat{p}_{\boldsymbol{\theta}^*}(\mathbf{y}_{0:\tau\triangle:\triangle})$.

   (c) With probability:

   $$1 \wedge \frac{\hat{p}_{\boldsymbol{\theta}^*}(\mathbf{y}_{0:\tau\triangle:\triangle})p(\boldsymbol{\theta}^*)}{\hat{p}_{\boldsymbol{\theta}^{m-1}}(\mathbf{y}_{0:\tau\triangle:\triangle})p(\boldsymbol{\theta}^{m-1})}\frac{q(\boldsymbol{\theta}^{m-1}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{m-1})},$$

   set $\boldsymbol{\theta}^m = \boldsymbol{\theta}^*$; $\mathbf{x}^m_{0:T-dt} = \mathbf{x}^*_{0:T-dt}$ and $\hat{p}(\mathbf{y}_{0:\tau\triangle:\triangle}|\boldsymbol{\theta}^m) = \hat{p}(\mathbf{y}_{0:\tau\triangle:\triangle}|\boldsymbol{\theta}^*)$.

   Otherwise, $\boldsymbol{\theta}^m = \boldsymbol{\theta}^{m-1}$; $\mathbf{x}^m_{0:T-dt} = \mathbf{x}^{m-1}_{0:T-dt}$ and $\hat{p}(\mathbf{y}_{0:\tau\triangle:\triangle}|\boldsymbol{\theta}^m) = \hat{p}(\mathbf{y}_{0:\tau\triangle:\triangle}|\boldsymbol{\theta}^{m-1})$.

4 Simulation and application

Five hundred datasets were simulated using the SIR model in (1) with parameter values selected to mimic the observed dynamic in the application. Two values for dt (1 and 1/4) were considered (with $\triangle = 1$ *day*) and three different numbers $(100, 500, 1000)$ of particles were tried in the inference procedure (see Section 3). It revealed that taking $P = 500$ *particles* enables to estimate the model parameters and the states consistently with no major effect of dt on the results. Application to the real data is made for $dt = 1/4$, $P = 500$ *particles* and $M = 500000$ *iterations* with a pilot run to select the variance-covariance matrix in the multivariate normal proposals for parameters. The mixing of the chains was very good with a non significant autocorrelation after a 20 iteration lag. The parameters were estimated as follows (with 95% credible intervals in brackets): $\hat{\psi} = 0.020\ (0.012, 0.033)$ (see Figure 2 *a.*); $\hat{\gamma} = 0.230\ (0.012, 0.611)$ (see Figure 2 *b.*). Figure 2 *c.* illustrates the scatterplot of the two parameters. These results suggest that a stochastic representation of dynamic models combined with a Bayesian filtering technique for estimation is a promising way to make inference with limited bias.

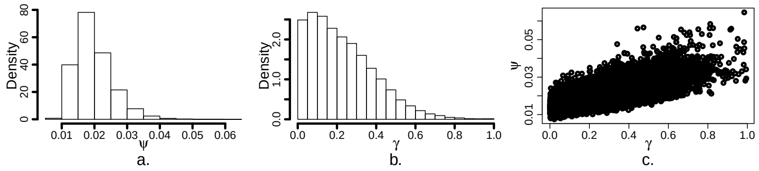


FIGURE 2. Histograms (*a.* and *b.*) and scatterplot (*c.*) of the posterior densities of parameters ψ and γ with $P = 500$ *particles* and $dt = 1/4$ from real Common-Cold data.

Acknowledgments: The authors acknowledge financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy).

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B* (with discussions), **72(3)**, 269 – 342.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing.* Institute of Mathematical Statistics Textbooks: Cambridge University Press, New York.

Shibli, M., Gooch, S., Lewis, H.E., Tyrrell, D.A.J. (1971). Common colds on Tristan da Cunha. *Journal of Hygiene*, **69**, 255 – 262.

Flexible Bayesian cure survival model with categorical time-dependent covariates : An application to fertility studies

Vincent Bremhorst[1], Philippe Lambert[1,2]

[1] Université catholique de Louvain, Belgium
[2] Université de Liège, Belgium

E-mail for correspondence: vincent.bremhorst@uclouvain.be

Abstract: Classical survival models cannot be used to study the transition to second birth. Indeed, they usually assume that any one-child mother under study will, later or sooner, become pregnant for a second time. However, one might expect that an unknown proportion of one-child mother will never have a second child. Cure survival models extend classical survival models by enabling to distinguish the women and/or family characteristics influencing the probability of having an extra child from those influencing the timing of an extra pregnancy. The promotion time cure model, first presented to handle survival data in cancer studies, argues that the observed time-to-event time is defined as the minimum required time to detect one of the underlying latent factors. These latent factors are assumed to be directly active at the beginning of the study. Since the probability of having a second child is a one-to-one function of the mean number of latent factors, only time-constant covariates can be used to model this probability. However, some women/family characteristics, such as the education attainments of the mother and of her partner for example, may vary over time. Therefore, in this work, we propose an extension of the promotion time cure model enabling to deal with categorical time-dependent covariates. Data from the German Socio-Economic Panel (GSOEP) are used to illustrate this new methodology.

Keywords: Cure fraction; Bayesian P-splines ; Categorical time-dependent covariates ; Power Variance Function ; Fertitily studies.

1 Introduction

Classical survival models cannot be used to study the transition to second birth. Indeed, they usually assume that any one-child mother under study will, later or sooner, become pregnant for the second time. However,

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).

one might expect that an unknown proportion of one-child mother will never have a second child. Cure survival models extend classical survival models by enabling to distinguish the women and/or family characteristics influencing the probability of having an extra child from those influencing its timing. In this work, we shall focus on the promotion time cure model (Chen et al., 1999). This model argues that the observed time-to-event time is defined as the minimum time for one of $N \sim \mathcal{P}(\theta)$ (Poisson distributed) latent factors to become detectable. These latent factors $(Y_1, \ldots, Y_N)$ are assumed to be directly active at the beginning of the study, independent and identically distributed (with a proper CDF $F(t)$ independent of N). The population survival and density functions can be shown to be

$$S_p(t|\theta) = \exp\left(-\theta F(t)\right) \ ; \ f_p(t|\theta) = \theta f(t) S_p(t|\theta). \tag{1}$$

As expected, since an unknown proportion of one-child mothers will never become pregnant for a second time, $S_p(.|\theta)$ is an improper survival function, i.e. $S_p(+\infty|\theta) = \exp(-\theta) = P[N = 0] \geq 0$. Note that this quantity coincides with the probability of never having a second child.

Independent baseline covariates, denoted by $\boldsymbol{x}$ (including an intercept) and $\boldsymbol{z}$ (without intercept), enter the model through a log-link on parameter θ and through a Cox model for $F(t)$, respectively:

$$\begin{aligned} \theta(\boldsymbol{x}) &= \exp(\alpha^T \boldsymbol{x}), & (2) \\ F(t|\boldsymbol{z}) &= 1 - S_0(t)^{\exp(\beta^T \boldsymbol{z})}. & (3) \end{aligned}$$

As suggested by Bremhorst and Lambert (2016) in that context, the baseline survival function $S_0(t)$ is modelled through the log-baseline hazard specified as a linear combination of a large number of cubic P-splines with estimation performed in a Bayesian framework (Eilers and Marx, 1996 and Jullion and Lambert, 2007).

This work is motivated by the analysis of data from the German Socio-Economic Panel (Wagner et al., 2007) studying the transition to second birth. In the data, some women/family characteristics, such as the education attainments of the mother and of her partner for example, may vary over time. However, since the latent factors are assumed to turn active only at the beginning of the follow-up, time-varying covariates cannot be used to model the probability of having a second child since it is a one-to-one function of the mean number of latent factors. Chi and Ibrahim (2006) proposed an extension to such covariates by allowing the latent factors to occur at any time during the follow-up. However, the assumptions of their extension lead to an increasing population hazard function which is not realistic in fertility studies.

2 Extension of the promotion time model

For simplicity, the principles underlying the extended promotion time model are explained on an example, pictured in Figure 1, dealing with a sin-

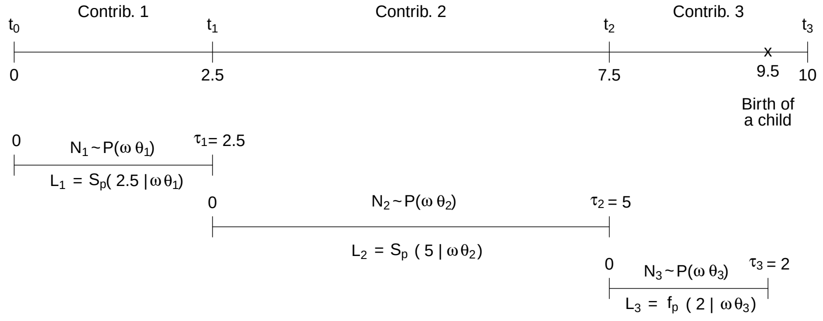


FIGURE 1. Example of contribution to the likelihood for a women with two variations of a single categorical covariate

gle woman and her educational attainment only influencing the pregnancy probability (and not its timing). Assume that a woman gave birth to her first child at 16 when she was still a student at the secondary school. Thus, when entering the study, her level of education (giving the last obtained diploma) is set to *at most primary education* ($x = 0$). As motivated by the promotion time model, it is assumed that she is directly exposed to $N_1 \sim \mathcal{P}(\omega\ \theta(x = 0))$ latent factors, where ω is a random effect to control the unobservable heterogeneity (with density function $g(\omega)$). In the realm of fertility studies, each latent factor could be seen as a potential decisive argument to decide to have an extra child and the time "to detection" as the time required for it to be convincing. After τ_1 years (2.5, say), she graduated from the secondary school without getting a new child. The contribution of this first period to the conditional likelihood is $S_p(\tau_1|\omega, \theta(x = 0))$ and the value of her level of education is updated to *Vocational degree* ($x = 1$). The proposed extension of the promotion time model assumes that when the characteristics of the women change, the N_1 preceding latent factors are replaced by N_2 new ones where $N_2 \sim \mathcal{P}(\omega\ \theta(x = 1))$. Five years ($\tau_2 = t_2 - t_1 = 5$) after secondary school, she gets a university degree, again without getting pregnant. The contribution of the completed period to the conditional likelihood is $S_p(\tau_2|\omega, \theta(x = 1))$. As assumed by our proposed extension of the promotion time model, the latent factors corresponding to this second period are replaced by N_3 new ones with $N_3 \sim \mathcal{P}(\omega\theta(x = 2))$. Two years later ($\tau_3 = 2$), she gave birth to her second child. The contribution to the conditional likelihood of this final event is $f_p(\tau_3|\omega, \theta(x = 2))$. Thus, the contribution to the marginal likelihood of this women is given by

$$L_{\text{birth}} = \int_0^{+\infty} S_p(\tau_1|\omega, \theta(x=0))S_p(\tau_2|\omega, \theta(x=1))f_p(\tau_3|\omega, \theta(x=2))g(\omega)d\omega.$$

Assume, now, that another women gets a university degree τ_1 years after entering the study (i.e. after the birth of her first child). Moreover, assume that τ_2 years later, she left the study without getting a second child. Then, her contribution to the marginal likelihood is given by

$$L_{\text{right cens.}} = \int_0^{+\infty} S_p(\tau_1|\omega, \theta(x=1))S_p(\tau_2|\omega, \theta(x=2))g(\omega)d\omega.$$

The power variance function (PVF) distribution is used for the random effect ω. This flexible distribution family contains the gamma, the inverse gaussian and the positive stable distributions as limiting cases. More informations on the PVF distribution can be found in Duchateau and Janssen (2008, Section 4.5.1).

3 Application

Bremhorst et al. (2016) studied the transition to second and third births when the time-dependent covariates (the level of education, for example) were frozen at the onset of the process (i.e. directly after the birth of the first or the second child, respectively). In this section, results accounting for possible evolution of the mother and father educational levels are reported, for second birth, in Table 1. Note that, for identification purpose, the calendar period was again frozen at the onset of the study. One finds that the probability of having a second child significantly increases with the education level of the mother. Regarding the timing, high educated susceptible women tend to have their second child significantly later than less educated ones. The education level of the partner seems to have no significant influence on the probability or on the timing of second birth. Not surprisingly, a single woman has a rather small probability to have a second child compared to a spouse one. Furthermore, the age at first birth has a significant negative (resp. positive) impact on the probability (resp. the timing) of a second birth. Figure 2 pictures the population baseline hazard function (left) and the baseline hazard function of susceptible women (right) with its 95% pointwise credible interval. Since the population is a mixture of susceptible and non-susceptible women, it was expected that the instantaneous risk of having a second child is smaller for the whole population than for susceptible women. The shapes of the two functions slightly differ. The baseline hazard function for susceptible women peaks 3.5 years after the update of mother and father education levels, then tends to slighlty decrease afterwards. On the other hand, the population hazard function shows a peak sooner (3 years after after the update of the parent education levels) and decreases thereafter. A more detailed explanation of the differences between the population and the susceptible hazard is available in Bremhorst et al. (2016) and will be discussed during the talk.

TABLE 1. Second birth. Estimate of the posterior median and of the posterior standard deviation for each regression parameter.
Signif. codes : * = 0.1 ; ** = 0.05 ; *** = 0.01

	Quantum			Timing		
	Est	sd_{post}		Est	sd_{post}	
Intercept	0.436	0.194		-	-	
Education (ref. Middle)						
Low	-0.474	0.192	**	0.303	0.212	
High	1.044	0.317	***	-0.924	0.334	***
Partner's education (ref. Middle)						
Low	-0.095	0.245		-0.033	0.273	
High	0.210	0.187		0.259	0.206	
No partner	-1.107	0.201	***	0.168	0.231	
Calendar Period (ref : 1998)	0.013	0.010		-0.010	0.012	
Age at first birth (Ref : 28.35 yrs)	-0.132	0.026	***	0.051	0.026	**

Acknowledgments: The authors acknowledge financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract Projet d'Actions de Recherche Concertées (ARC) 11/16-039 of the Communauté française de Belgique, granted by the Acadmie universitaire Louvain. The authors also thank Michaela Kreyenfeld for the data.

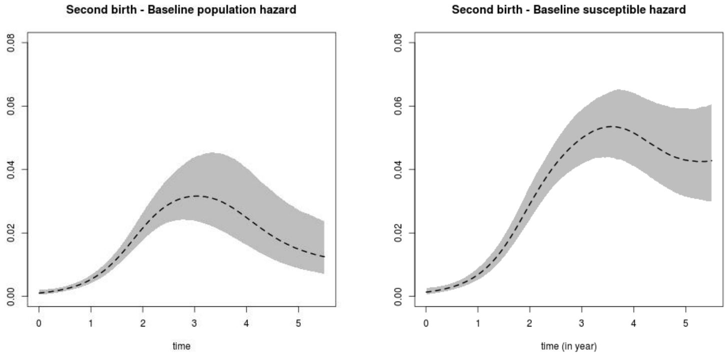


FIGURE 2. Second birth - Fitted baseline population hazard (left) and fitted baseline hazard for suceptible women (right) with 95% pointwise credible intervals.

References

Bremhorst V. and Lambert P. (2016). Flexible estimation in cure survival models using Bayesian P-splines. *Computational Statistics and Data Analysis.* **93**, 270 – 284.

Bremhorst V., Kreyenfeld, M. and Lambert P. (2016). Fertility progression in Germany : An analysis using flexible non parametric cure survival models. *Submitted to Demographic Research.*
Preprint available at http://hdl.handle.net/2078.1/170303

Chen, M.-H., Ibrahim, J.G. and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association.* **94**, 909 – 919.

Chi, Y., and Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics*, **7,** 432 – 445.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model.* New York:Springer.

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

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spacially adaptive Bayesian P-splines models. *Computational Statistics and Data Analysis.* **51**, 2542 – 2558.

Wagner, G. G., Joachim R. F., and Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP): Scope, evolution and enhancements. *Schmollers Jahrbuch.* **127**, 139 – 169.

Trends in digit preference: Mastering a challenge with the penalized composite link model

Carlo G. Camarda[1], Paul H.C. Eilers[2], Jutta Gampe[3]

[1] Institut National d'Études Démographiques, Paris, France
[2] Dept. of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands
[3] Max Planck Institute for Demographic Research, Rostock, Germany

E-mail for correspondence: `carlo-giovanni.camarda@ined.fr`

Abstract: We present a two-dimensional generalization of a penalized composite link model for modelling latent distributions of birth weights with digit preference. We allow weights at multiples of 10 to attract from neighbouring categories, however, different multiples of 10 may attract counts differently. Moreover, we are able to measure improvement in registration of birth weights by modulating the strength of this misreporting pattern over time. Data are taken from Emmerson & Roberts (2013) and they pose major challenges due to their size and sparseness. We achieve a feasible solution by aggregating data when details are secondary, i.e. in the estimation of the smooth latent distribution of birth weights. Smoothness is enforced by a difference penalty on neighbouring coefficients, and both misreporting pattern and its development over time are estimated by iteratively weighted least-squares systems. We provide uncertainty measures for each elements of the model structure by bootstrap.

Keywords: birth weight; composite link model; digit preference; penalized likelihood.

1 Introduction

In neonatal intensive care units the birth weight of a newborn is an important determinant of drug prescriptions. Especially for very small babies accuracy is important. Unfortunately, digit preference (DP), the rounding of weights to multiples of 10 is very common. Emmerson & Roberts (2013) collected 9170 birth weights over a period of 19 years and they presented an analysis. Figure 1 presents proportions of births by last two digits of

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).

their registered birth weights in two different periods. It shows an evident attraction to weights ending with 0 and 5 as well as changes in the pattern over time.

It is thus of interest to quantify developments over time of DP, i.e. to determine how far improved hospital policies have reduced it. Camarda et al. (2009) presented a model for trends in DP, based on the penalized composite link model. The data of Emmerson & Roberts looked like the perfect real-life test bed for our ideas, and luckily we got access to them. The exercise turned out to be more challenging than expected, since we had to deal with a distribution in steps of one gram over a region from 500 to 4500 grams. Here we describe how we came to a workable solution.

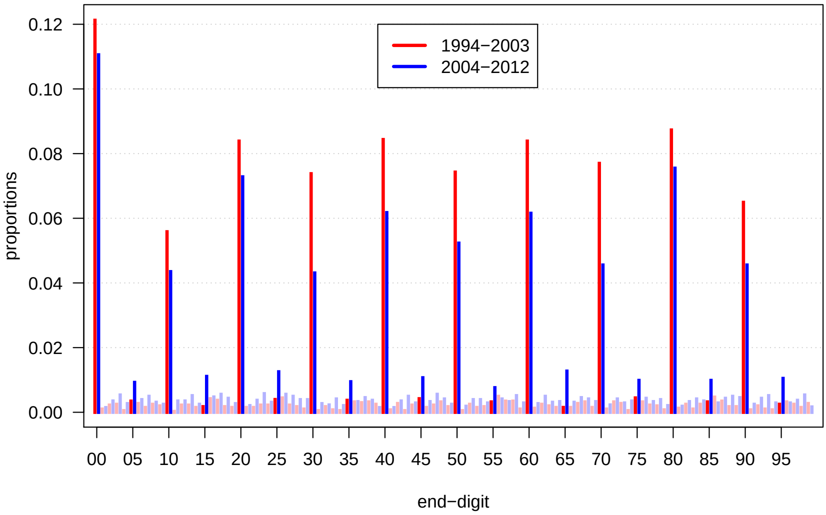


FIGURE 1. Proportion of births by last two-digits of their registered birth weights in two different periods. Darker colors depict end-digits at multiples of 5.

2 The model

The observed data, which we denote by y_{ij}, are the counts of birth weights (index i), ranging from 500 to 4499, in the years 1994 to 2012 (index j). Hence the total number of counts is $4000 \times 19 = m \times n$ (91% of which are zero). The vector $\boldsymbol{y} = (y_{1,1}, \ldots, y_{4000,1}; \ldots; y_{1,19}, \ldots, y_{4000,19})^T$ holds these counts, year after year. The total number of births in each year is given by the vector $\check{\boldsymbol{e}} = (\check{e}_j)$, that is $\check{e}_j = \sum_i y_{ij}$. Likewise we arrange these exposures as vector $\boldsymbol{e} = \mathtt{vec}(\boldsymbol{E})$, where $\boldsymbol{E} = \check{\boldsymbol{e}}\, \mathbf{1}_{1,m}$.

The observed counts arise from a true, but latent distribution (assumed to be smooth), modified by a digit preference mechanism that leads to heaping of counts at preferred end-digits. The observed counts $\boldsymbol{y}$ follow a Poisson distribution $\mathcal{P}(\boldsymbol{\theta})$, where the means $\boldsymbol{\theta} = \boldsymbol{e} \cdot \boldsymbol{\mu}$ incorporate the exposure numbers. The true (latent) distributions are denoted by $\boldsymbol{\gamma} = \texttt{vec}(\gamma_{ij})$, the digit preference is expressed by a misreporting pattern embodied in a matrix $\boldsymbol{C}$ so that the vector $\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma}$. This is a composite link model (CLM, Thompson & Baker, 1981), and the goal is to estimate the misreporting pattern in $\boldsymbol{C}$ and to see whether it became less strong over time.

Following the results reported in Emmerson & Roberts we focus on a model producing heapings at multiples of 10. The following assumptions were made: Counts at, for example, ␣␣20 arise from misreporting of a proportion of counts at ␣␣16 to ␣␣19 as well as ␣␣21 to ␣␣24. The proportion depends on the distance to the target. These proportions are denoted by p_1^{20} (for ␣␣19 and ␣␣21) to p_4^{20} (for ␣␣16 and ␣␣24). In the current version of the model we include that different multiples of 10 have different p_w^d (d for weight decade, ranging from 00 to 90; $w = 1, \ldots, 4$), however, we do not discriminate between, say, 3420 and 3520.

The composition matrix is constructed from $\boldsymbol{C}_0$, which incorporates this misreporting pattern. $\boldsymbol{C}_0$ is a block-diagonal matrix over i,

$$\boldsymbol{C}_0 = \texttt{diag}\left(\ldots, \boldsymbol{C}^{00}, \boldsymbol{C}^{10}, \ldots, \boldsymbol{C}^{d}, \ldots \boldsymbol{C}^{90}, \boldsymbol{C}^{00}, \boldsymbol{C}^{10}, \ldots\right) ,$$

where the superscript denotes the weight decade attracting counts from the neighbouring four categories on both sides. A generic $\boldsymbol{C}^d$ is given by

$$\boldsymbol{C}^d = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -p_4^d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -p_3^d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -p_2^d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -p_1^d & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & p_4^d & p_3^d & p_2^d & p_1^d & \cdot & p_1^d & p_2^d & p_3^d & p_4^d \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -p_1^d & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -p_2^d & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -p_3^d & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -p_4^d \end{bmatrix} .$$

In this way, 40 different misreporting probabilities, 4 for each weight decade, constitute the misreporting pattern.

The misreporting pattern in $\boldsymbol{C}_0$ may, however, vary over the years. This is expressed by a vector $\boldsymbol{g} = (g_j)$ so that the final the composition matrix is

$$\boldsymbol{C} = \boldsymbol{I}_{m \cdot n} + [\texttt{diag}(\boldsymbol{g}) \otimes \boldsymbol{C}_0] ,$$

where $\boldsymbol{I}_{m \cdot n}$ is an identity matrix with mn rows. Smoothness is assumed both for the distributions in $\boldsymbol{\gamma}$ within one year, but also across adjacent years.

The latent distributions in $\boldsymbol{\gamma}$ are estimated by supplementing the CLM with a two-dimensional smoothness penalty (Eilers, 2007). Moreover we use tensor products of B-splines to reduce the size of the system of equations in the iterative re-weighted least-squares algorithm. The latent $\boldsymbol{\gamma}$ are needed at a 1g resolution to estimate the misreporting probabilities, however, reliable estimates of the γ_{ij} can be achieved even when the weights are binned in intervals of length 100g. This only changes the composition matrix to

$$\boldsymbol{C}_G = \boldsymbol{Q}\,\boldsymbol{C}\,,$$

where $\boldsymbol{Q} = \boldsymbol{I}_{n\,K} \otimes \mathbf{1}_{1,100}$, with K denoting the number of 100g-intervals. The elements q_{il} of $\boldsymbol{Q}$ are equal to 1, if weight i is contained in class l, and zero otherwise. Again $\boldsymbol{\mu} = \boldsymbol{C}_G\,\boldsymbol{\gamma}$ and a CLM results. Therewith we still can estimate the γ_{ij} at 1g resolution but are able to reduce the computation time by a factor of 20.

Both the misreporting probabilities in $\boldsymbol{C}_\mathbf{0}$ and the modulating vector $\boldsymbol{g}$ are estimated by iteratively weighted least-squares (WLS) systems. Specifically we approximate $(\boldsymbol{y} - \boldsymbol{\gamma})$ as

$$(\boldsymbol{y} - \boldsymbol{\gamma}) \approx N(\boldsymbol{X}^p\boldsymbol{p}, \mathtt{diag}(\breve{\boldsymbol{\mu}}))$$

and

$$(\boldsymbol{y} - \boldsymbol{\gamma}) \approx N(\boldsymbol{X}^g\boldsymbol{g}, \mathtt{diag}(\breve{\boldsymbol{\mu}}))\,.$$

for fitting p_w^d and g_j, respectively. $\boldsymbol{X}^p$ and $\boldsymbol{X}^g$ are the associated design matrices and $\breve{\boldsymbol{\mu}} = \boldsymbol{C}\boldsymbol{\gamma}$, i.e. expected values at a 1g resolution.

We enforce smoothness of $\boldsymbol{g}$ with a second-order difference penalty. Alternatively each g_j can be estimated independently by solving a system of equations for each j.

The amount of smoothness of the latent surface $\boldsymbol{\gamma}$ as well as of the modulating vector $\boldsymbol{g}$ is determined by minimizing the Akaike Information Criterion. Standard errors for the model components are obtained by bootstrap. We resample from the data 500 times with replacement, then estimate $\hat{\gamma}_{ij}$, $\hat{p}_w^d$ and $\hat{g}_j$. Confidence intervals for the misreporting pattern and modulating vector were derived directly from the estimated values. Pointwise confidence intervals for the γ_{ij} were computed by drawing random Poisson counts from the fitted latent distributions.

3 Results for the birth weight data

When we fit the model to the data of Emmerson & Roberts we obtain the results presented in Figure 2. We analyzed the full data set as well as the subset of low-weight babies below 2500g. Multiples of 100g (end-digits 00) clearly attract more observations than the other weight decades. Misreporting probabilities are higher for categories next to the target decades. They are lowest for the weight decades 10 and 90, while for weights ending with 20 and 80 the p_w^d are noticeably higher. The misreporting pattern for infants with low birth weight is not much different from the pattern for the complete data set.

The strength of this pattern changed over time, and there has been a substantial improvement in the accuracy of birth weight measurements over the period 1994-2012. The improvement was particularly strong for the last years and stronger for the infants with low birth weights.

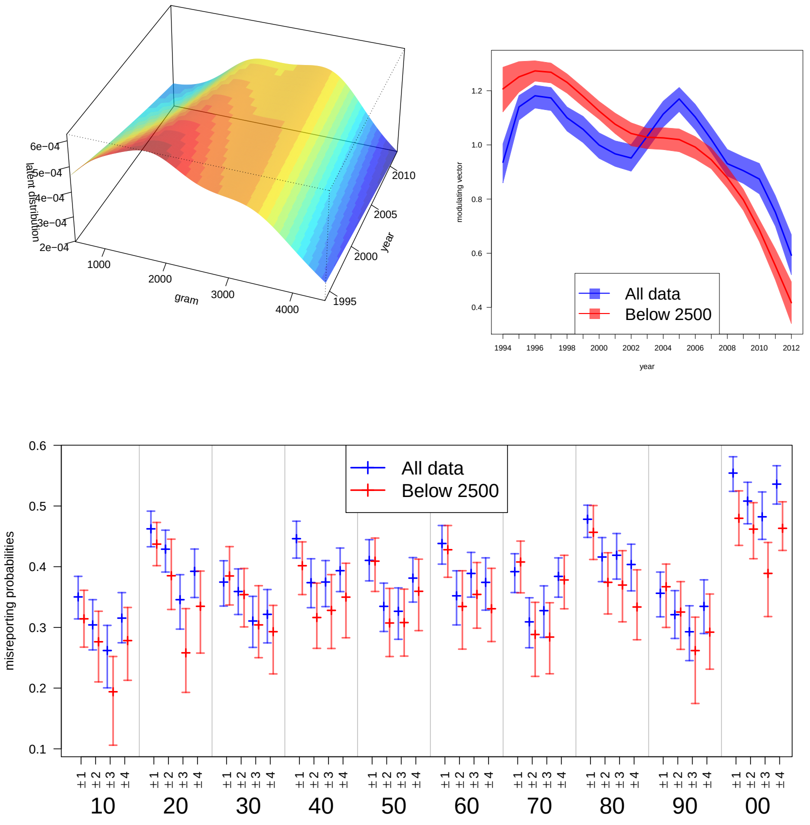


FIGURE 2. Top left: Estimated surface of the true birth weight distributions over time (full data set). Top right: Strength of DP pattern over time, full data and low birth weights (<2500g). The curves represent the estimated scaling vector $\hat{\boldsymbol{g}}$ with 95% confidence intervals. Bottom: Estimated misreporting probabilities p_w^d for weight decades $d = 10, \ldots, 00$ with 95% confidence intervals.

4 Outlook

The estimated misreporting pattern in Figure 2 calls for extension of the model to allow a even more flexible setting.

For instance, the increase of the p_w^d at distance $w = 4$ presumably is due to the fact that we did not incorporate the end-digits 05 in our model. We attempted to generalize our model toward this aspect. We allow the same category to exchange to either the closest digit ending with 0 or ending with 5, e.g. latent counts in 2343 could have been misreported to either 2340 or 2345. Whereas this model worked well on simulated data, first results on birth weights were not satisfactory. Likely the actual data do not contain sufficiently large information to inform the model over this double option and therefore final outcomes favored always digits ending with 0. A possible solution would be to allow end-digits 0 and 5 into the model, but without any overlapping structure in the $\boldsymbol{C}$ matrix.

A more challenging, though flexible approach could be to modulate a single preference pattern for 00, 10, ..., 90 across weight (i.e. multiples of 100) and time. If attainable, this approach would allow to study misreporting patterns over time and subsets of the data without splitting of the weight axis into groups.

We plan to consider both ideas in a future extension.

References

Camarda, C. G., Eilers, P. H. C. and Gampe J. (2009). Modelling trends in digit preference patterns. In J. Booth (Ed.), *Proceedings of the 24th International Workshop on Statistical Modelling*. 81-88.

Eilers, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, **7**, 239–254.

Emmerson A. J. and Roberts S. A. (2013). Rounding of birth weights in a neonatal intensive care unit over 20 years: an analysis of a large cohort study *BMJ Open*, **3**, 1–5.

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

Forecasting and representation of the intensity of an ETAS model: a visualization approach

Angela Carollo[1], Francesco Libasci[2], Giada Adelfio[2], Marcello Chiodi[2], Orietta Nicolis[3]

[1] Max Planck Institute for Demographic Research, Rostock, Germany
[2] Dipartimento di Scienze Economiche, Statistiche e Aziendali, Universitá degli Studi di Palermo, Palermo, Italia
[3] Department of Mathematics and Statistics, Universidad de Valparaíso, Valparaíso, Chile

E-mail for correspondence: carollo@demogr.mpg.de

Abstract: A forecasting approach, based on *Forward Likelihood for Prediction* which aims to forecast the daily intensity of space-time point process, is applied to the seismic sequence of L'Aquila earthquake. Estimates of *total* intensity are represented through the interface between R and Google Earth, and also results of the forecast experiments are shown.

Keywords: Earthquakes; Forward Likelihood for Prediction; ETAS; plotKML; Google Earth.

1 Introduction

Earthquakes are identified by points in space and time through their space-time coordinates. For this reason *point processes* are mostly used to represent and describe earthquakes.

A point process is a random collection of points, each one representing the time and space coordinates of a single event. Analytically, the *conditional intensity function* $\lambda(t, x, y|\mathcal{H}_t)$ uniquely characterizes any space-time point process (Daley and Vere-Jones, 2003). It is defined as the frequency with which events are expected to occur around a particular location in time and space, conditional on the prior history $\mathcal{H}_t = \{(t_i, x_i, y_i, M_i) : t_i < t\}$ of the point process up to time t, where (x_i, y_i) are the spatial coordinates and M_i is the magnitude of the i-th event.

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 *Epidemic Type Aftershock Sequences* (ETAS) model is a point process, introduced for the first time by Ogata in 1988 and extended in 1998. The model represents the intensity of earthquakes of magnitude $M \geq M_0$, where M_0 is the magnitude threshold, in a well defined spatio-time region. The model includes *background* activity with occurrence rate $\mu(x, y)$ that is constant in time. The background activity is described by a homogeneous Poisson process. The model also includes *aftershock* activity represented by a non-stationary Poisson process according to a branching-type structure. Given a seismic catalog $\{(t_i, x_i, y_i, M_i); i = 1, \ldots, n\}$, the space-time conditional intensity function of an ETAS model is written as follows:

$$\lambda(t, x, y|\mathcal{H}_t) = \mu f(x, y) + \sum_{t_j < t} \frac{\kappa_0 e^{(\alpha-\gamma)(m_j - m_0)}}{(t - t_j + c)^p} \left\{ \frac{(x - x_j)^2 + (y - y_j)^2}{e^{\gamma(m_j - m_0)}} + d \right\}^{-q} \tag{1}$$

The total intensity $\lambda(t, x, y|\mathcal{H}_t)$ is obtained as sum of two parts: the background intensity $\mu f(x, y)$ and the triggered intensity, the sum on the right side. In equation (1) κ_0 measures the strength of the aftershock activity, c and p are characteristic parameters of the seismic activity of a given region, α and γ measure the efficiency of an event of given magnitude in generating aftershocks, m_j is the magnitude of the j-th event while m_0 is the threshold magnitude and d and q control for the spatial influence of the mainshock.
The *Forward Likelihood for Prediction* (FLP) is a semi-parametric estimation technique proposed by Chiodi and Adelfio (2011 and 2015) that allows to obtain simultaneous estimates of the *background* and *triggered* intensity components of a branching-type point process, such as the ETAS model. It estimates backround intensity through a weighted gaussian kernel estimator and the triggered intensity through maximum likelihood.
In this paper we aim to estimate the Italian seismic activity and then to represent the estimates interfacing R and Google Earth. Moreover we intend to forecast the daily *total* and *triggered* intensity of the seismic sequence of L'Aquila earthquake, in the context of the *Collaboratory for the Study of Earthquakes Predictability* experiment (Zechar et al, 2010).

1.1 Data

We study the seismic activity of Italy. The data have been collected in ISIDe (Italian Seismic Instrumental and parametric Data-base), provided by the INGV (Istituto Nazionale di Geofisica e Vulcanologia). It gathers the earthquake's parameters integrating data provided in quasi real-time from localization performed by the Italian earthquake surveillance service. The spatio-temporal window is from April 16th, 2005 to November 1st, 2013, in the observed rectangular space area with degrees `35 0'0.00''N` (latitude) `6 16'19.20''E` (longitude) and `47 58'12.00''N` (latitude) `18 57'39.60''E` (longitude), which covers the Italian territory.

Table 1 presents the data structure:

TABLE 1. Data structure

Time	Latitude	Longitude	Depth	Magnitude
$2013/11/01 - 23:41:05.180$	40.600	15.159	3.6	2.2
$2013/11/01 - 23:40:56.300$	42.878	12.906	10.1	1.3
$2013/11/01 - 22:59:26.960$	38.651	15.435	214.5	2.1
$2013/11/01 - 22:51:57.530$	43.351	12.524	8.1	1.2
$2013/11/01 - 21:10:41.200$	43.403	12.571	5.4	0.6

2 Method

2.1 Forecasting the intensity - Forward Likelihood for Prediction

We aim to obtain daily estimates for the probability to observe at least one event of magnitude greater than the *threshold magnitude* m_0 in each of the cells of the testing area. We then present the estimates through a dynamic representation (see next section).

We employ an ETAS model, with conditional intensity function as in equation (1). Background seismicity $\mu f(x, y)$ is estimated through FLP, while parametric components are estimated through maximum likelihood. The two estimation steps are alternated until convergence (Adelfio and Chiodi, 2015).

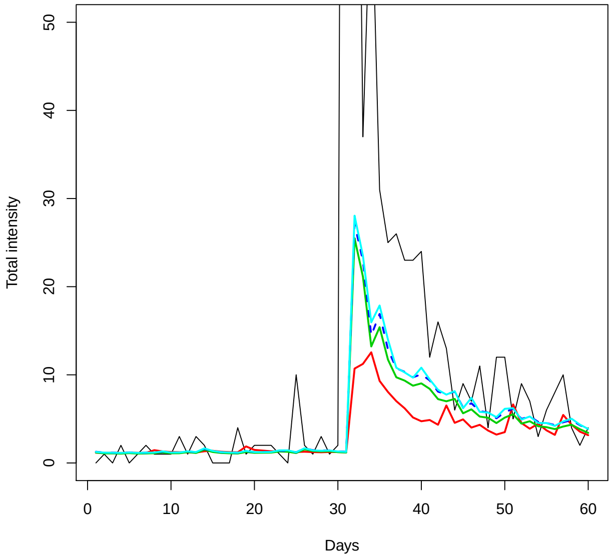
Estimates were obtained using the R-package *etasFLP* (Chiodi and Adelfio, 2015), (R Core Team, 2015). As an example of the performance of the proposed technique, we report a short forecast experiment conducted for the temporal period of the catastrophic seismic sequence that hit the region surrounding the city of *L'Aquila* in April 2009. We first estimate the total, background and triggered intensity using all observations, with magnitude greater than 2.5, from April 16^{th}, 2005 to March 6^{th}, 2009. Then we forecast the daily intensity forward for 60 days.

Staring from the estimated model, we forecast the intensity of the following day. Then we re-estimate the model based on the predicted intensity and we repeat these steps for a period of 60 days.

Results of the experiment are presented in Figure 1. It compares the number of events observed with the number of events predicted by the model, for the whole region of test.

As shown in Figure 1, the blue and light blue lines are closer to the black line, which implies that the forecast obtained combining the only time model and the amplifier effect of magnitude, is the closest one to the real

FIGURE 1. Results of forecasting experiment: 6 March 2009 - 5 May 2009



The black line represents the total observed intensity, the red line represents the spatio-temporal prediction of the total intensity, the green line represents the temporal prediction; the blue (broken) and light blue lines are temporal predictions plus an amplifier effect due to magnitude of past events.

situation. However it is evident lack of fit in the forecasting when considering also the spatial dimension. We suggest to improve the model in terms of spatial adjustment.

2.2 Interactive scientific visualization of spatio-temporal data

When analyzing complex spatio-temporal phenomena, such as earthquakes, it is important to have an appropriate scientific visualization approach. In this paper we use the R-package *plotKML*. It provides methods to represent the most common spatial classes of R and is easily executable with Google Earth. Therewith spatio-temporal data can be visually explored easily.
The advantages of this interface with Google Earth are to visually extend the key concept to audiences not trained in using GIS. Additionally it allows researches to qualitatively interpret the results of spatial analysis and to easily detect the performance of complex spatial models by exploring data in multiple domains.
In this section we provide estimation results of the ETAS model to compare the observed seismic intensity through *etasFLP* in two time periods, 30

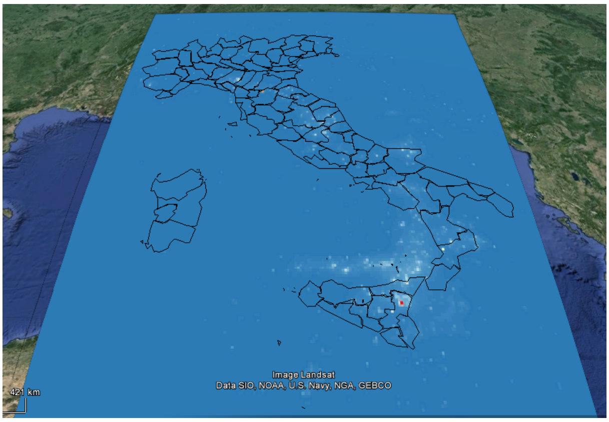


FIGURE 2. Output ETAS model on Google Earth: Total Intensity - March 2009

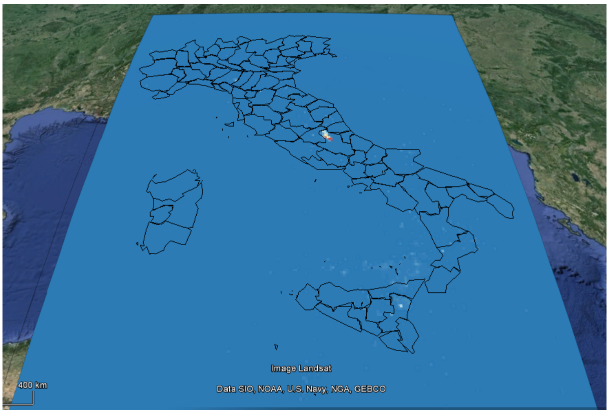


FIGURE 3. Output ETAS model on Google Earth: Total Intensity - May 2009

days before the mainshock of L'Aquila seismic sequence, and 30 days later. We visualize how the Italian seismicity changed due to L'Aquila earthquakes. As an example of interactive visualization (although this tool is more appreciated if used interactively) here we show the result of the ETAS model estimated with all observations from April 16^{th}, 2005 to March 6^{th}, 2009 (Figure 2).

Figure 3 represents, instead, the result of the ETAS model estimated with all observations from April 16^{th}, 2005 to May 6^{th}, 2009 .

By comparing the two figures, it is clear how the seismic sequence of L'Aquila had a strong effect on the Italian seismicity. In March 2009 the total observed intensity was concentrated in the region of the southern Tyrrhenian Sea. After the events of L'Aquila sequence a concentration around the territory of L'Aquila is visible.

3 Future developments and discussion

In this work we first performed a short forecast experiment concerning the sequence of L'Aquila earthquake. Then we proposed a dynamic visualization of space-time data given by the interface of R and Google Earth to represent the results of the forecast experiments obtained by the *etas-FLP* package. Future developments are needed to improve the results of the forecast experiments in terms of spatial adjustment by considering a more flexible version of the ETAS model and its variants accounting for the geophysics structures observed in space (Siino et al., 2016).

References

Adelfio, G. and Chiodi, M. (2015). *Alternated Estimation in semi-parametric space-time branching-type point process with application to seismic catalogs*, Stoch. Environ. Res. Risk. Assess., **29**, 443 – 450.

Chiodi, M. and Adelfio, G. (2011). *Forward likelihood-based predictive approach for space-time point processes*, Environmetrics, **22**, 749 – 757.

Chiodi M. and Adelfio G. (2015). *Package etasFLP*, *Url = https://cran.rproject.org/web/packages/etasFLP/etasFLP.pdf*

Delay, D.J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Process: Volume I: Elementary Theory and Methods. Second Edition.*. Springer.

Hengl T., Roudier P., Beaudette D., Pebesma E., Blaschek M. (2015). *plotKML: Scientific Visualization of Spatio-temporal Data. Package plotKML, URL http://plotKML.R-Forge.R-project.org.*

Ogata, Y. (1998). *Space-time point process models for earthquake occurrences.* Ann. Inst. Statist. Math., **2**, **50**, 379 – 402 .

R Core Team (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Wien, Austria. URL = https://www.R-project.org/.

Siino, M. et al. (2016)*Spatial pattern analysis using hybrid models: an application to the Hellenic seismicity.* Submitted 2016.

Zechar, J.D. et al. (2009) *The Collaboratory for the Study of Earthquak Predictability perspective on computational earthquake science.*, Wiley Interscience.

Signal detection in Event-Related Potentials data

David Causeur[1], Ching-Fan Sheu[2], Emeline Perthame[3]

[1] Agrocampus Ouest and IRMAR (UMR 6625 CNRS), France
[2] National Cheng-Kung University, Taiwan
[3] Laboratoire Jean Kuntzmann, MISTIS, INRIA Grenoble, France

E-mail for correspondence: `david.causeur@agrocampus-ouest.fr`

Abstract: Event-related potentials (ERPs) are recordings of electrical activity along the scalp time-locked to perceptual, motor and cognitive events. Because significant association between ERPs and behavioral variables of interest are often rare and weak, detection of ERP signals poses major challenges to statistical analysis. In this 'rare-and-weak' paradigm, the Higher Criticism method was shown in a number of recent papers to be optimal to determine signal detection threshold.
However, ERP time dependence exhibits a block pattern suggesting strong local and long-range autocorrelation components which violates the mild dependence assumption under which signal detection can be achieved efficiently. In high throughput settings, a variety of decorrelation approaches have been developed to counter those detrimental effects of dependence. The presentation first highlights the impact of dependence in terms of instability of signal detection by Higher Criticism Thresholding. A second objective is to revisit the decorrelation issue using a flexible factor modeling for the covariance. The present method, and variants introducing penalized estimation of the inverse covariance of the process of test statistics, are compared to recent other decorrelation approaches either based on a shrinkage estimation of the inverse covariance or on its Cholesky decomposition.

Keywords: Correlated noise; Event-Related Potentials; High dimension; Higher Criticism; Signal detection.

1 Scientific Background

Event-related potentials (ERPs) are recordings of electrical activity along the scalp time-locked to perceptual, motor and cognitive events. Such high-throughput instrumental data provide high temporal resolution to chart

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 time course of mental processes. With the routine collection of massive amounts of data from ERP studies, researchers must face the challenge of signal detection, which shall guarantee a low false positive error rate while maintaining sufficient power. How to achieve this objective for ERP data exhibiting arbitrarily strong temporal dependence is the focus of the present paper.

In the 'Rare and Weak' (RW) paradigm introduced by Donoho and Jin (2004), signal detection is based on a $m-$vector T of test statistics $T = (T_{t_1}, \dots, T_{t_m})'$ where m is the number of time frames, for the collection of corresponding null hypotheses H_{0,t_i} of no association between the ERP measured at time t_i and the response variable. The RW setup is defined in Donoho and Jin (2004) as the following sparse normal mixture model for T: for all t,

$$T_t \sim (1-\varepsilon)\mathcal{N}(0,1) + \varepsilon\mathcal{N}(\delta,1),$$

where the mixing parameter $0 \leq \varepsilon \leq 1$ is the proportion of non-null features and $\delta \geq 0$ is the signal amplitude. Note that the normality assumption introduced above holds for most ERP studies in which the tests for the association between the ERPs and the response variable is handled by t-tests for the significance of a single parameter. The alternative parameterization $\beta_\varepsilon = -\log(\varepsilon)/\log(m),\ r_\delta = (\delta^2/2)/log(m)$ is often preferred because it maps both the sparsity parameter β_ε and the amplitude parameter r_δ into $[0;1]$, if we observe that the expectation of the maximum test statistics under the null is bounded by $\sqrt{2\log m}$. Sparsity of the signal is characterized by $1/2 \leq \beta_\varepsilon \leq 1$ and weakness by $r_\delta < 1$. Many ERP studies fall into this situation of a rare and weak signal.

In the former RW paradigm, Donoho and Jin (2004) demonstrates that a detection method called Higher Criticism Thresholding, which is based on a distance between the empirical probability distribution function of the p-values and the uniform null distribution, achieves the theoretically optimal decision limits. As reported in Causeur *et al.* (2012), the pronounced auto-correlation observed in ERP data can however induce a long-range regularity for the test statistics, resulting in spuriously low p-values outside of the support of the signal, which in turn can affect the control of type-I error rate of signal detection procedures. Equivalently, Hall and Jin (2010) reports that the theoretical detection bounds derived in the RW framework are markedly modified by a strong dependence among the test statistics. Therefore, Hall and Jin (2010) proposes to extend the RW framework as follows:

$$T = \delta + T_0,$$

where $T_0 \sim \mathcal{N}(0;\Sigma)$. If U is the inverse of the Cholesky factorization of Σ, namely $U\Sigma U' = I$, Hall and Jin (2010) introduces the so-called innovated HCT (iHCT) as the HCT procedure applied on the uncorrelated vector of

innovations $UT = U\delta + UT_0$ and shows that iHCT restores the effectiveness of the HCT procedure in situations of strong dependence.
As in Perthame *et al.* (2015) and Sheu *et al.* (2016), we propose an alternative approach of innovated HCT based on a flexible factor model for Σ. The complex dependence pattern observed in the correlation structure of test statistics derived from ERP data can indeed be well approximated using the factor decomposition of Σ, often with a moderate number of factors. Moreover, it provides simple and efficient algebraic tools to derive the decorrelation matrix operator $\Sigma^{-1/2}$. A Cyclic-Coordinate Descent (CCD) algorithm is also presented for a sparse penalized estimation of $\Sigma^{-1/2}$.

2 Materials and Methods

In ERP studies, perhaps the most commonly used experimental task is the oddball paradigm. In this paradigm, typically two classes of stimuli are presented, one occurring frequently and the other occurring infrequently. The subject is required to distinguish between the two stimuli and to respond to the rare stimuli. In an auditory ERP study performed at Kaohshung Medical University in Taiwan, the task uses two pure tones of 500 Hz and 1,000 Hz. The former is presented 120 out of 150 trials, whereas the latter is presented only for 30 trials. At given electrode locations on the scalp, ERP waveforms were obtained from each of the two tone conditions. Many studies have demonstrated that an ERP waveform across the parieto-central area of the skull is usually observed around 300 ms (the so-called P300 component) and is larger after the target event. The question to be addressed is whether it is possible to detect a significant difference between the mean ERP curves observed for the two stimuli. This verification is of fundamental importance if the P300 component is to be considered as an electrophysiological marker for further assessment of psychiatric and neurological disorders.

The t-test process of no difference between the two conditions along time shows a strong regularity which is not consistent with the expected profile of a process of independently distributed Student variables. This suggests a strong time-dependence among tests, which is known to affect the joint null distribution of test statistics. We propose the following RW framework for the $m-$vector of test statistics:

$$T = \delta + T_0,$$

where $T_0 \sim \mathcal{N}(0; \Sigma = \Psi + BB')$, where Ψ is a $m \times m$ diagonal matrix of specific variances whose diagonal elements ψ_t^2 are in $[0; 1]$ and B is a $m \times q$ matrix of factor loadings, with, for all t, $||b_t = (b_{t,1}, \ldots, b_{t,q})'||^2 = \sum_{l=1}^{q} b_{tl}^2 = 1 - \psi_t^2$. Note that the above parameterization provides a closed-form expression for the decorrelation matrix $\Sigma^{-1/2}$, which can be estimated either using the EM algorithm presented in Sheu *et al.* (2016) or an alternative penalized ML estimation of Σ^{-1} which leads to a sparse estimate.

3 Results and partial conclusion

Some variants of HCT procedures are compared hereafter, including the method proposed by Hall and Jin (2010) based on the Cholesky decomposition of Σ (iHCT for innovated HCT), the correlation-adjusted t-tests introduced by Ahdesmäki and Strimmer (2010), based on a James-Stein Shrinkage estimator and the Factor-innovated HCT (F-iHCT) method presented above, taking advantage of a factor structure for Σ. The comparison is both based on the application of the HCT and iHCT procedures to the auditory oddball ERP dataset introduced above and on intensive simulations under various dependence patterns. We particularly focus on two criteria: the prediction performance based on the selected features and the number of selected features. Only partial simulation results are reported here, demonstrating that an innovated HCT procedure based on a factor decomposition of Σ^{-1} shows desirable properties in a simulation scenario which mimics the auditory oddball ERP data introduced above.
1,000 datasets with dimensions 30×800 are generated according to a multivariate normal distribution. Both the correlation structure and the within-condition variances are estimated from the auditory oddball ERP data introduced in section 2. Each dataset is split into two balanced groups. The normal distribution has expectation zero for the first 15 subjects (group 1) and the expectation for the 15 last subjects (group 2) is a waveform with various amplitudes and non-null features in $[150ms, 200ms]$. $1,000$ training datasets are generated for each signal strength. Eight corresponding testing data of size 1000×800 with two balanced groups are also generated according to the same simulation plan for a prediction purpose. The RW model parameters for this simulation plan are $\varepsilon_T = 12\%$ and $A_T = \sqrt{2rlog(T)}$ with r taking 8 equally distributed values in $[0.004; 0.688]$. According to the RW setup, the present combination of r and β characterizes a not very sparse signal, with a weak to large strength.
As in Donoho and Jin (2008), the variable selection step by different versions of HCT is followed by a supervised classification on the subset of selected variables. Four methods are compared in this simulation study:

- *Standard HCT*: variable selection by standard HCT on raw p-values, classification by Naives Bayes (see Bickel and Levina, 2004);

- *CAT-scores*: variable selection by HCT on decorrelated test statistics using a shrinkage estimator of the whitening matrix (see Ahdesmäki and Strimmer, 2010), classification by diagonal Shrinkage Discriminant Analysis (SDA, see Ahdesmäki and Strimmer, 2010);

- *F-iHCT*: variable selection by Factor-innovated HCT, classification by conditional Bayes classifier (proposed by Perthame *et al.*, 2015);
- *AFA*: variable selection by standard HCT performed on p-values adjusted for effects of latent factors as returned by the AFA (see Sheu *et al.*, 2016) procedure, classification by conditional Bayes classifier (see Perthame *et al.*, 2015).

For all the methods described above, the proportion of signal recovery, called precision, the false discovery rate (FDR), the number of selected features and the prediction error rate are computed. For all datasets, variable selection and estimation of classification rule are performed on training data (including the optimization of meta-parameters) and prediction error is computed on testing data.

Figure 1 shows that selection by CAT-scores appears to be the most efficient to catch weak signals, with both the smallest FDR and the largest precision for small amplitudes of signal. Even if CAT-scores does not achieve the best performance for large signal strengths, the FDR, precision and number of selected variables are remarkably stable. Standard HCT seems robust to dependence as the method performs well in term of FDR but its precision is small regarding methods based on decorrelation. Moreover, the number of selected variables is also small, which suggests that HCT is conservative under dependence. Lastly, classification by Naive Bayes fails as the error rates are the largest for weak to moderate strengths of signal. Variable selection and classification procedures based on the factor model assumption (AFA and F-iHCT) provide the best results both in terms of false positive, recovery of the signal and prediction error. FDR turns out to be small for moderate to high signal strengths and a correct power of signal identification is achieved.

References

Ahdesmäki, M and Strimmer, K. (2010) Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Annals of Applied Statistics*, vol. **4**, pp. 503 – 519.

Bickel, P.J. and Levina, E. (2004) Some theory for Fisher's Linear Discriminant function, naive Bayes, and some alternatives when there are many more variables than observations. *Bernoulli*, vol. **10**, 6, pp. 989-1010.

Causeur, D., Chu, M.C., Hsieh, S. and Sheu, C.F. (2012) A factor-adjusted multiple testing procedure for ERP data analysis. *Behavior Research Methods*, vol. **44**, pp. 635 – 643.

Donoho, D. and Jin. J. (2008) Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, vol. **105**, 39, pp. 14790-14795.

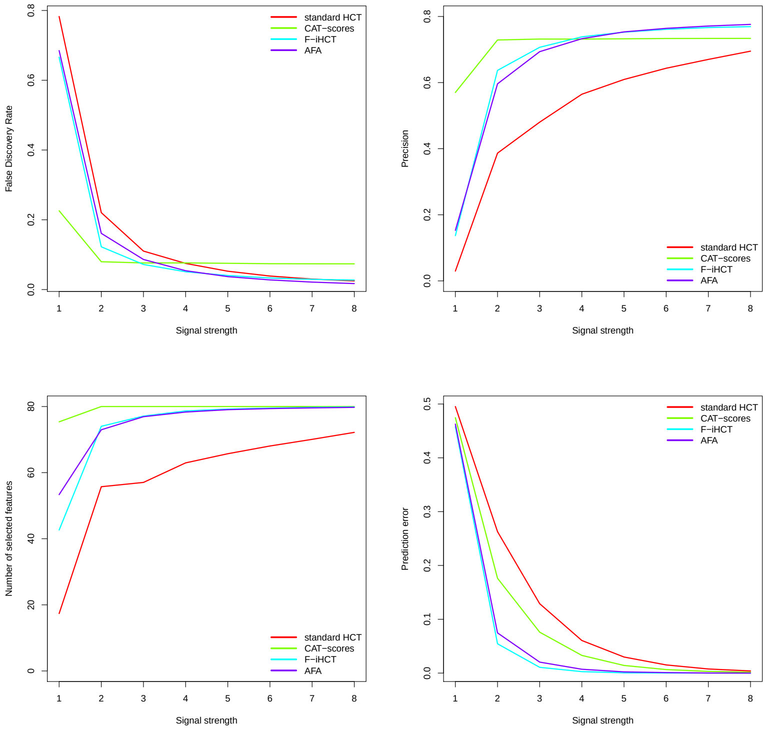


FIGURE 1. Results of the simulation study depending on signal strength: False Discovery Rate (top left), Precision (top right), Number of selected features (bottom left), Prediction error (bottom right).

Donoho, D and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, vol. **32**, no. 3, pp. 962 – 994.

Hall, P. and Jin, J. (2010) Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, vol. **38**, no. 3, pp. 1686 – 1732.

Perthame, E., Friguet, C. and Causeur, D. (2015) Stability of feature selection in classification issues for high-dimensional correlated data. *Statistics and computing*, DOI 10.1007/s11222-015-9569-2.

Sheu, C.F, Perthame, E., Lee, Y.S. and Causeur, D. (2016) Accounting for time dependence in large-scale multiple testing of event-related potential data. *Annals of Applied Statistics.* Vol. **10**, 1, 219-245.

The induced smoothed LASSO

Giovanna Cilluffo[1,2], Salvatore Fasola[1,2], Vito M.R. Muggeo[1], Stefania La Grutta[2]

[1] Department of Economical, Business and Statistical Sciences, University of Palermo (Italy),
[2] Institute of Biomedicine and Molecular Immunology A Monroy (IBIM) - National Research Council (CNR) of Palermo (Italy)

E-mail for correspondence: giovanna.cilluffo@unipa.it

Abstract: We propose a new lasso-type estimator of regression coefficients for regression models. Our proposal relies on the recent idea of induced smoothing and leads to estimators with sampling distribution somewhat close to the Normal one, regardless of their true value, along with the corresponding reliable covariance matrix. As a consequence inference (e.g. p-values) may be carried out relatively easily. We present results from some simulation experiments.

Keywords: Induced smoothing; LASSO; Standard Error.

1 Introduction

Regression models are widely used and well-established statistical tool in many fields. They allow to estimate the covariate effect by returning point estimates and (reliable) standard errors to compute confidence intervals and p-values. However high dimensional regression models pose some issues connected to the complexity of the model and/or to the presence of uninformative variables. The Least Absolutes Shrinkage and Selection Operator (LASSO) represents a very elegant and relatively widespread solution to carry out variable selection and parameter estimation simultaneously. While point estimation can be performed quite straightforwardly, a possible current limitation is computation of standard errors. Tibshirani (1996) proposed to use as approximation $\sum|\beta_j| \approx \sum \beta_j^2/|\beta_j|$, but this approach returns zero standard error when the corresponding point estimate equals zero. Also, the bootstrap is far from being helpful as it is inconsistent (Kyung et al., 2010). Osborne et al. (2010) derived a formula for covariance matrix which ensures positive standard errors for all coefficient estimates. Beside of computation of standard errors, an additional issue comes from the sampling distribution of the regression coefficients which is not Normal when at least one coefficient is equal to zero. Typically, the sampling distribution with null parameter has positive probability mass at zero (Knight and Fu, 2000: Pötscher and Leeb, 2009; Kyung *et al.*, 2010). In the Robert Tibshirani discussion paper at annual conference of RSS 2010, Peter Bühlmann discusses: *The issue of assigning uncertainty and variability in high dimensional statistical inference deserves further research. For example, questions about power are largely unanswered.*

This paper focuses on setting up of a lasso estimator which allows reliable computation of covariance matrix and standard errors. Our proposal relies on the recent idea of Induced Smoothed (IS) wherein the unsmooth estimating functions are replaced by the naturally smoothed counterparts. Section 2 describes the idea of IS and Section 3 reports some simulation evidence.

2 Methods

The idea of natural smoothing was introduced by Brown and Wang (2005) to deal with unsmooth estimating equations $U(\beta)$, say. Assuming a multinormal distribution for $\hat{\beta}$, i.e. $V^{-1/2}(\hat{\beta}-\beta) \sim z$, the smoothed estimating function is obtained via

$$\widetilde{U}(\beta) = E_z[U(\beta + V^{1/2}z)], \tag{1}$$

where $E_z[\ \cdot\ ]$ represents expectation over $z \sim N(0, I_p)$, standard multinormal random perturbations. $\widetilde{U}$ is smooth, thus the slope matrix $\tilde{U}' = \frac{\partial}{\partial\beta}\widetilde{U}(\beta)$ exists and the usual sandwich formula applies to compute the covariance matrix of estimator $\hat{\beta}$,

$$V = \tilde{U}'^{-1}\ \mathcal{I}\ \tilde{U}'^{-1} \tag{2}$$

where $\mathcal{I} = \mathrm{cov}(U)$ is the usual information matrix.

Clearly $\widetilde{U}$ requires V (see (1)), and in turn V needs $\widetilde{U}$ (see (2)). Hence an iterative procedure is called for, alternating computation of $\widetilde{U}$ and V. More specifically:

1. fix initial guesses for V and β; in particular, we set $V^{(0)} = I_J/n$
2. compute $\widetilde{U}(\beta^{(0)})$ according to (1)
3. compute $\widetilde{V}(\beta^{(0)})$ according to (2)
4. update $\hat{\beta}$ via a Newton-Raphson step $\beta^{(0)} - \widetilde{U}'^{-1}\widetilde{U}$

5. set $\beta^{(0)} = \hat{\beta}$ and repeat steps 2-4 till convergence

We propose to apply the natural smoothing to the lasso estimating equations. Figure 1 portrays the effect of the induced smoothing on the lasso penalty $\sum_{j=1}^{2} |\beta_j|$ for two different standard error estimates.

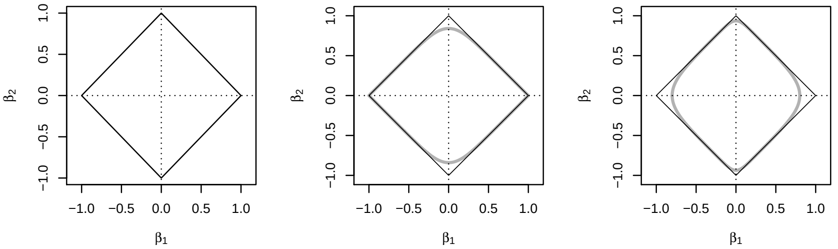


FIGURE 1. Contrasting the lasso penalty (diamond, black thin lines) with the induced smoothed counterpart (thick gray lines). The amount of smoothing at kink depends on the variance of the corresponding estimator and it is determined automatically by data.

3 Simulation Evidence

We compare Lasso and IS-Lasso estimators in a (limited) simulation study. Standard errors for the Lasso are computed according to Osborne et al. (2010), while standard errors for IS-Lasso come from the sandwich formula described in method section. We generate 1000 replications from a linear regression model with sample size $n = 50$ and number of parameters $p = 20$ (only 4/20 informative covariates); the tuning parameter λ is fixed at 4 throughout replicates.

Table 1 shows summary of sampling distributions for the first 8 coefficients. In terms of bias and variance of estimators, results are quite similar between Lasso and IS-Lasso: both naive and IS lasso estimators are biased if the corresponding coefficient is nonzero, however SE from IS-Lasso always appear to be more reliable, especially when the true parameter is zero: on the other hand, the Osborne approach heavily overestimates uncertainty in the estimate.

We assess performance of the IS-LASSO approach in hypothesis testing problems. We assume a standard Normal null distribution for the Wald statistic based on IS-Lasso estimator $\hat{\beta}_j/\mathrm{SE}(\hat{\beta}_j)$; note, under H_0, $\hat{\beta}_j$ is unbiased, the SE correctly estimates the estimator variability, and therefore good performance is expected. We compare the IS-Lasso Wald statistic with two recent proposals: i) covTest (Lockhart et al., 2014) which does

TABLE 1. Mean and standard deviation of the sampling distributions in the simulation study, obtained with LASSO and IS-LASSO. $\overline{\text{SE}}$ is the average of the standard errors. $\lambda = 4$ at each replicate and true values $\beta = (0.7, 0.5, 1, 0.8, 0, 0, \ldots, 0)^T$, only the first 8 reported.

	mean		*sd*		$\overline{\text{SE}}$	
	Lasso	IS	Lasso	IS	Lasso	IS
β_1	0.617	0.613	0.146	0.148	0.168	0.148
β_2	0.443	0.444	0.153	0.152	0.165	0.145
β_3	0.927	0.921	0.183	0.186	0.199	0.178
β_4	0.761	0.758	0.138	0.141	0.146	0.134
β_5	-0.001	-0.004	0.080	0.095	0.178	0.087
β_6	-0.003	-0.005	0.089	0.100	0.151	0.089
β_7	-0.019	-0.026	0.089	0.101	0.161	0.091
β_8	-0.021	-0.026	0.094	0.106	0.177	0.090

not depend on the lambda value since it is based on the difference between the fitted values of the models with and without the relevant covariate entering into active set at proper lambda value; ii) selective or post-selection inference (Lee et al., 2013), namely a conditional (to the selected model) approach using the truncated Normal distribution for the parameter estimators with a fixed lambda value. We generate 1000 replications from a linear regression model with sample size $n = 50$ and number of parameters $p = 20$ (3 true nonzero coefficients); at each replicate the tuning parameter λ is optimized through 5 fold Cross Validation. Table 2 shows the results of the first 10 coefficients only. When the true coefficient is different from zero, the Wald statistic based on IS-Lasso estimator leads to a (remarkably) more powerful test with respect to covTest and the conditional approaches; moreover under the null hypothesis (i.e. when the true coefficient is zero), the size is quite close to the nominal 0.05 level.

4 Application

We use the IS-Lasso to the well-known Prostate Cancer dataset analyzed in Tibshirani (1996). There are $n = 97$ subjects, $p = 8$ covariates (see Table 3) and the response variable is the log of prostrate specific antigen. Table 3 reports the estimates from naive Lasso and IS-Lasso along with p-values returned by covTest for the naive Lasso estimates and the Wald statistic for IS-Lasso estimates. It is noteworthy that svi variable results to be statistically significant for IS-Lasso but not for covTest.

5 Conclusions

We have presented a smooth approximation for the Lasso regression. It is based on the recent idea of induced smoothing (IS) and leads to estimators having a sampling distribution closer to the Normal one; moreover the method allows to gain reliable standard errors which correctly quantify the estimator variance, even for estimator of zero coefficient. We have compared the IS-Lasso in hypothesis testing assuming a standard Normal distribution for the traditional Wald statistic. Interestingly, results get better than competitors in terms of power. There are several sides to be further investigated, for instance the IS-Lasso does not return *exactly* zero estimates, although the resulting p-value can be used to discard or not the covariates.

References

Brown, B.M., and Wang, Y.G. (2005). Standard errors and covariance matrices for smoothed rank estimators. *Biometrika*, **92**, 149 – 158.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics*, **32**, 407–489.

Kyung, M., Gilly, J., Ghoshz, M.,and Casella, G. (2010) Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis*, **5**, 1–44.

Knight K. and Fu W. (2000) Asymptotics for lasso-type estimators. *Annals of Statistics*, **28**, 1356-1378.

TABLE 2. Power functions (at 5% level) of the statistic tests (IS-Lasso, covTest and postSel, see text for details). At each replicate the optimal lambda employed by IS and postSel has been obtained via cross validation. True coefficients $\beta = (0.5, 0.4, -0.4, 0, 0 \ldots, 0)^{\mathrm{T}}$, only the first 10 reported.

IS	covTest	postSel
0.823	0.539	0.579
0.622	0.183	0.285
0.568	0.134	0.199
0.039	0.030	0.016
0.051	0.000	0.026
0.059	0.002	0.019
0.060	0.003	0.051
0.045	0.000	0.024
0.037	0.002	0.024
0.042	0.002	0.026

TABLE 3. Estimates and p-values of Prostate Cancer data. Estimates come from naive Lasso and IS-Lasso, p-values come from covTest for Lasso and Wald test for IS-Lasso.

	Lasso		IS-Lasso	
covariate	Est	p-value	Est	p-value
lcavol	0.594	0.0000	0.587	0.0000
lweight	0.228	0.0003	0.231	0.0043
age	−0.046	0.2779	−0.067	0.2423
lbph	0.079	0.4472	0.090	0.1626
svi	0.240	0.2118	0.239	0.0056
lcp	0.000	0.2499	−0.001	0.9424
gleason	0.000	0.9703	0.027	0.4817
pgg45	0.059	0.8778	0.059	0.2704

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). On the LASSO and its dual. *Journal of Computational Graphical Statistics*, **9**, 319–337.

Pötscher, B. M., and H. Leeb (2009) On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, **10**, 2065–2082.

Tibshirani (1996) Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, **58**, 267–288.

Zou, H., Hastie, T., and Tibshirani, R. (2007) On the 'degrees of freedom' of the lasso. *Annals of Statistics*, **35**, 2173–2192.

Lee, J.D., Sun, D.L., Sun, Y., & Taylor, J. E. (2016) Exact post-selection inference, with application to the lasso. *Annals of Statistics*, **44**, 907–927.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014) A significance test for the lasso. *Annals of statistics*, **42**, 413.

Ordinal Responses with Latent Uncertainty: a Bivariate Model

Roberto Colombi[1], Sabrina Giordano[2], Anna Gottard[3], Maria Iannario[4]

[1] Department of Management, Information and Production Engineering, University of Bergamo, Italy
[2] Department of Economics, Statistics and Finance, University of Calabria, Italy
[3] Department of Statistics, Computer Science, Applications, University of Florence, Italy
[4] Department of Political Sciences, University of Naples Federico II, Italy

E-mail for correspondence: sabrina.giordano@unical.it

Abstract:
In responding to a rating question, an individual may give an answer according to his/her knowledge (*feeling*) or to his/her level of indecision (*uncertainty*) and such behavior can be modelled through a latent variable. In this paper, two latent binary variables are assumed to rule the answers to two rating questions. The joint distribution of the ordinal variables, describing the two responses, is modelled by a mixture of four components corresponding to the cases of uncertainty in both the answers, feeling in both the answers and uncertainty in only one of them.

Keywords: Ordinal data; Mixture models; Latent variables, Uncertainty.

1 Introduction

According to the CUB models (D'Elia and Piccolo, 2005), both individual *feeling* (personal perception of an item) and *uncertainty* (intrinsic indecision) determine the choice among ordered alternatives of a rating question. In this context, the distribution of the answer to a single item is a mixture of feeling and uncertainty components, where the first is modelled by a (shifted) Binomial distribution, the latter by a discrete Uniform distribution. Tutz et al. (2014), as an alternative, use a general ordinal response model (Tutz, 2012, Agresti, 2010) as feeling component. Their proposal is extended in our approach to the multivariate case to model the association among the respondent's ratings on several items taking into account the dependence of the answers on subject's features. In this paper, we illustrate the idea in the bivariate case and use real data to prove the usefulness of our model.

2 A mixture model for bivariate ordinal responses

Let R_1 and R_2 be two ordinal variables, with support $\{1, 2, \ldots, m_1\}$ and $\{1, 2, \ldots, m_2\}$, respectively. We assume the existence of two latent variables, C_l, $l = 1, 2$, such that the respondent answers the $l - th$ question according to his/her feeling when $C_l = 1$ or his/her uncertainty when $C_l = 0$. Moreover, the ordinal variable R_l is assumed to depend only on the latent variable C_l, $l = 1, 2$. Consequently, we suppose that:

1) $R_1 \perp\!\!\!\perp C_2 | C_1$;

2) $R_2 \perp\!\!\!\perp C_1 | C_2$;

3) given $C_l = 0$, R_l has a Uniform distribution, $l = 1, 2$.

Under these assumptions, the marginal distribution of R_l, $l = 1, 2$, is:

$$P(R_l = r_l) \;=\; \pi_l \, P(R_l = r_l \mid C_l = 1) \;+\; (1 - \pi_l) \, v_l(r_l), \quad r_l = 1, 2, \ldots, m_l, \tag{1}$$

where $\pi_l = P(C_l = 1)$ and $v_l(r_l)$ is the discrete Uniform distribution over $\{1, 2, \ldots, m_l\}$, as considered by Tutz et al. (2014) and D'Elia and Piccolo (2005) in the univariate case. Moreover, to specify the joint distribution of the two responses, it is reasonable to assume that R_1 and R_2 are independent whenever $C_1 \cdot C_2 = 0$. This is equivalently expressed by the conditions:

4) $R_1 \perp\!\!\!\perp R_2 | C_1 = 0, C_2 = 0$;

5) $R_1 \perp\!\!\!\perp R_2 | C_1 = 0, C_2 = 1$;

6) $R_1 \perp\!\!\!\perp R_2 | C_1 = 1, C_2 = 0$.

If $\pi_{ij} = P(C_1 = i, C_2 = j)$, $i = 0, 1$, $j = 0, 1$, are the joint probabilities of the latent variables, conditions $1 - 6$ imply that the joint distribution of (R_1, R_2) is a mixture of four conditional distributions:

$$\begin{array}{rcl} p(R_1 = r_1, R_2 = r_2) & = & \pi_{00} v_1(r_1) v_2(r_2) \\ & + & \pi_{01} v_1(r_1) P(R_2 = r_2 \mid C_2 = 1) \\ & + & \pi_{10} P(R_1 = r_1 \mid C_1 = 1) v_2(r_2) \\ & + & \pi_{11} p(R_1 = r_1, R_2 = r_2 \mid C_1 = 1, C_2 = 1). \end{array} \tag{2}$$

3 A parametrization for the distribution of $(\boldsymbol{R_1}, \boldsymbol{R_2})$

Two marginal logits and a log odds ratio are used for specifying π_{ij}, $i = 0, 1$, $j = 0, 1$ in order to derive a simple parametric expression for $\pi_l = P(C_l = 1)$, involved in the marginal probabilities of R_l given in (1).
The vectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, including $(m_1 - 1)$ and $(m_2 - 1)$ logits (local, global, continuation, reverse continuation), are used to parameterize $P(R_l = r_l \mid C_l = 1)$, $l = 1, 2$. In addition, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, together with the $(m_1-1)(m_2-1)$ log odds ratios (local, global, continuation, reverse continuation) of the vector $\boldsymbol{\beta}_{12}$, parameterize the joint distribution $P(R_1 = r_1, R_2 = r_2 \mid C_1 = 1, C_2 = 1)$.
This parameterization includes $m_1 m_2 - 1 + 3$ parameters, so that identifiability constraints are necessary. For instance, the presence of covariates may serve this need. Given a set of covariates $\mathcal{X}$, the vectors of logits $\boldsymbol{\beta}_1^{\mathcal{X}}$, $\boldsymbol{\beta}_2^{\mathcal{X}}$, and of log odds ratios $\boldsymbol{\beta}_{12}^{\mathcal{X}}$ are defined for every configuration x of the covariates in $\mathcal{X}$. Heterogeneity can be modelled through:

$$\boldsymbol{\beta}_1^{\mathcal{X}} = \mathbf{X}_1\boldsymbol{\alpha}_1, \quad \boldsymbol{\beta}_2^{\mathcal{X}} = \mathbf{X}_2\boldsymbol{\alpha}_2, \quad \boldsymbol{\beta}_{12}^{\mathcal{X}} = \mathbf{X}_{12}\boldsymbol{\alpha}_{12}$$

The entries of the matrices $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_{12}$ are functions of the covariates of $\mathcal{X}$. Constraints on the parameters of the above linear models can solve the identifiability issue. For instance, given a covariate X with J categories, we denote the marginal logits and log odds ratios by $\beta_1^{X=j}(i_1)$, $\beta_2^{X=j}(i_2)$ and $\beta_{12}^{X=j}(i_1, i_2)$, $i_1 = 1, ..., m_1 - 1, i_2 = 1, ..., m_2 - 1, j = 1, 2, ..., J$. Hence, we consider the following proportional logit models:

$$\begin{aligned} \beta_1^{X=j}(i_1) &= \alpha_1(i_1) + \delta_{1j}, \qquad j = 1, 2, ..., J, \\ \beta_2^{X=j}(i_2) &= \alpha_2(i_2) + \delta_{2j}, \qquad j = 1, 2, ..., J, \\ \delta_{11} &= \delta_{21} = 0, \end{aligned}$$

together with the hypothesis of homogeneous association (Kateri, 2014) $\beta_{12}^{X=j}(i_1, i_2) = \alpha_{12}(i_1, i_2)$, $i_1 = 1, ..., m_1 - 1, i_2 = 1, ..., m_2 - 1, j = 1, 2, ..., J$. For this model, the number of parameters $(m_1 m_2 - 1) + 2(J - 1) + 3$ is less than the number $J(m_1 m_2 - 1)$ of independent observable frequencies, so that the necessary condition for identifiability is always satisfied.

4 An example

As an example, we apply the proposed models to data from the 6^{th} round of the European Social Survey, collected in 2012. Data are available at http://ess.nsd.uib.no/ess/round6/. The sample size is 3189.
Among the variables recorded by the survey, we want to study *Happiness* and *life Satisfaction*, here measured on a 5 points scale.
Happiness and *life Satisfaction* are ambiguous concepts (Feldman, 2008), thus it seems intriguing to evaluate in which extent people are able to provide an accurate evaluation of these two variables.

TABLE 1. The fitted models are specified under different hypotheses of: covariates effect on the observed responses (column 1) and on latent variables (column 2), and the independence/no independence of the two latent variables (column 2). The number of parameters (n.par.) and the BIC value are reported in the last 2 columns.

Effect of covariates on observed var.	Hypothesis on latent var.	n.par.	BIC
C, G on H, S	indipendence	30	13489.98
C on H, S	indipendence	28	**13475.64**
G on H, S	indipendence	28	13512.10
C, G on H, S	no indipendence	31	13496.09
C on H, S	no indipendence	29	13481.71
G on H, S	no indipendence	29	13520.27
C on H, S	no indipendence, G	31	13497.27
C on H, S	no indipendence, C	31	13489.60
C on H, S	indipendence, G	30	13490.71
C on H, S	indipendence, C	30	13480.13

In particular, taking into account the uncertainty component in the answers, we investigate whether there is a relationship between *Happiness* and *life Satisfaction* and if the *Gender* (G) and the *Country* (UK, Italy) of residence (C) have any influence on these aspects of the life and/or on the uncertainty in expressing an opinion on them.
Some models, aimed at giving a first response to these questions, are reported in Table 1.
The model specified by the hypotheses of row 2 in Table 1 shows the best fit (lowest BIC value) to the analyzed data.
Thus, the results seem to be coherent with the hypotheses that people tend to give an answer at random to the questions about *Happiness* and *Satisfaction* independently, but this uncertain behavior does not vary remarkably between men or women, or for the English and Italian people (hypotheses in rows 7-10). While the *Country* where the respondents live is a relevant factor in discriminating their level of *Happiness* and *Satisfaction* for the life, showing the British unexpectedly happier and more satisfied than the Italians.

Acknowledgments: This research has been partially supported by the frame of SHAPE project in STAR Programme (CUP E68C13000020003) at University of Naples Federico II, financially supported by UniNA and Compagnia di San Paolo, and by project PRIN-SURWEY (grant 2012F42NS8,

Italy).

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, 2nd edition. J. Wiley & Sons, Hoboken.

Feldman, F. (2008). Whole life satisfaction concepts of happiness. *Theoria*, **74**, 219 – 238.

D'Elia, A. and Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, 49, 917 – 934.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. Birkhuser, New York.

Tutz, G (2012). *Regression for Categorical Data.* Cambridge University Press.

Tutz, G., Schneider, M., Iannario, M. and Piccolo, D. (2014). Mixture models for ordinal responses to account for uncertainty of choice. *Technical Report*, **175**, Department of Statistics LMU.

Optimal Maintenance Policy under Imperfect Repair: A Case Study of Off-Road Engines

Enrico A. Colosimo[1], Maria Luíza G. de Toledo[2], Marta A. Freitas[1], Gustavo L. Gilardoni[3]

[1] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil,
[2] Escola Nacional de Estatística, Rio de Janeiro, Brazil
[3] Universidade de Brasília, Brasília, Brazil

E-mail for correspondence: enricoc@est.ufmg.br

Abstract: In the repairable systems literature one can find a great number of papers that propose maintenance policies under the assumption of minimal repair after each failure (such repair leaves the system in the same condition as it was just before the failure - *as bad as old*). This paper derives a statistical procedure to estimate the optimal Preventive Maintenance (PM) periodic policy, under the following extended two assumptions: (1) perfect repair at each PM action (i.e., the system returns to the *as good as new* state) and (2) imperfect system repair after each failure (the system returns to an intermediate state between *as bad as old* and *as good as new*). This work was motivated by a real situation involving off-road engines maintenance.

Keywords: imperfect repair; minimal repair; ARA models; power law process.

1 Introduction

Off-road trucks are designed to operate in harsh conditions and, consequently, they are used in every conceivable industry where rough terrain goes with the territory (mining, drilling, etc.). In mining companies particularly, off-road trucks are used to transport high production between the front mining and the cell homogenization and, for that matter, the good performance of this equipment is essential to the financial health of this kind of business. As a matter of fact, since the treatment plant needs to work with a constant supply of ore, it is crucial to bring back an off-road truck to its operational state as soon as a failure occurs. These failures cost millions of dollars to the global mining industry directly (replacement and

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016.

corrective repair actions) and indirectly through the inconveniences caused by those failures, such as loss of production, security risks, and reallocation of maintenance resources. Due to the high cost of these systems, one great concern is the implementation of good maintenance policies in order to prolong their life and reduce any expenses generated by the occurrence of unexpected failures.

The approach proposed in this paper takes into account the effect of repair actions implemented after each failure (repair efficiency) in order to determine the optimal periodicity of PMs. The PM actions are supposed here to be perfect.

Models for imperfect repair have already been presented in the literature. However inference procedures for the quantities of interest have not yet been fully studied. In the present paper, statistical methods, including the likelihood function, Monte Carlo simulation, and bootstrap resample methods, are used in order to: (1) estimate the degree of efficiency of repair and (2) obtain the optimal preventive maintenance check points that minimize the expected total cost.

2 Imperfect repair and maintenance cost

From a modelling point of view, $\{N(t)\}_{t\geq 0}$ (where $N(t)$ denotes the number of observed failures up to time t) is a stochastic point process, with mean function $\Phi(t) = E[N(t)]$ and failure intensity function

$$\lambda(t) = \lim_{\delta t \to 0} \frac{P(N(t+\delta t) - N(t) = 1 | \Im_t^-)}{\delta t}, \quad \forall t \geq 0 \tag{1}$$

where $\Im_t^-$ represents the history up to time t (informally, one could think of $\Im_t^-$ as the information provided by the failure times $0 < t_1 < \cdots < t_{N(t)} < t$).

Assume that PM is performed every τ units of time. The expected maintenance cost per unit time for the system is given by (Gilardoni and Colosimo, 2007)

$$C(\tau) = \frac{C_{PM} + C_{IR}E[N(\tau)]}{\tau}, \quad \tau > 0, \tag{2}$$

where C_{PM} and C_{IR} are fixed costs of PM and imperfect repair, respectively. The objective here is to find τ which minimizes $C(\tau)$. Since $E[N(t)] = \Phi(t)$, deriving Equation (2) in respect to τ and equating to zero we obtain

$$D(\tau) = \tau\phi(\tau) - \Phi(\tau) = \frac{C_{PM}}{C_{IR}}, \tag{3}$$

where $\phi(\tau) = \frac{d}{d\tau}\Phi(\tau)$ is the ROCOF function for the system.

3 Statistical inference

A likelihood function appropriate to model this process considers the observed k systems failure times $t_{ij}; i = 1, \ldots, k; j = 1, \ldots, n_i$, and is given by

$$L(\mu) = \prod_{i=1}^{k_1} [f(t_{i,1}, \ldots, t_{i,n_i} | N(t_i^*) = n_i) P(N(t_i^*) = n_i)].$$

ARA$_1$ (Kijima, 1989) model, $\lambda(t) = \lambda_R(t - (1 - \theta)T_{N(t)})$, where T_n is a random variable representing the real age of the system at the n^{th} failure and θ, the repair efficiency parameter. Power Law Process (PLP - Crow, 1974) is used as the initial intensity function $\lambda_R(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}, \quad \eta, \beta, t > 0$.

The steps of the proposed method are illustrated using the PLP but it can be applied to any other parametric form chosen for the initial intensity. The steps are described below:

- Step 1: Maximum Likelihood estimation of the model parameters: $\hat{\beta}$, $\hat{\eta}$ (PLP parameters) and $\hat{\theta}$ (repair efficiency).
- Step 2: Estimation of the mean function: Monte Carlo simulation of failure histories and calculation of the MCF.
- Step 3: Estimation of the optimal periodicity τ. In order to solve Equation (3), it is necessary to find estimates for the functions $\phi(t)$ and MCF. In Step 2, the MCF was used as an estimate for $\Phi(t)$. However, the MCF is a step function, so its derivative is almost everywhere zero, and an estimate for $\phi(t)$ cannot be directly obtained from this. So, we use here the nonparametric estimate given by the right derivative of the Greatest Convex Minorant (GCM) (Boswell, 1966):

4 Off-Road Engines Maintenance Data Revisited

In this section, we return to the situation described in Section 1, i.e, the off-road engines maintenance. Table 1 exhibits the point and interval (95% confidence intervals based on Normal approximation) MLEs for parameters. Under both models, $\hat{\beta}$ is consistently greater than 1, indicating that the engines tend to fail more frequently with age. Also, it is noteworthy that the estimated value for θ, and the corresponding confidence interval suggest that the repair actions taken after failures are neither minimal ($\theta = 1$) nor perfect ($\theta = 0$). According to the intensity function for ARA$_1$, the inclusion of θ is what differentiates the IR from the MR modelling. So, $\hat{\theta} \neq 1$ means that $T_{N(t)}$ has to be taken into consideration in the intensity function after each failure. In other words, this is a strong evidence that the repair effects must be considered in the problem, supporting the use of ARA$_1$ instead of the MR model.

TABLE 1. Point and bootstrap interval (95% confidence level) estimates for PLP (β, η) and effect of repair (θ) parameters, and values of the maximum of the log-likelihood function ($\hat{l}$), AIC e BIC under minimal and imperfect repair models for off-road data.

Model:	Minimal Repair	Imperfect Repair (ARA$_1$)
$\hat{\beta}$	$2.125(1.916; 2.357)$	$2.458(2.185; 2.765)$
$\hat{\eta}$	$16,715(15,604; 17,905)$	$15,582(14,601; 16,628)$
$\hat{\theta}$	-	$0.471(0.330; 0.672)$
$\hat{l}$	-2126.74	-2118.59
AIC	4257.48	4243.18
BIC	4264.15	4253.19

5 Acknowlegments

This work was partially supported by FAPEMIG grant.

References

Boswell, M.T. (1966). Estimating and Testing Trend in a Stochastic Process of Poisson Type. *The Annals of Mathematical Statistics*, **37**, 1564 – 1573.

Crow, L.R. (1974). Reliability Analysis of Complex Systems. In: *F. Proschan and J. Serfling (editors), Reliability and Biometry: Statistical Analysis of Lifelength*, SIAM, Philadelphia, PA, 379– 140.

Doyen, L., Gaudoin, O. (2004). Classes of Imperfect Repair Models Based on Reduction of Failure Intensity or Virtual Age. *Reliability Engineering & System Safety*, **84**, 45 – 56 .

Gilardoni, G.L., Colosimo, E.A. (2007). Optimal Maintenance Time for Repairable Systems. *Journal of Quality Technology*, **39**, 48– 53

Kijima, M. (1989). Some Results for Repairable Systems with General Repair. *Journal of Applied Probability*, **26**, 89– 102.

Bias-corrected z-tests for regression models

Claudia Di Caterina[1], Ioannis Kosmidis[2]

[1] University of Padua, Italy
[2] University College London, United Kingdom

E-mail for correspondence: dicaterina@stat.unipd.it

Abstract: In regression settings the effect of a covariate, accounting for all the others, on the dependent variable is typically tested by using a z-statistic. Under regularity conditions on the model and assuming the null hypothesis holds, the associated Wald pivot is asymptotically normally distributed. However, its finite-sample distribution can be far from Gaussian when the sample size is small or moderate relative to the dimension of the global parameter. In this work, asymptotic bias correction of the Wald z-statistic is proposed as a means to improve the accuracy of first-order inference for the regression coefficients.

Keywords: Asymptotic bias correction; First-order asymptotic inference; Generalized linear model; Wald z-statistic.

1 Wald pivots in regression settings

Consider a standard regression framework with p covariates and p corresponding scalar unknown coefficients $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, all assumed to be identifiable. For the sake of generality, suppose the model involves also a vector $\lambda = (\lambda_1, \ldots, \lambda_q)^{\mathrm{T}}$ of nuisance parameters (e.g. dispersion/precision). Furthermore, denote by $\theta = (\beta^{\mathrm{T}}, \lambda^{\mathrm{T}})^{\mathrm{T}}$ the full parameter vector and by $i(\theta)$ the expected information matrix, with inverse having the following blocking structure:

$$\{i(\theta)\}^{-1} = \begin{bmatrix} i^{\beta\beta}(\theta) & i^{\beta\lambda}(\theta) \\ i^{\lambda\beta}(\theta) & i^{\lambda\lambda}(\theta) \end{bmatrix} .$$

The most widespread way to investigate the impact of a particular regressor on the response variable, taking into consideration all the other covariates, is via z-tests, mainly because of their simplicity and the facility of their implementation. Formally, the procedure to test the hypothesis $H_0 : \beta_j = \beta_{0j}\,(j = 1, \ldots, p)$ while accounting for the remaining model parameters, consists of comparing the quantiles of the standard normal distribution to the value of the Wald z-statistic

$$T^j(\hat{\theta};\beta_{0j}) = \frac{\hat{\beta}_j - \beta_{0j}}{\sqrt{\nu_j(\hat{\theta})}}, \tag{1}$$

where $\hat{\theta} = (\hat{\beta}^{\mathrm{T}}, \hat{\lambda}^{\mathrm{T}})^{\mathrm{T}}$ is the full maximum likelihood (ML) estimate and $\nu_j(\theta)$ indicates the (j,j)th element of the matrix $i^{\beta\beta}(\theta)$. Indeed, under mild regularity conditions on the model and if the null hypothesis holds, usual likelihood-based asymptotic arguments (see, for instance, Pace and Salvan, 1997, Chapter 4) can be employed to show that the large-sample distribution of (1) is standard normal. Such limiting result may however lead to statistical tests of poor performance when the sample size is moderate and/or small with respect to p.

2 Location-adjusted z-statistics

In order to make the normal approximation more reliable in these situations, one can attempt to reduce the asymptotic bias of $T^j(\hat{\theta};\beta_{0j})$ by similar arguments as in Efron (1975, Remark 11) for correcting the bias of the ML estimator. The key step is to consider $T^j(\theta;\beta_{0j})$ as a non-singular transformation of the parameter vector θ. Then, the ML estimator of $T^j(\theta;\beta_{0j})$ is simply $T^j(\hat{\theta};\beta_{0j})$.

Let $z^j(\theta;\beta_{0j}) = T^j(\theta;\beta_{0j})/\sqrt{n}$, so that it has the same asymptotic order as θ. Assuming the function $z^j(\cdot;\beta_{0j})$ is at least three times differentiable and starting from Remark 3 of Section 4.3 in Kosmidis and Firth (2010), it may be shown that, given the consistency of the ML estimator and adopting the Einstein summation convention, $z^j(\hat{\theta};\beta_{0j})$ admits the asymptotic bias expansion

$$\begin{aligned} E_\theta\big[z^j(\hat{\theta};\beta_{0j}) - z^j(\theta;\beta_{0j})\big] &= B_z(\theta;\beta_{0j}) + O(n^{-2}) \\ &= z^j_r(\theta;\beta_{0j})B^r(\theta) + \frac{1}{2} z^j_{rs}(\theta;\beta_{0j})\kappa^{r,s}(\theta) + O(n^{-2}), \end{aligned}$$

where $B^r(\theta)$ is such that $E_\theta\big[\hat{\theta}^r - \theta^r\big] = B^r(\theta) + O(n^{-2})$, $z^j_r(\theta;\beta_{0j})$ and $z^j_{rs}(\theta;\beta_{0j})$ are the gradient and the hessian, respectively, of $z^j(\cdot;\beta_{0j})$ evaluated at θ and $\kappa^{r,s}(\theta)$ is the (r,s)th element of $\{i(\theta)\}^{-1}(r,s = 1,\ldots,p+q)$. It is then possible to estimate the first-order bias of $z^j(\hat{\theta};\beta_{0j})$ by $B_z(\hat{\theta};\beta_{0j})$ and derive the location-adjusted Wald z-statistic

$$T^{j,*}(\hat{\theta};\beta_{0j}) = T^j(\hat{\theta};\beta_{0j}) - \sqrt{n}B_z(\hat{\theta};\beta_{0j}). \tag{2}$$

In the following, we will refer to the test based on $T^{j,*}(\hat{\theta};\beta_{0j})$ as the bias-corrected z-test.

3 Illustrations on the performance of the location-adjusted z-statistic

3.1 Inference about the logarithm of an exponential mean

Consider independent random variables $Y_1, \dots, Y_n$, each having an exponential distribution with mean $E(Y_i) = e^{\beta}$ $(i = 1, \dots, n)$. Under this assumption, we have that $T(\hat{\beta}; \beta_0) = -\sqrt{n}(\log \bar{y} + \beta_0)$ and $T^*(\hat{\beta}; \beta_0) = T(\hat{\beta}; \beta_0) - 1/(2\sqrt{n})$. The performance of the location-adjusted z-statistic in such framework is remarkable: when testing $\beta = \beta_0$ against different alternatives (two- or one-sided) at a level α, the test based on $T^*(\hat{\beta}; \beta_0)$ has size which is closer to the nominal level than the one based on $T(\hat{\beta}; \beta_0)$ for any value of β_0, n and α. This can be easily checked by comparing the null distributions of both statistics with the standard normal. Figure 1 shows such a comparison for $n = 5$: it is clear that the distribution of $T^*(\hat{\beta}; \beta_0)$ is closer to that of a $N(0, 1)$.

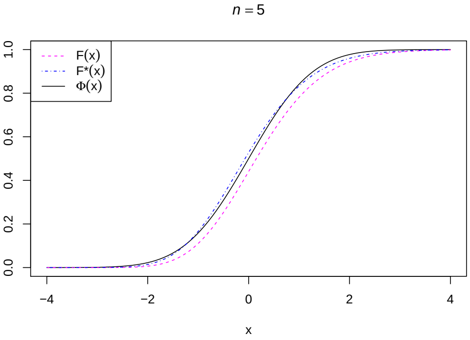


FIGURE 1. Comparison of the null cumulative distribution functions $F(x) = P_{\beta_0}(T(\hat{\beta}; \beta_0) \leq x)$ and $F^*(x) = P_{\beta_0}(T^*(\hat{\beta}; \beta_0) \leq x), \forall \beta_0 \in \mathbb{R}$, to the standard normal distribution $\Phi(x)$, in the case $n = 5$.

3.2 Gamma regression

A simulation study can be set up as follows: starting from $n = 8$, for every ith unit, covariates x_i and z_i $(i = 1, \dots, n)$ are generated as independent realizations of a $N(1, 1)$. The corresponding dependent variable y_i in each of the 2000 simulated datasets is then obtained by random generation from a Gamma distribution with shape parameter $\nu = 2$ and rate $\lambda_i = \nu/\mu_i$, where $\mu_i = \exp(\tilde{\beta}_1 + \tilde{\beta}_2 x_i + \tilde{\beta}_3 z_i)$ with $\tilde{\beta}_1 = 1, \tilde{\beta}_2 = 1, \tilde{\beta}_3 = 2$. On every sample, statistics (1) and (2) are used to test $H_0 : \beta_j = \tilde{\beta}_j$ $(j = 1, \dots, 4)$ versus the two-tailed alternative, taking into account the other regressors, so that

empirical sizes of the corresponding tests can be estimated at nominal levels $\alpha = 0.01, 0.05$. This procedure is repeated for $n = 16, 32, 64, 128, 256$, but instead of generating a new set of regressors every time, the same x_i and z_i $(i = 1, \dots, 8)$ are used for adjacent blocks of 8 units.
Partial results of the study are available in Table 1, which also displays estimated sizes for tests based on the score statistic s, profile likelihood ratio statistic r and its modification r^* (Brazzale et al., 2007, Chapter 8). As can be seen, for small values of n (especially $n = 8, 16$) the bias-corrected z-test has empirical sizes much closer to α than its standard version, and does also better than the test associated with the likelihood ratio statistic. Among the first-order tests, s appears to have the best general performance, even comparable to the second-order accurate r^*. Not surprisingly, such discrepancies tend to disappear as n increases.

TABLE 1. Empirical sizes at nominal levels $\alpha = 0.01, 0.05$ of the tests related to T^j, its adjusted version $T^{j,*}$, the score statistic s^j, the likelihood ratio statistic r^j and its modification $r^{j,*}$ $(j = 1, 2, 3)$ in the Gamma regression model, estimated by a study based on 2000 simulated datasets of size $n = 8, 16, 32, 64$.

	$\alpha = 0.01$					$\alpha = 0.05$				
$n = 8$	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$
$j = 1$	0.109	0.040	0.015	0.051	0.014	0.178	0.096	0.074	0.135	0.060
$j = 2$	0.113	0.048	0.004	0.062	0.015	0.199	0.105	0.068	0.147	0.072
$j = 3$	0.107	0.046	0.005	0.057	0.016	0.200	0.099	0.066	0.144	0.064
$n = 16$	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$
$j = 1$	0.043	0.026	0.015	0.027	0.015	0.107	0.068	0.062	0.087	0.057
$j = 2$	0.046	0.020	0.008	0.023	0.009	0.112	0.071	0.057	0.083	0.057
$j = 3$	0.039	0.020	0.006	0.024	0.011	0.116	0.068	0.051	0.081	0.051
$n = 32$	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$
$j = 1$	0.023	0.013	0.010	0.014	0.010	0.072	0.058	0.051	0.061	0.054
$j = 2$	0.022	0.014	0.008	0.013	0.011	0.076	0.059	0.048	0.061	0.049
$j = 3$	0.024	0.017	0.011	0.018	0.013	0.074	0.056	0.043	0.061	0.045
$n = 64$	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$
$j = 1$	0.020	0.016	0.013	0.014	0.018	0.071	0.063	0.058	0.062	0.065
$j = 2$	0.014	0.013	0.009	0.011	0.012	0.061	0.052	0.049	0.056	0.050
$j = 3$	0.014	0.011	0.008	0.010	0.009	0.063	0.056	0.050	0.058	0.053

A more realistic scenario is considered in the next simulation experiment, involving the *clotting* dataset in McCullagh and Nelder (1989, p. 300). The data record observations of $n = 18$ mean clotting times in seconds of blood (y) for nine percentage concentrations of normal plasma (x) and two lots of clotting agent ($z = 1, 2$). Assuming $Y_1, \dots, Y_n$ are independent Gamma random variables with mean $\mu_i = \exp(\beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 x_i * z_i)$ ($i = 1, \dots, n$), a Gamma regression model with log link is fitted to the data and 2000 samples of size n are simulated under the ML fit. To test $H_0 : \beta_j = \hat{\beta}_j$, where $\hat{\beta}_j$ is the estimate of β_j obtained from the original ML fit ($j = 1, \dots, 4$), the standard z-statistic and its location-adjusted version, the score statistic, the likelihood ratio one and its modification are computed on every dataset.

Table 2 reports empirical sizes of the associated two-tailed tests at nominal levels $\alpha = 0.01, 0.05$. For each regression coefficient, the bias-corrected z-test results in sizes closer to α than (1). Moreover, the normal Q-Q plots in Figure 2 illustrate how the adjustment in location improves the standard normal approximation to the null distribution of the z-statistic in case of testing $H_0 : \beta_4 = \hat{\beta}_4$. From Table 2 we can also see that (2) does not perform as well as r^*, but performs always better than the likelihood ratio statistic and better than s when the nominal level is 0.05.

TABLE 2. Empirical sizes at nominal levels $\alpha = 0.01, 0.05$ of the tests related to T^j, $T^{j,*}$, s^j, r^j and $r^{j,*}$ ($j = 1, \dots, 4$) in the Gamma regression model. The figures are based on a simulation study with 2000 replications.

	$\alpha = 0.01$					$\alpha = 0.05$				
	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$	T^j	$T^{j,*}$	s^j	r^j	$r^{j,*}$
$j = 1$	0.036	0.016	0.006	0.023	0.006	0.106	0.059	0.070	0.089	0.051
$j = 2$	0.039	0.015	0.010	0.023	0.008	0.108	0.060	0.071	0.088	0.052
$j = 3$	0.035	0.015	0.010	0.024	0.008	0.092	0.056	0.064	0.076	0.046
$j = 4$	0.034	0.014	0.010	0.019	0.008	0.105	0.054	0.067	0.082	0.045

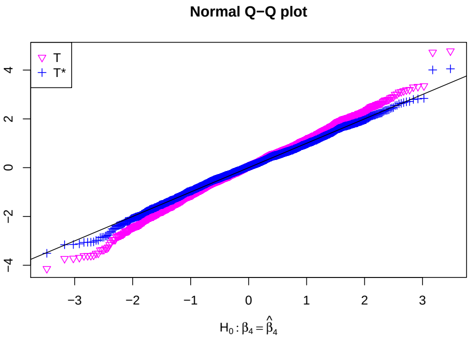


FIGURE 2. Normal Q-Q plots based on 2000 values of $T^4(\hat{\theta}; \hat{\beta}_4)$ and $T^{4,*}(\hat{\theta}; \hat{\beta}_4)$ computed under the null hypothesis $H_0 : \beta_4 = \hat{\beta}_4$.

4 Remarks and future work

Results obtained in the Gamma regression setting suggest that location adjustment of the z-statistic can appreciably improve testing at no extra computational cost, especially for small sample sizes. In addition, some initial experiments not shown here give evidence that adopting a parametric bootstrap to correct the scale of (2) can lead to a performance even comparable to second-order tests. Obviously, this improvement has to be evaluated taking also into account the increase in computational effort. Applications to other generalized linear models need undoubtedly to be explored, since the availability of explicit formulae for the first-order asymptotic bias of the ML estimator in such framework (Cordeiro and McCullagh, 1991) makes the adjustment in location of the Wald z-statistic appealing in terms of required calculations. A specific model we plan to consider is the Cox proportional hazards model (see, for example, Ma, 2008, Chapter 4, for a different method to correct the bias of Cox estimates).

References

Brazzale, A.R., Davison, A.C., and Reid, N. (2007). *Applied Asymptotics Case Studies in Small-Sample Statistics.* Cambridge University Press.

Cordeiro, G.M. and McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society B*, **3**, 629 – 643.

Efron, B. (1975). *The Annals of Statistics*, **3**, 1189 – 1217.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag.

Kosmidis, I. and Firth, D. (2010). A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, **4**, 1097 – 1112.

Ma, L. (2008). *Improved Methods for the Analysis of Time-to-Event Data.* PhD dissertation, Temple University. Ann Arbor: ProQuest/UMI.

McCullagh, P. and Nelder, J.A (1989). *Generalized Linear Models.* London: Chapman and Hall.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective.* London: World Scientific.

Fast stable relative risk regression using an overparameterised EM algorithm

Mark W. Donoghoe[1,2], Ian C. Marschner[1,2]

[1] Macquarie University, Sydney, Australia
[2] NHMRC Clinical Trials Centre, University of Sydney, Australia

E-mail for correspondence: mark.donoghoe@mq.edu.au

Abstract: Relative risk regression models can be fitted using a log-link binomial GLM, however standard algorithms can suffer convergence problems. Combinatorial EM (CEM) algorithms that provide stable convergence can be computationally intensive, particularly for large models. We present a new approach using an EM algorithm with an overparameterised model that retains the stability of the CEM algorithm but greatly reduces computing time. This is demonstrated with a small example in which modified Fisher scoring fails to converge to the MLE, and a bootstrap analysis of data from a clinical trial in heart attack patients.

Keywords: Binomial regression; EM algorithm; Relative risk.

1 Introduction

In biostatistics, regression models for relative risks can be fitted using a binomial generalised linear model (GLM) with a log link function. These are often preferred to a standard logistic regression analysis, but the standard method used to fit such models in many statistical packages — Fisher scoring — can fail to converge to the maximum likelihood estimate (MLE). Modifications such as step-halving can help avoid such issues in some cases, but cannot guarantee convergence in general.

Stable methods based on the EM algorithm have been presented for fitting such models, but they can be computationally expensive when the number of covariates is large. We present a novel method that can greatly reduce computational time, without compromising the stability of the algorithm.

2 Combinatorial EM algorithm

Consider a log-link binomial GLM with A categorical covariates (each with k_a levels) and B linear covariates; that is $Y_i \sim \text{Bin}(N_i, \exp\{\Lambda(\mathbf{u}_i, \mathbf{v}_i; \boldsymbol{\theta})\})$, where

$$\Lambda(\mathbf{u}_i, \mathbf{v}_i; \boldsymbol{\theta}) = \alpha_0 + \sum_{a=1}^{A} \sum_{k=2}^{k_a} \alpha_a(k) 1(u_{ia} = k) + \sum_{b=1}^{B} \beta_b v_{ib},$$

and we have imposed the identifiability constraint $\alpha_a(1) = 0$. The total number of parameters that need to be estimated in this model is $1 + \sum_a (k_a - 1) + B$.

The exponentiated parameters in this model represent adjusted relative risks. Specifically, $\exp(\alpha_a(k))$ is the relative risk associated with a change in the a^{th} categorical covariate from level 1 to level k, keeping all other covariates fixed. Likewise $\exp(\beta_b)$ is the relative risk associated with a one-unit increase in the b^{th} continuous covariate, with all else staying constant.

As described by Marschner and Gillett (2012), the model can be viewed as a missing data problem, in which the complete data are independent latent Bernoulli random variables underlying the observed binomial outcomes. That is,

$$Y_i = \sum_{j=1}^{N_i} \mathcal{Y}_{ij} \qquad \text{where} \qquad \mathcal{Y}_{ij} = \mathcal{A}_{ij0} \times \prod_{a=1}^{A} \mathcal{A}_{ija} \times \prod_{b=1}^{B} \prod_{k=1}^{v_{ib}} \mathcal{B}_{ijbk} \tag{1}$$

with

$$\begin{aligned} \mathcal{A}_{ij0} &\sim \text{Bernoulli}(\exp(\alpha_0)) \\ \mathcal{A}_{ija} &\sim \text{Bernoulli}(\exp(\alpha_a(u_{ia}))) \\ \mathcal{B}_{ijbk} &\sim \text{Bernoulli}(\exp(\beta_b)), \end{aligned}$$

and we have assumed that all v_{ib} are non-negative integers. This can be done without loss of generality because continuous covariates must be measured to a finite number of decimal places and can therefore be rescaled appropriately.

Stable maximum likelihood estimation can be performed by implementing an EM algorithm based on this complete-data model. However, since each of the Bernoulli probabilities must not exceed 1, the complete-data model imposes non-positivity constraints on each of the parameters, and hence the observed-data log-likelihood will be maximised over a subspace of the parameter space.

To overcome this, a combinatorial EM (CEM) algorithm can be used, in which the constrained maximisation is sequentially performed on each of a collection of subspaces that together cover the entire parameter space (Marschner, 2014). The constrained estimate with the highest likelihood is the global MLE for the log-binomial model. However, this requires up to $\prod_{a=1}^{A} k_a \times 2^B$ applications of the EM algorithm, and can be computationally expensive for large models.

3 Overparameterised EM algorithm

To overcome the computational complexity of the CEM algorithm, we consider an overparameterised model. We remove the identifiability constraint, allowing all of the categorical parameters to be non-zero, and express each of the continuous covariates as the difference of two dummy covariates: $\beta_b = \beta_b^- - \beta_b^+$. This is a similar idea to an algorithm suggested for the lasso problem (Tibshirani, 1996, p. 279), and somewhat resembles, but is not equivalent to, the parameter expanded EM (PX-EM) algorithm presented by Liu et al. (1998).

The overparameterised model is $Y_i \sim \text{Bin}(N_i, \exp\{\Lambda^*(\mathbf{u}_i, \mathbf{v}_i; \boldsymbol{\theta}^*)\})$, where

$$\Lambda^*(\mathbf{u}_i, \mathbf{v}_i; \boldsymbol{\theta}^*) = \alpha_0^* + \sum_{a=1}^{A} \sum_{k=1}^{k_a} \alpha_a^*(k) 1(u_{ia} = k) + \sum_{b=1}^{B} \left(\beta_b^- v_{ib} + \beta_b^+ v_{ib}^*\right),$$

with $v_{ib}^* = v_b^{(1)} - v_{ib}$, $v_b^{(1)} = \max_i v_{ib}$. This observed-data model is equivalent to the original in the sense that the original parameter vector can be recovered using $\boldsymbol{\theta} = R(\boldsymbol{\theta}^*)$ where R is the many-to-one reduction function defined by

$$\begin{aligned} \alpha_0 &= \alpha_0^* + \sum_{a=1}^{A} \alpha_a^*(1) + \sum_{b=1}^{B} \beta_b^+ v_b^{(1)} \\ \alpha_a(k) &= \alpha_a^*(k) - \alpha_a^*(1) \\ \beta_b &= \beta_b^- - \beta_b^+. \end{aligned}$$

such that

$$\Lambda^*(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}^*) = \Lambda(\mathbf{u}, \mathbf{v}; R(\boldsymbol{\theta}^*)).$$

A complete-data model analogous to that used for the CEM algorithm (1) can be defined for the overparameterised model, which includes a total of $1 + \sum_a k_a + 2B$ parameters. The resulting collection of non-positivity constraints on the expanded set of parameters is equivalent to a constraint that the largest fitted probability for any covariate combination is 1, as desired. Thus maximum likelihood estimation over the entire parameter space can now be performed by a single application of the associated EM algorithm. Upon convergence, the estimate of the expanded parameter vector $\hat{\boldsymbol{\theta}}^*$ is converted to an estimate for the original parameter vector using $\hat{\boldsymbol{\theta}} = R(\hat{\boldsymbol{\theta}}^*)$. This method retains the stability of the EM algorithm while potentially greatly reducing the time needed to fit the model.

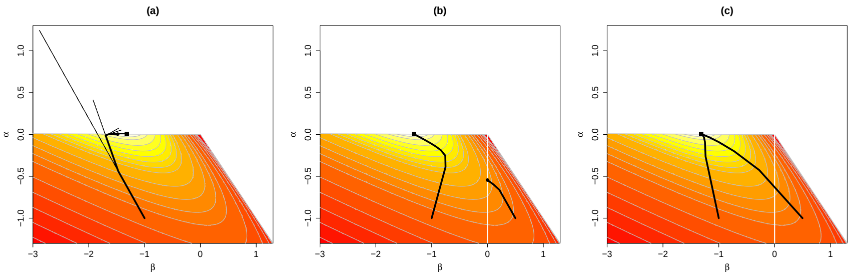


FIGURE 1. Log-likelihood surface for a log-binomial GLM with the example data; white areas are outside the parameter space, and the square marks the MLE. Dark lines are the paths followed by (a) `glm` / `glm2`, (b) the CEM algorithm, and (c) the overparameterised EM algorithm.

4 Example

Marschner (2015) discussed various types of convergence problems that can occur with log-binomial GLMs, but here we will consider an example of an additional type of convergence problem. We use a simple dataset of 100 observations with a single continuous covariate. Figure 1 shows the log-likelihood with respect to the intercept (α) and slope (β) parameters for a log-binomial GLM on this data. The MLE is marked with a square, and is on the boundary of the parameter space.

The standard implementation of Fisher scoring for GLMs in R is the `glm` function, and the `glm2` package (Marschner, 2011) provides additional stability in some scenarios. From a starting estimate of $(\hat{\alpha}, \hat{\beta}) = (-1, -1)$, the path of the modified Fisher scoring algorithm used by `glm` and `glm2` is shown in Figure 1(a). At every iteration, the full Fisher scoring step (thinner line) produces an estimate outside the parameter space, and step-halving is used to find a valid estimate. Close to the boundary of the parameter space, the gradient of the Fisher scoring step is almost parallel to the boundary, and the step-halving process only produces very small changes in the estimate until convergence is declared at a suboptimal estimate.

The CEM algorithm partitions the parameter space into two subspaces, corresponding to positive and negative values of the slope parameter. Figure 1(b) shows the path taken by the EM algorithm from $(-1, -1)$, which reaches the MLE after 81 iterations. However, because this estimate is not a stationary point of the log-likelihood, we must search the other parameter subspace. The path taken from $(-1, 0.5)$ is also shown on the same graph, converging to the constrained MLE after 41 iterations. The likelihood for this estimate is lower than that in the first parameter subspace, and so we know that the first constrained MLE is the global maximum.

Our overparameterised EM algorithm requires only one application of the EM algorithm, letting $\beta = \beta^- - \beta^+$ and searching the resulting three-dimensional parameter space. Figure 1(c) shows the path taken by the overparameterised EM from $(\hat{\alpha}, \hat{\beta}^{-}, \hat{\beta}^{+}) = (-1, -2, -1)$ and $(-1, -0.5, -1)$, projected back onto the original two dimensions, to the global MLE. This shows two advantages of the overparameterised EM approach: the path from $(-1, -1)$ is a more direct route to the MLE, requiring just 72 iterations to converge, while the path from $(-1, 0.5)$ does not stop at the 'boundary' between positive and negative values of β, converging to the global MLE in 73 iterations. Thus our proposed algorithm has addressed the false convergence of `glm` and `glm2` in this scenario, while providing faster convergence than the CEM algorithm from either starting estimate.

5 Application

ASSENT-2 (ASSENT-2 Investigators, 1999) was a randomised trial that studied 30-day mortality in 16949 patients with acute myocardial infarction (MI). A log-binomial model allows us to estimate the impact of age, MI severity, treatment delay and geographical region on death, expressed as relative risks. But the MLE is on the boundary of the parameter space, and so we cannot use the information matrix to estimate standard errors. The stability of our EM algorithm allows us to use bootstrap resampling to construct confidence intervals for the parameter estimates, avoiding any bias due to failed convergence in some samples. With three levels for each of the covariates, the CEM algorithm must search up to $3^4 = 81$ parameter subspaces for each resampled dataset. Our overparameterised EM approach, by contrast, requires just a single application of the EM algorithm. We used these algorithms to fit models on 1000 bootstrap resamples from the ASSENT-2 data. Table 1 shows a summary of the stability, number of iterations and time needed for convergence of each algorithm. The overparameterised EM approach retained the stability of the CEM algorithm, but required far fewer iterations of the EM algorithm (6970 versus 27120, on average). This translated to an average acceleration that was greater than 3-fold (1.9 versus 7.0 seconds), adding up to a difference of 85 minutes over the entire analysis. General acceleration methods for the EM algorithm could be used to improve this further still.

TABLE 1. Summary of results from fitting log-binomial GLMs to 1000 bootstrap samples from the ASSENT-2 data, using the CEM and overparameterised EM (OP EM) algorithms

	% conv.	Iterations (1000s)				Time (sec)			
Method	to MLE	Q1	Med	Mean	Q3	Q1	Med	Mean	Q3
CEM	100	6.2	7.9	27.1	13.0	1.6	2.1	7.0	3.4
OP EM	100	2.2	2.7	7.0	7.8	0.6	0.8	1.9	2.2

6 Conclusions

We have presented a novel method for maximum likelihood estimation in relative risk regression models that retains the stability of the EM algorithm but considerably reduces computational time compared to the CEM algorithm for large models. The application considered here is not of particularly high dimension. In other situations the required number of CEM subspaces may be much larger, in which case the improvements would be even more dramatic. A proof of the convergence of the overparameterised algorithm is a subject of further research. The approach could also be applied in a similar way to improve on CEM algorithms that have been published for rate difference (Marschner, 2010) and risk difference (Donoghoe and Marschner, 2014) regression models.

References

ASSENT-2 Investigators (1999). Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: the ASSENT-2 double-blind randomised trial. *The Lancet*, **354**, 716 – 722.

Donoghoe, M.W. and Marschner, I.C. (2014). Stable computational methods for additive binomial models with application to adjusted risk differences. *Computational Statistics and Data Analysis*, **80**, 184 – 196.

Liu, C., Rubin D.B., and Wu, Y.N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, **85**, 755 – 770.

Marschner, I.C. (2010). Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics*, **19**, 666 – 683.

Marschner, I.C. (2011). glm2: Fitting generalized linear models with convergence problems. *The R Journal*, **3**, 12 – 15.

Marschner, I.C. (2014). Combinatorial EM algorithms. *Statistics and Computing*, **24**, 921 – 940.

Marschner, I.C. (2015). Relative risk regression for binary outcomes: Methods and recommendations. *Australian & New Zealand Journal of Statistics*, **57**, 437 – 462.

Marschner, I.C. and Gillett, A.C. (2012). Relative risk regression: Reliable and flexible methods for log-binomial models. *Biostatistics*, **13**, 179 – 192.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, **58**, 267 – 288.

Functional data analysis of juggling records with the smooth complex logarithm

Paul H.C. Eilers[1]

[1] Erasmus University Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: `p.eilers@erasmusmc.nl`

Abstract: Multi-dimensional recordings of the movements of the hand of a juggler show a quasi-periodic pattern, with slowly changing frequency and amplitude. The smooth complex logarithm is a suitable model, especially when extended with the second harmonic.

Keywords: Analytic signal, frequency, penalty

1 Introduction

Seeing someone juggling balls is a fascinating sight. It becomes even more interesting if the movements of the juggler have been measured and are available for statistical modeling. The data I will work with here have been collected by Jim Ramsay and colleagues (Ramsay et al., 2013). They consist of ten trials of approximately ten seconds. The movements of a light emitting diode on the index finger of the juggler were recorded with high precision, in three directions: left-right (x), forward-backward (y) and up-down (z). Figure 1 shows eight seconds of trial 2.
A special section on functional data analysis in the Electronic Journal of Statistics (Volume 8, part 2, 2014) contains nine papers on the analysis of these juggling data. They all first split the data in cycles, based on chosen landmarks, and then apply different alignment or warping procedures within the cycles. In contrast, I model the series with extensions of the smooth complex logarithm model (Eilers, 2010). This model was developed for the analysis of chirp-like signals, like the sounds of crickets or bats. It fits a sine wave with variable frequency and amplitude to a time series. The juggling data, especially the x and y components have a more complex shape, but by adding harmonics with multiples of the fundamental frequency we get a very good fit to the 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.

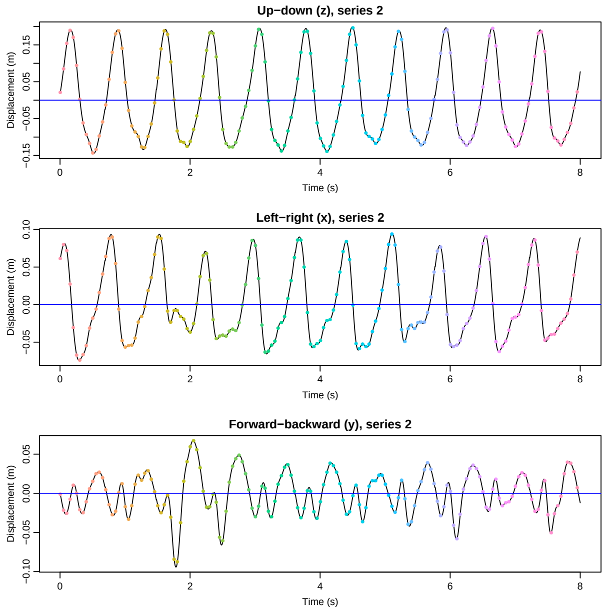


FIGURE 1. The juggling movements in three directions, from top to bottom: up-down, left-right, forward-backward.

2 The extended smooth complex logarithm

One of the most beautiful formulas in complex analysis says that

$$\exp(\alpha + i\phi) = e^{\alpha}(\cos\phi + i\sin\phi), \tag{1}$$

where $i = \sqrt{-1}$. If we have functions $\alpha(t)$ and $\phi(t)$, then the real part is a cosine with momentary amplitude $a(t) = \exp(\alpha(t))$ and momentary frequency $f(t) = d\phi/dt$. Conversely, if we observe a signal $u(t)$ that looks like a cosine with (smoothly) changing amplitude and frequency, we can try to estimate its complex logarithm. In practice we have to work with discrete data, so we observe, say, u_i for $i = 1 : n$.

An attractive recipe is to augment the observed real component with its imaginary part, v, using the Hilbert transform (see Wikipedia, under the lemma *analytic signal*). Then we can compute the momentary amplitude $\tilde{a}_i = \sqrt{u_i^2 + v_i^2}$ and the momentary phase as $\psi_i = \text{atan2}(y_i, x_i)$. However, ψ is the so-called reduced phase, always lying between $-\pi$ and π. There are a number of downward jumps, as illustrated in Figure 2. These jumps are not hard to locate and if we add 2π to ψ at each jump cumulatively, we get the continuous phase $\tilde{\phi}$.

It is possible to smooth $\tilde{a}$ and $\tilde{\phi}$ directly. A smooth estimate of ϕ is important if we want to compute its derivative, to obtain the local frequency.

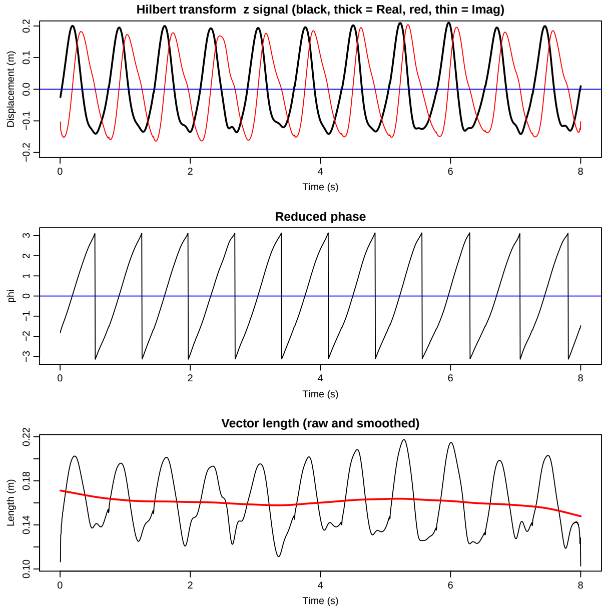


FIGURE 2. Top: the original z signal (thick black line)and its imaginary component (thin red line), obtained with the Hilbert transform. Middle: the momentary reduced phase. Bottom: the momentary amplitude and its trend.

A more refined approach is to use smooth non-linear regression, fitting $a_i \cos\phi$, with discrete roughness penalties on a and ϕ (Eilers, 2010), by minimizing

$$S = \sum_i (y_i - a_i \cos\phi)^2 + \lambda \sum_i (\Delta^2 a_i)^2 + \lambda' \sum_i (\Delta^2 \phi)^2.$$

Here Δ^2 is the operator that forms second order differences. The model is strongly non-linear in ϕ, but the phase obtained from the Hilbert transform gives us excellent initial values. So linearization works well, using

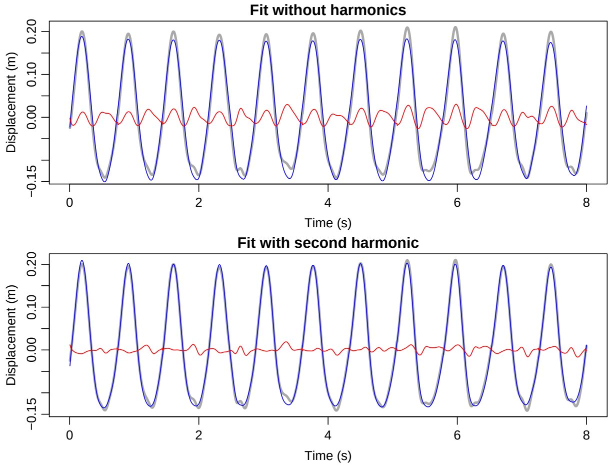
$$a_i \cos(\phi_i) \approx a_i \cos(\breve{\phi}_i) - a_i \sin(\breve{\phi}_i)(\phi_i = \breve{\phi}_i),$$

where $\breve{\phi}_i$ is an approximation to the solution.

The top panel of Figure 3 presents the result of fitting this model. A careful look shows systematic undershoots in both valleys and peaks. This is caused by the deviation from a cosine waveform. To extend the model, I introduce the second harmonic:

$$\mu_i = a_i(q_1 \cos\phi_i + q_2 \sin\phi_i + q_3 \cos 2\phi_i + q_4 \sin\phi_i), \quad (2)$$

using a and ϕ as estimated with the initial model. The coefficients q_1 to q_4 can be estimated by linear regression. The bottom panel of Figure 3 shows the result of fitting this extended model It is very well possible to give each of the sines and cosines its own smoothly changing amplitude, but the advantage of the model in (2) is its ease of interpretation: each cycle has the same basic shape, modulated in strength by the amplitude and stretched or shrunk by changes in the speed of the phase.


Fit without harmonics

Fit with second harmonic

FIGURE 3. Data (thick gray), fitted curves (thin blue), and residuals (red), with (bottom) and without (top) the second harmonic.

The movements in three directions are synchronized, so the next stage is a model in which the estimated phase for the up-down signal is taken as the basis for what could be called a bilinear or modulated factor complex logarithm model. For $j = 1:3$ it states that

$$\mu_{ij} = E(y_{ij}) = a_{ij}(q_{1j}\cos\phi_i + q_{2j}\sin\phi_i + q_{3j}\cos(2\phi_i) + q_{4j}\sin(2\phi_i)).$$

There is one series for the phase and for each direction there are four coefficients, that determine the optimal combination of sine and cosine of the fundamental and the doubled phase. This combination is modulated by a separate amplitude function for each direction. Given $a_{.j}$, the $q_{.j}$ are found by linear regression. Given these coefficients, we find $a_{.j}$ by minimizing

$$\sum_i (y_{ij} - a_{ij}f_{ij})^2 + \lambda \sum_i (\Delta^2 a_{ij})^2 + \lambda' \sum_i (\Delta^2 a_{ij})^2,$$

where $\lambda = 10^5$ and

$$f_{ij} = q_{1j}\cos\phi_i + q_{2j}\sin\phi_i + q_{3j}\cos(2\phi_i) + q_{4j}\sin(2\phi_i).$$

This is a type of varying-coefficient model. Experience has shown that alternating between updating amplitude and coefficients converges quickly

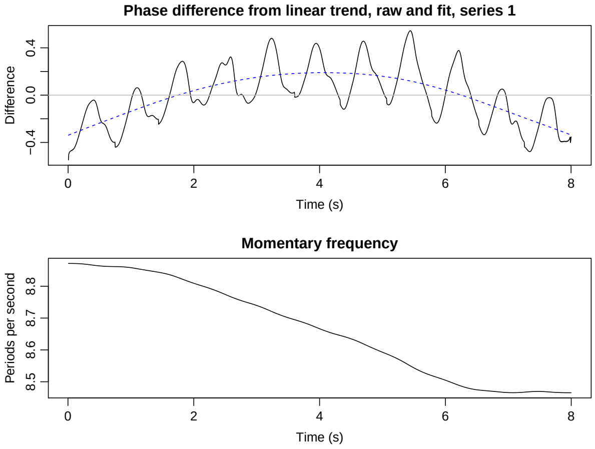


FIGURE 4. Phase and frequency for the z signal. Top panel: continuous phase as constructed from the Hilbert transform (black line) and the estimate form the model; shown are the differences with a fitted linear regression line. Bottom panel: the derivative of the estimated continuous phase, giving the momentary frequency.

to the solution. In principle it is also possible to update ϕ in each iteration, but I did not go that far.

Results are shown in Figure 5. The fit is quite good for the up-down and left-right signal. It is a bit worse for the forward-backward signal, but there is no clear indication of systematic deviations. What remains are small unsystematic variations; after all the juggler is not a machine.

3 Discussion

I have presented a model for quasi-periodic juggling data that breaks away from the popular approach that splits them in cycles, followed by warping. A common time series for the continuous phase is obtained and a separate amplitude series for each of the three movement directions. For each of them four coefficients describe how the shape of one cycle is formed from sines and cosines of the phase and its first harmonic. No landmarks are needed and at any moment in time a smooth estimate of the local frequency is available.

The smooth complex logarithm gives a more parsimonious model than a set of some 30 warping functions, one for each cycle of each coordinate. Dissection into cycles, and mapping these onto a domain from 0 to 1, effectively removes the momentary frequency, losing important information. Yet, if dissection into cycles would still be desired, one can use the crossings

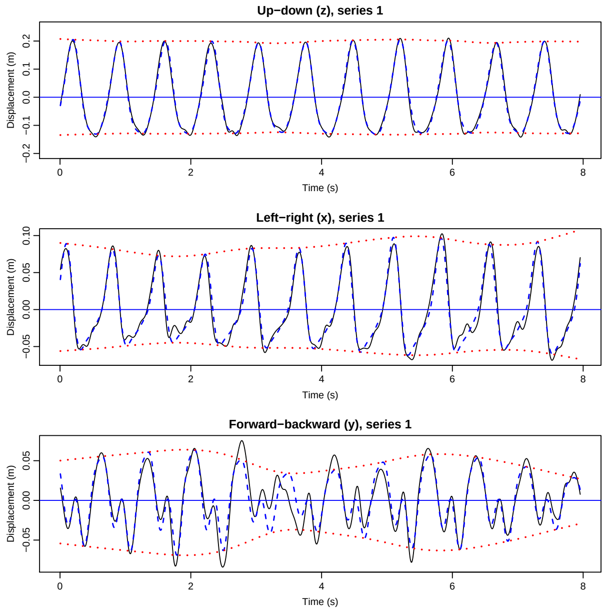


FIGURE 5. Fit with varying amplitude and second harmonic. The black lines represent the observed data and the broken blue lines the fitted values. The dotted red lines show the estimated amplitudes.

of the reduced phase with any chosen level between $-\pi$ and π to mark their start and end.

The juggling records are a good example, but it seems that the model is useful in many more situations. With modern technology it is very easy to record movements of humans or animals with high precision at low cost (think of the Kinect that comes with some Microsoft Xbox game consoles). The juggling data analyzed here are of exceptional quality, but the model is robust enough to be applied to noisy data too.

I used my carpenter's eye to set the amount of smoothing. More research is needed to determine how well automatic methods like cross validation perform for the present model.

References

Eilers, P.H.C. (2010) The Smooth Complex Logarithm and Quasi-Periodic Models. In *Statistical Modelling and Regression Structures* Kneib, T. and Tutz, G. (eds.) Springer.

Ramsay, J.O., Grible, P. and Kurtek, S. (2014) Description and processing of functional data arising from juggling trajectories. *Electronic Journal of Statistics* 8, 1811–1816.

A diagnostic plot for assessing model fit in count data models

Jochen Einbeck[1], Paul Wilson[2]

[1] Department of Mathematical Sciences, Durham University, UK
[2] School of Mathematics and Computer Science, University of Wolverhampton, UK

E-mail for correspondence: jochen.einbeck@durham.ac.uk

Abstract: Whilst many numeric methods, such as AIC and deviance, exist for assessing model fit, diagrammatic methods are few. We present here a diagnostic plot, to which we refer as 'Christmas tree plot' due its characteristic shape, that may be used to visually assess the suitability of a given count data model.

Keywords: Diagnostic plot, model fit, count data.

1 Introduction

Consider univariate count data $Y_1, \ldots, Y_n$, which are supposedly distributed according to some count distribution $F(\mu_i, \theta)$, with mean parameters $\mu_i = E(Y_i|x_i)$ possibly depending on covariates x_i (which may be vector–valued). We assume that a routine to obtain estimates $\hat{\mu}_i = \hat{E}(Y_i|x_i)$ and $\hat{\theta}$ is readily available, and we are interested in assessing graphically the quality of the resulting model fit. The idea is to check whether, for each count k, the number $N(k)$ of observed counts k is consistent with the suspected count distribution F. More precisely, denote $p_i(k) = P(k|\mu_i, \theta)$ the probability of observing the count k under covariate x_i and model F, which can be estimated by $\hat{p}_i(k) = P(k|\hat{\mu}_i, \hat{\theta})$ from the fitted model. For instance, in the special case that $F(\mu_i, \theta)$ corresponds to Pois(μ_i), one has $\hat{p}_i(k) = \exp(-\hat{\mu}_i)\hat{\mu}_i^k/k!$. This scenario is discussed in Wilson and Einbeck (2015, 2016) with focus on the case $k = 0$. This abstract generalizes those ideas to general k and F and proposes a generic diagrammatic tool.

The random variable $N(k)$ follows a Poisson–Binomial distribution with parameters $p_1(k), \ldots, p_n(k)$ (Chen and Liu, 1997). Hence, for any choice of k and F, a range of plausible values of $N(k)$ can be obtained by confidence

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).

intervals from this distribution, which can be computed using the R package `poibin` (Hong, 2013). By doing this for a range of values of k, one can draw diagrams which give envelopes for plausible values of $N(k)$ which can then be compared to the true values. Since these diagrams resemble Christmas trees, we refer to them as 'Christmas tree plots' from now on. We explain the construction of the diagram in systematic form in the next section, and give examples in the final sections.

2 The Christmas tree plot

For count data $Y = (Y_1, \dots Y_n)$, we will typically be interested in the range of counts $K = [0, \max(Y)]$, though in some applications, where very small counts are not to be expected, one may prefer using $K = [\min(Y), \max(Y)]$. Denote the chosen range by $K = [k_a, k_b]$. We construct the diagnostic plot as follows.

(i) Fit the model $F(\mu_i, \theta)$ to the data Y.

(ii) For k in $k_a ... k_b$, obtain estimates $\hat{p}_i(k)$. Use a Poisson-Binomial distribution to estimate the median $m(k) = \text{med}(N(k))$ under count data model F, as well as lower and upper limits, say $\underline{c}_\alpha(k)$ and $\bar{c}_\alpha(k)$ of a $(1-\alpha)\%$ confidence interval for $N(k)$.

(iii) Compute the median–adjusted bounds $\underline{b}_\alpha(k) = \underline{c}_\alpha(k) - m(k)$ and $\bar{b}_\alpha(k) = \bar{c}_\alpha(k) - m(k)$.

(iv) Plot the functions $\underline{b}_\alpha(k)$ and $\bar{b}_\alpha(k)$ versus k.

(v) Add to the plot the observed adjusted counts, $A(k) = N(k) - m(k)$ of the observed data Y.

If the data is consistent with the distribution fitted, the curve $A(k)$ should (largely) stay within the adjusted bands $\underline{b}_\alpha(k)$ and $\bar{b}_\alpha(k)$. If the data is *not* consistent with the distribution fitted then $A(k)$ is likely not stay within these bands. Additionally, when interpreting the bands as a measure of typical variation of $N(k)$, we can use this plot to diagnose whether the counts exhibit less random variation than expected under model F.

One may argue that due to the consideration of a sequence of confidence intervals for $k_a ... k_b$ one has to account for multiple testing issues. It should be stressed, however, that we do not consider the proposed plot as a *testing* procedure, but as a simple diagrammatic tool which supports the data analyst in identifying potential model inadequacies, similar in spirit to a QQ plot.

TABLE 1. Simulated data with upper and lower confidence intervals for $N(k)$ and $A(k)$.

k	$N(k)$	$\underline{c}_{0.1}(k)$	$\bar{c}_{0.1}(k)$	$m(k)$	$A(k)$	$\underline{b}_{0.1}(k)$	$\bar{b}_{0.1}(k)$
0	38	19	33	26	12	-7	7
1	28	27	43	35	-7	-8	8
2	15	17	31	24	-9	-7	7
3	7	6	16	10	-3	-4	6
4	8	1	7	3	5	-2	4
5	1	0	3	1	0	-1	2
6	2	0	1	0	2	0	1
7	1	0	0	0	1	0	0

3 Simulation example

Consider a covariate–free data set of size $n = 100$ drawn from a zero-inflated Poisson (ZIP) distribution with Poisson parameter 1.5 and zero-inflation parameter 0.2, that is overall mean equal to 1.2. The data are given in terms of $N(k)$ in the 2nd column of Table 1. Following the procedure outlined in Section 2 with $F \sim \text{Pois}(\mu)$ yields 90% confidence intervals for $N(k)$ (displayed in the 3rd and 4th column of Table 1), resulting in the Christmas tree plot displayed in the left hand panel of Figure 1. This plot indicates that the Poisson model is not suitable, as visible by the number of zero-observations falling well above the upper confidence band, as well as by the adjusted count $A(2)$ falling below the lower band. The right hand plot is constructed similar to that of the left, except that here the zero-inflated Poisson (ZIP) model serves as model F. Clearly this plot indicates that a ZIP model is suitable for the data.

4 Application on biodosimetry data

We consider data consisting of $n = 14430$ chromosome aberration counts previously studied by Oliveira et al. (2016). The covariate *dose*, with values between 0 and 4.5Gy, gives the radiation dose applied to blood sample cells, causing DNA damage in form of double–strand breaks. When incorrectly repaired by the cellular DNA–damage response mechanism, this can lead to dicentric chromosomes which can be counted under a microscope. That is, each examined blood sample cell contributes, for known covariate dose, exactly one count observation. For this data set, the counts take values in the range from 0 to 5. Data of this type have been fitted traditionally through Poisson regression models, though the presence of excess zero counts has been regularly reported in the literature.

FIGURE 1. Christmas tree plots for simulated covariate–free data. The dashed curve corresponds to $A(k)$ and the dotted curves give the median–adjusted bounds.

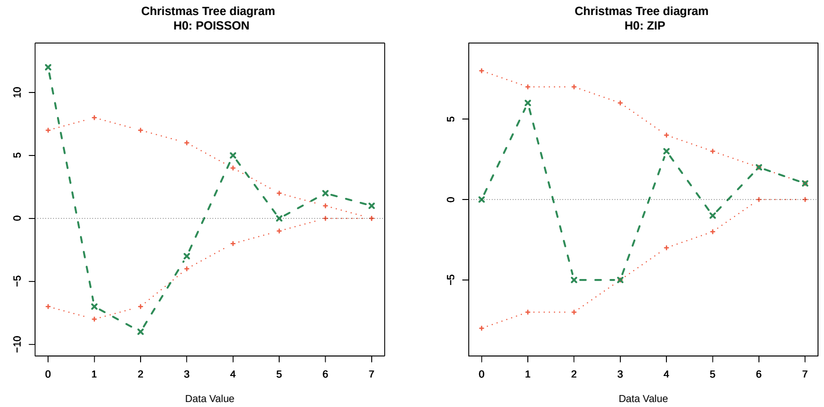


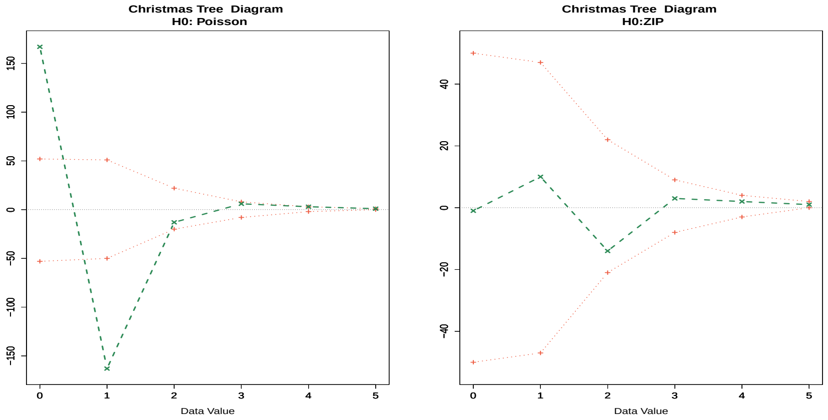
Table 2 displays the data under investigation, and Figure 2 contains the Christmas Tree diagrams obtained when Poisson and zero–inflated Poisson models, using a log–link and quadratic polynomial for dose, are fitted to these data. The left hand plot clearly indicates the unsuitability of the Poisson model, whereas the right hand plot indicates that ZIP is suitable. Oliveira et al. (2016) carried out an extensive analysis of this data set, applying several statistical tests and model selection criteria in order to decide for an adequate modelling strategy. Specifically, they found that a negative binomial type 2 model returned the lowest AIC (7489.1), closely followed by a ZIP model (AIC=7490.4). Other models considered included the Poisson as reference model (AIC=7504.7), and a Poisson Inverse Gaussian (AIC=7495.2).

The two plots in Figure 3 corresponding to the NB2 and PIG models, respectively, illustrate cases where the adjusted observed data line, $A(k)$, remains close to the centre line. For the NB2, all observations lie between the 43rd and 57th quantiles of their respective Poisson–Binomial distribution. Hence, there is less random variation amongst observed counts than would be expected under NB2, most likely indicating that the variance of the fitted model is inflated in order to accommodate the number of observed zeros. A similar effect is observed for the PIG model. In summary, these plots suggest that the ZIP model is the most adequate model for these data, deviating from what would be concluded by looking at a single–number model selection criterion such as AIC.

TABLE 2. Frequency of dicentric chromosomes after acute whole body *in vitro* exposure to doses between 0 and 4.5Gy of Cobalt-60 γ-rays. (This corresponds to data set A1 in the notation of Oliveira et al. (2016), where also the reference for the data source is provided.)

	Frequency of counts					
dose	0	1	2	3	4	5
0.00	2591	1	0	0	0	0
0.25	2185	8	0	0	0	0
0.75	2550	44	1	0	0	0
1.00	2231	54	2	0	0	0
1.50	1712	96	3	0	0	0
2.50	1196	123	7	1	0	0
3.00	1070	320	41	6	1	0
4.50	895	360	110	25	5	1

FIGURE 2. Christmas tree plots for biodosimetry data, with the hypothesized distribution F corresponding to Poisson and ZIP, respectively.



References

Chen, S.X. and Liu, J.S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875–892.

Hong, Y. (2013). poibin: The Poisson Binomial Distribution. R package version 1.2. https://CRAN.R-project.org/package=poibin

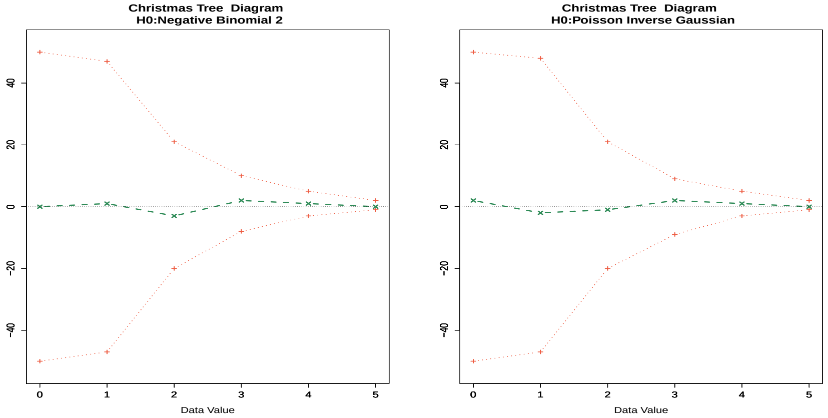
Oliveira, M., Einbeck, J., Higueras, M., Ainsbury, E., Puig, P. and Rothkamm, K.

(2016). Zero-inflated regression models for radiation–induced chromosome aberration data: A comparative study. *Biometrical Journal* **58**, 259-279.

Wilson, P. and Einbeck, J. (2015). A simple and intuitive test for number–inflation or number–deflation. In: Wagner, H. and Friedl, H. (Eds). Proc's of the 30th IWSM, Linz, Austria, Vol 2, pages 299–302.

Wilson, P. and Einbeck, J. (2016). On statistical testing and mean parameter estimation for zero–modification in count data regression. Proc's of the 31st IWSM, Rennes, France, *to appear*.

FIGURE 3. Christmas tree plots for biodosimetry data, with the hypothesized distribution F corresponding to NB2 and PIG, respectively.



Estimation in nonlinear dynamic systems using Linearized ODE penalized splines

Gianluca Frasso[1], Philippe Lambert[1,2]

[1] Faculté des sciences humaines et sociales, Méthodes quantitatives en sciences sociales, Université de Liège, Belgium.

[2] Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Belgium

E-mail for correspondence: `Gianluca.Frasso@ulg.ac.be`

Abstract: We propose an efficient penalized spline framework for estimation in dynamic systems described by nonlinear ordinary differential equations.

Keywords: nonlinear ODEs; penalized splines; quasilinearization.

1 Introduction

Dynamic systems are observed within many scientific fields and can be compactly described by (systems of) ordinary differential equations (ODEs) of the form

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}(t) = f(t, \boldsymbol{x}, \boldsymbol{\theta});\ t \in [0, T] \\ \text{s.t. } \boldsymbol{x}(t_0) = \boldsymbol{s}_{t_0} \text{ for } t_0 \in \mathcal{T}_0 \subset [0, T], \end{cases} \tag{1}$$

where $f(t, \boldsymbol{x}, \boldsymbol{\theta})$ is a known nonlinear function of the state function(s) $\boldsymbol{x}(t)$ and of the (unknown) ODE-parameters $\boldsymbol{\theta}$. If we consider the observed (noisy) data ($\boldsymbol{y}$) as realizations of a process driven by state functions solving a given ODE, we can estimate the data signal and the ODE-parameters using statistical techniques. In particular, in what follows we suppose that

$$\boldsymbol{y} = \boldsymbol{x}(t|\boldsymbol{\theta}) + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon}$ is a random error term. In this framework, we introduce the LODE-PS (Linearized Ordinary Differential Equation P-splines) estimation strategy for $\boldsymbol{\theta}$ and $\boldsymbol{x}(t)$ based on iterative (penalized) least squares.

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 Estimation procedure

In order to simplify our presentation, we assume here Gaussian distributed errors with constant variance τ^{-1} and a one-dimensional $\boldsymbol{x}(t)$. We propose to approximate the ODE solution using a linear combination of B-spline functions: $\boldsymbol{x}(t|\boldsymbol{c}) = \sum_k^K \boldsymbol{b}_k(t)c_k$. In order to connect the estimated spline coefficients to the ODE-model, we penalize for violations of Eq. (1). This leads to the following penalized log-likelihood function:

$$2\mathcal{L}(\boldsymbol{c}, \boldsymbol{\theta}, \tau, |\gamma, \boldsymbol{y}) = N\log(\tau) - \tau\sum_{i=1}^{N}(y_i - x_i(t|\boldsymbol{c}))^2 - \gamma\text{PEN}(\boldsymbol{c}, \boldsymbol{\theta}), \tag{3}$$

where $\text{PEN}(\boldsymbol{c}, \boldsymbol{\theta}) = \int \|\dot{\boldsymbol{x}}(s|\boldsymbol{c}) - f(s, \boldsymbol{x}(s|\boldsymbol{c}), \boldsymbol{\theta})\|^2 \,\mathrm{d}s$, $\boldsymbol{c}$ is a K-vector of spline coefficients, $\boldsymbol{\theta}$ is a D-vector of unknown ODE-parameters and γ is an ODE-compliance parameter such that $\hat{\boldsymbol{x}}(t)$ is forced to solve (1) if $\gamma \to \infty$.

It is not straightforward to maximize Eq. (3) since the optimal spline coefficients depend in a complicated way on $\boldsymbol{\theta}$ and PEN is nonlinear in $\boldsymbol{c}$ and/or $\boldsymbol{\theta}$. Suppose now that approximations to the state function $\tilde{\boldsymbol{x}}(t)$ and to the ODE-parameters $\tilde{\boldsymbol{\theta}}$ are available. Following Bellman and Kalaba (1965), Eq. (1) can be linearized as

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}(t) \approx \tilde{f}(t) + \sum_{d=1}^{D}(\theta_d - \tilde{\theta}_d)\frac{\partial\tilde{f}}{\partial\theta_d}(t) + (\boldsymbol{x}(t) - \tilde{\boldsymbol{x}}(t))\frac{\partial\tilde{f}}{\partial\boldsymbol{x}}(t), \tag{4}$$

where $\tilde{f}(t) = f(t, \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\theta}})$. Eq. (4) is a non-homogeneous linear ODE with solution:

$$\boldsymbol{x}(t) = \boldsymbol{p}(t) + \sum_{d=1}^{D}\theta_d\boldsymbol{q}_d(t).$$

Functions $\boldsymbol{p}(t)$ and $\boldsymbol{q}_d(t)$ solve the linear ODEs

$$\begin{aligned}
&\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t}(t) = \tilde{f}(t) - \sum_{d=1}^{D}\tilde{\theta}_d\frac{\partial\tilde{f}}{\partial\theta_d}(t) + (\boldsymbol{p}(t) - \tilde{\boldsymbol{x}}(t))\frac{\partial\tilde{f}}{\partial\boldsymbol{x}}(t);\ p(t_0) = \tilde{x}(t_0),\\
&\frac{\mathrm{d}\boldsymbol{q}_d}{\mathrm{d}t}(t) = \frac{\partial\tilde{f}}{\partial\theta_d}(t) + \boldsymbol{q}_d(t)\frac{\partial\tilde{f}}{\partial\boldsymbol{x}}(t);\ q_d(0) = 0,
\end{aligned} \tag{5}$$

and hence

$$\begin{aligned}
&\boldsymbol{p}(t) = \Lambda(t)\int_{t_0}^{t}\Lambda^{-1}(s)\left(-\frac{\partial\tilde{f}}{\partial\boldsymbol{x}}(s)\tilde{\boldsymbol{x}}(s) - \sum_{d=1}^{D}\tilde{\theta}_d\frac{\partial\tilde{f}}{\partial\theta_d}(s) + \tilde{f}(s)\right)\mathrm{d}s + k,\\
&\boldsymbol{q}_d(t) = \Lambda(t)\int_{t_0}^{t}\frac{\partial\tilde{f}}{\partial\theta_d}(s)\Lambda^{-1}(s)\mathrm{d}s,
\end{aligned}$$

with $\Lambda(s) = \mathrm{e}^{\int_{t_0}^{t} \frac{\partial \tilde{f}}{\partial \boldsymbol{x}}(s)\mathrm{d}s}$ and k a constant depending on $\tilde{x}(t_0)$. Therefore, we can maximize Eq. (3) and select γ by iterating between the following steps:
0) Set initial values for all the unknowns: $\tilde{\boldsymbol{c}}$, $\tilde{\boldsymbol{\theta}}$, $\tilde{\gamma}$ and $\tilde{\tau}$.
1) For fixed $\{\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\theta}}, \tilde{\gamma}, \tilde{\tau}\}$, update the vector of spline coefficients by minimizing the (quadratic) penalized criterion:

$$J(\boldsymbol{c}|\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\theta}}, \tilde{\gamma}, \tilde{\tau}) = \tilde{\tau}\, \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{c}\|^2 + \tilde{\gamma}\mathrm{PEN}_\ell(\boldsymbol{c}|\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\theta}}), \text{ where}$$
$$\mathrm{PEN}_\ell(\boldsymbol{c}|\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\theta}}) = \int \left\| \sum_k^K \dot{\boldsymbol{b}}_k(s)c_k - \tilde{f}(s) - \left(\sum_k^K \boldsymbol{b}_k(s)c_k - \tilde{\boldsymbol{x}}(t) \right) \frac{\partial \tilde{f}}{\partial \boldsymbol{x}}(s) \right\|^2 \mathrm{d}s,$$

and $\dot{\boldsymbol{b}}_k(s)$ is a first derivative of B-spline vector.
2) Update $\tilde{\boldsymbol{x}}(t) = \boldsymbol{x}(t, \tilde{\boldsymbol{c}})$, compute $\boldsymbol{p}(t)$ and $\boldsymbol{q}_d(t)$ and update $\tilde{\boldsymbol{\theta}}$ and τ by maximizing the following criterion (quadratic in $\boldsymbol{\theta}$):

$$H(\boldsymbol{\theta}, \tau|\tilde{\boldsymbol{c}}) = \frac{N}{2}\log\tau - \frac{\tau}{2}\left\| \boldsymbol{y} - \boldsymbol{p}(t) - \sum_{d=1}^{D} \theta_d \boldsymbol{q}_d(t) \right\|^2 .$$

3) Update $\gamma = (\mathrm{ED})/\|\mathrm{PEN}_\ell(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{\theta}})\|^2$ (see Ruppert et al., 2003) where ED is the effective model dimension of the smoother (see Hastie and Tibshirani, 1990).
LODE-PS generalizes the QL-ODE-P-spline framework (Frasso et al., 2015). It can easily deal with unknown state (initial and/or boundary) conditions and, by modifying Eqs. (5), can be adapted to handle systems of ODEs (as illustrated in Section 3). In addition the proposed (profiled) likelihood optimization simplifies the one introduced by Ramsay et al. (2007) when dealing with nonlinear systems since all the unknowns in (3) are estimated by solving simple least squares problems.
The presented framework can be extended to handle non Gaussian data (such as in GLM settings). In the next section we show an example based on Poisson distributed data.

3 Two real data examples

We first analyze the dynamic of two competing species by modeling the population density of snowshoe hares and Canadian lynx over time as described by the Lotka-Volterra model:

$$\begin{cases} \dfrac{\mathrm{d}x_1}{\mathrm{d}t}(t) = x_1(t)\,[\theta_1 - \theta_2 x_2(t)], \\ \dfrac{\mathrm{d}x_2}{\mathrm{d}t}(t) = -x_2(t)\,[\theta_3 - \theta_4 x_1(t)]. \end{cases}$$

Figure 1 shows the raw data and the estimates obtained using the LODE-PS approach introduced in Section 2. The estimated state functions appropriately describe the observed dynamics and they are close to the numerical solutions computed using a Runge-Kutta scheme for the LODE-PS parameter estimates.

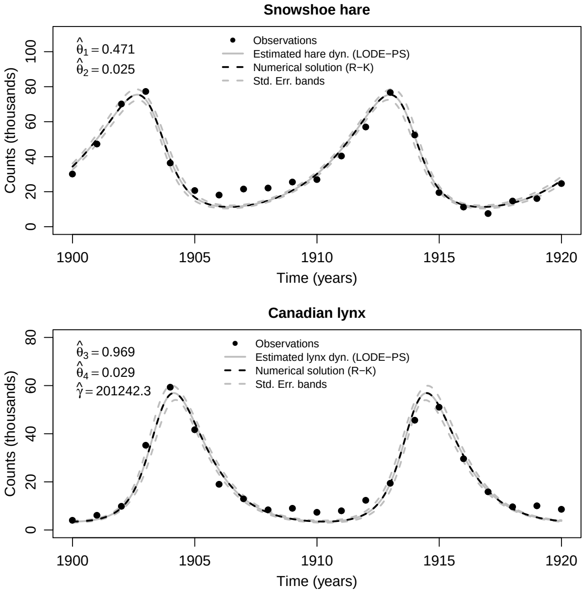


FIGURE 1. Raw data and estimates obtained for the Canadian lynx vs snowshoe hare predator-pray dynamics. The estimated system parameters are reported in the legend.

As second example we analyze infectious disease epidemic data. Figure 2 shows the total number of infectious and recovered subjects (black dots) reported for a common-cold outbreak observed during 21 days (from October 1967) in the Tristan da Cunha island (see e.g. Shibli et al., 1971). We analyze these data using a Susceptible-Infectious-Removed epidemic compartmental model:

$$\begin{cases} \dfrac{\mathrm{d}S}{\mathrm{d}t}(t) = -\beta \dfrac{I}{N}(t)S(t), \\ \dfrac{\mathrm{d}I}{\mathrm{d}t}(t) = \beta \dfrac{I}{N}(t)S(t) - \delta I(t), \\ \dfrac{\mathrm{d}R}{\mathrm{d}t}(t) = \delta I(t). \end{cases}$$

Here we assume a constant population size (N) and consider the number of new infectious subjects at each time $\mathrm{d}I^{+}(t)$ as Poisson distributed. This implies that the optimal spline coefficients can be estimated by maximizing:

$$J(\boldsymbol{c}|\tilde{\beta}, \tilde{\delta}, \tilde{\gamma}) = \sum_{i=1}^{21} \Big(\mathrm{d}I^{+}(t_i) \ \log \boldsymbol{\mu}(t_i|\boldsymbol{c}, \tilde{\beta}, \tilde{\delta}) - \boldsymbol{\mu}(t_i|\boldsymbol{c}, \tilde{\beta}, \tilde{\delta}) \Big) - \tilde{\gamma} \mathrm{PEN}_{\ell}(\boldsymbol{c}|\tilde{\boldsymbol{c}}, \tilde{\beta}, \tilde{\delta}).$$

where $\mu(t|\boldsymbol{c},\tilde{\beta},\tilde{\delta}) = \mathrm{E}(\mathrm{d}I^{+}(t)|\boldsymbol{c},\tilde{\beta},\tilde{\delta}) = \beta\boldsymbol{S}(t|\boldsymbol{c},\tilde{\beta},\tilde{\delta})\boldsymbol{I}(t|\boldsymbol{c},\tilde{\beta},\tilde{\delta})$. For $\boldsymbol{c}$ estimated in Step 1, the ODE parameters can be estimated by maximizing:

$$H(\beta,\delta|\hat{\boldsymbol{c}}) = \sum_{i=1}^{21}\left(\mathrm{d}I^{+}(t_i)\ \log\left(\beta\boldsymbol{\nu}_S(t_i)\boldsymbol{\nu}_I(t_i)\right) - \left(\beta\boldsymbol{\nu}_S(t_i)\boldsymbol{\nu}_I(t_i)\right)\right),$$

with

$$\begin{aligned}\boldsymbol{\nu}_S(t) &= \boldsymbol{p}_S(t) + \beta\boldsymbol{q}_{S,\beta}(t) + \delta\boldsymbol{q}_{S,\delta}(t),\\ \boldsymbol{\nu}_I(t) &= \boldsymbol{p}_I(t) + \beta\boldsymbol{q}_{I,\beta}(t) + \delta\boldsymbol{q}_{I,\delta}(t).\end{aligned}$$

Functions $\boldsymbol{p}_i(t), \boldsymbol{q}_{i,j}(t)$ for $i \in \{S, I\}$ and $j \in \{\beta,\delta\}$ are the solutions of the linearized ODEs defined for the first two equations of the SIR model in analogy with Eqs. 5.

The estimated state functions in Figure 2 (dashed black lines) describe the observed epidemic states sufficiently well and are compliant with the ODE numerical solutions (gray solid lines). The optimal ODE parameters appear consistent with the results presented in the literature and obtained by using different approaches (see e.g. Toni et al., 2009).

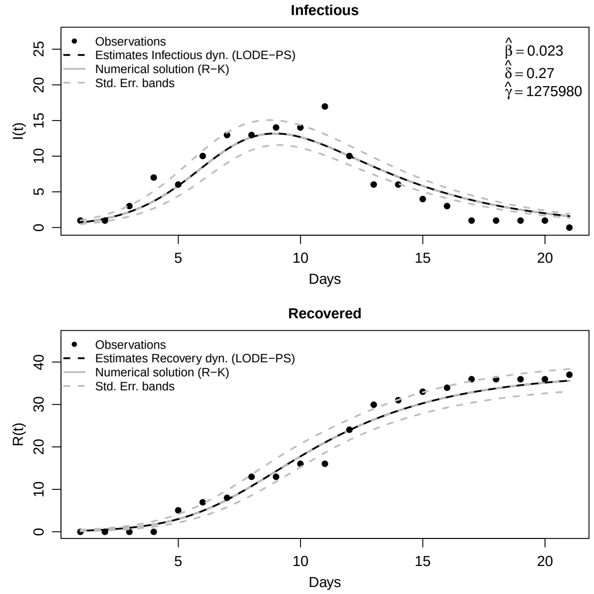


FIGURE 2. Raw data and estimates obtained for the common cold infectious dynamics. The estimated SIR system parameters are reported in the legend.

4 Discussion

We presented the LODE-PS smoothing approach for the estimation of dynamic systems described by nonlinear (systems of) ODEs. The estimates

are obtained by means of a profiled likelihood maximization procedure. Our approach exploits a quasi-linearization of the ODE problem and enables to estimate all the unknowns using simple least squares procedures. This represents a valuable simplification of the optimization task in both Gaussian and non Gaussian settings.
We have evaluated the performances of the proposed methods by dealing with two real data example. In both cases the LODE-PS approach ensured satisfactory estimates.
The proposed framework can also be generalized. For example, it can be extended to handle time-varying ODE-parameters (modeled, for example, through P-spline smoothers) by modifying the profiled likelihood optimization task in step 2).

Acknowledgments: The authors acknowledge financial support from IAP research network P7/06 of the Belgian Government (Belgian Science Policy) and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique'.

References

Bellman, R. and Kalaba R. (1965) *Quasilinearization and nonlinear boundary-value problems.* Modern analytic and computational methods in science and mathematics. American Elsevier Pub. Co.

Frasso, G., Jaeger, J., and Lambert, P. (2015). Inference in dynamic systems using B-splines and quasilinearized ODE penalties. *Biometrical Journal*, in press.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models.* London: Chapman & Hall.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society, Series B*, **69**, 741 – 796.

Ruppert, D., Wand, P., and Carroll, R. (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Shibli, M., Gooch, S., Lewis, H. E., and Tyrrell, D. A. J. (1971). Common colds on Tristan da Cunha. *Journal of Hygiene*, **69**, 255.

Toni, T., Welch, D., Strelkowa, N., Ipsen, E., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, **6**, 31, 178 – 202.

A Linear Mixed Models with Functional Covariable Applied to Chlorophyll Concentration Data

Gustavo Adolfo Gómez Escobar[1], Mercedes Andrade-Bejarano[1], Ramn Giraldo-Henao[2]

[1] Universidad del Valle, Colombia
[2] Universidad Nacional de Colombia, Colombia

E-mail for correspondence: gustavo.gomez@correounivalle.edu.co

Abstract: This research have the aim to model chlorophyll concentration in Tabasco papper plant through a mixed model with functional covariable. The plants have been effected by differentlevels of stress caused by type of fertilizer and irrigation levels we also used as a covariable functional the spectral signature

Keywords: Functional Covariable; Mixed Model; Bootstrap; Spectral Asignature; Chlorophyll.

1 Case Study

The spectral behavior of vegetation, that is, the amount of reflective energy measured in each individual or plant along the spectrum, depends of the nature of the same, their interactions with solar radiation, other climatic factors, availability of nutrients and water in their environment(Jensen, 2000). Warner et al, (2009) indicate that the information provided by the spectral signature is the fraction of vegetation cover, chlorophyll content, the Green Index of leaf area and other biophysical parameters of the plants. The spectral signature in accordance with its resolution contains information of reflectance for diferent wavelengths. In our research this consists of 900 wavelengths. As we has a data sequence which can be perceived almost as a function, is such a case that such data can be addressed through Functional Data Analysis (FDA) (Ramsay & Silverman, 2005) (Ferraty & Vieu, 2006).

The aim of this research is to evaluate the concentration of chlorophyll in Tabasco pepper plants under levels of stress caused by the type of fertilizer applied and levels of irrigations. We used a factorial design with four levels per factor. The levels for fertilizer are: F1 (Solution without Boron), F2 (Solution without Iron), F3 (Solution without Manganese) y F4 (complete solution). For irrigation: R1 (15 ml), R2 (75 ml), R3 (115 ml) y R4 (225 ml)

We had eight replications per treatment, for a total of 128 experimental units, which are Tabasco pepper plants. Each treatment was assigned randomly to the experimental units in order to reduce experimental error. Chlorophyll concentration was measured to the experimental units during seven times (weeks) = 1; 2; 4; 5; 6; 7; 8. ,

2 Material and Methods

Chlorophyll data correspond to longitudinal measurements. It is intended to capture the dependence between the measurements of the specific subject through a Mixed Model with Functional Covariable (MMFC), where the functional covariable is included by the operator $\Psi(\chi(t)) = \int_T \psi(t)\chi(t)dt$ where $\psi(t), \chi(t) \in H$ being H separable Hilbert space (Goldsmith, et al. 2012). For an individual i with n_i repeated measurements over time, we have an array of responses $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})^T$, which can be modeled as

$$\mathbf{y}_i = \int_T \boldsymbol{\chi}_i(t)\tilde{\beta}(t)dt + X_i\beta + Z_i\mathbf{b}_i + \mathbf{e}_i, \tag{1}$$

where $\boldsymbol{\chi}_i(t)$ and $\tilde{\beta}(t)$ belong to H. X_i y Z_i are matrices of fixed and random effects, respectively. β, $\mathbf{b}_i$ correspond to vector, for the fixed and random effects like a traditional mixed model $\mathbf{N}(\mathbf{0}, D)$ (Verbeke & Molenberghs, 2000) y e_i is a vector of random errors $\mathbf{N}(\mathbf{0}, \Sigma)$. Assuming the functional data is $\chi_{il}(t) = \sum_{k=1}^{k_\chi} c_{ijk}\phi_j(t), \;\; \chi_i(t) = \mathbf{c}_{il}^T\boldsymbol{\phi}$ and the functional parameter can be expressed as $\tilde{\beta}(t) = \sum_{j=1}^{k_\beta} \upsilon_j\theta_j(t), \;\; \tilde{\beta}(t) = \boldsymbol{\theta}^T\boldsymbol{\upsilon}$, where $\boldsymbol{\phi}$ y $\boldsymbol{\theta}$ are: basis functions of sizes k_χ y k_β. We reconstructed the matrix of fixed effects for an individual i as $\tilde{X}_i = (\mathbf{c}_i^T\mathbf{J}_{\phi\theta}, \; X)$, matrix $\mathbf{J}_{\phi\theta}$ is of size k_χ, k_β It is conformed by $\int_T \phi_i(t)\theta_j(t)dt$. The model is expressed as

$$\mathbf{y}_i = \tilde{X}_i\beta^* + Z_i\mathbf{b}_i + \mathbf{e}_i, \tag{2}$$

where $\beta^{*T} = (\boldsymbol{\upsilon}^T\beta^T)$is the vector of fixed effects must be estimated. In our case the MMFC is estimated by Restricted Maximum Likelihood (REML) (Patterson & Thompson, 1971)for variance components and the fixed effects by Generalized Least Squares (Verbeke & Molenberghs, 2000). For confidence bands of functional parameter we use fully parametric Bootstrap for Mixed Models.

3 Results

The concentration of chlorophyll have an increasing linear behavior in terms of weeks, for this reason we include time (in weeks) as a fixed effect. The first level of the model under consideration is given by

$$\textbf{Chlorophyll}_i = \beta_{0i} + \beta_{1.} + \beta_{2.} + \beta_{3i}t + \int_T \boldsymbol{\chi}_i(t)\tilde{\beta}(t)dt + \mathbf{e}_i \tag{3}$$

In the second level we considered the change of the specific subject intercept and the variation of slope in terms of time

$$\begin{aligned}
\beta_{0i} &= \beta_0 + b_{0i} \\
\beta_{1.} &= \beta_1 F_1 + \beta_2 F_2 + \beta_3 F_3 + \beta_4 R_1 + \beta_5 R_2 + \beta_6 R_3 \\
\beta_{2.} &= \beta_7 F_1 R_1 + \beta_8 F_1 R_2 + \beta_9 F_1 R_3 + \beta_{10} F_2 R_1 + \beta_{11} F_2 R_2 + \beta_{12} F_2 R_3 \\
&\quad + \beta_{13} F_3 R_1 + \beta_{14} F_3 R_2 + \beta_{15} F_3 R_3 \\
\beta_{3i} &= \beta_{16} + b_{1i}
\end{aligned}$$

The spectral signatures were smoothed using 20 *B-Spline* (Figure 1) which were chose through the Generalized Cross Validation Criteria and the number of basis functions for functional parameter $\tilde{\beta}(t)$ were determined through marginal Akaike Information Criterion. According to ANOVA results only irrigation factor was significant for explaining the chlorophyll concentration, we also found that the trend over time is significant. Furthermore Figure 2 shows that there are wavelengths significantly different from zero.

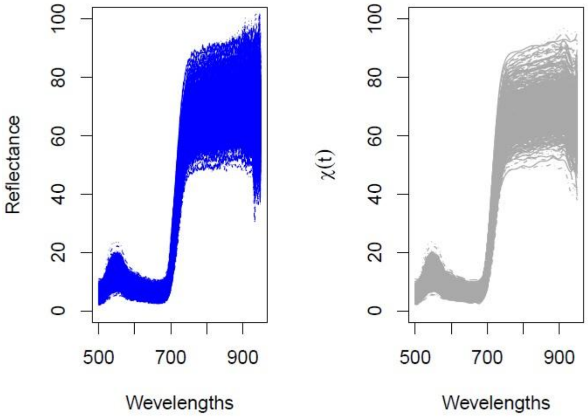


FIGURE 1. Left Spectral signatures and Rigth Spectral signature smoothed

TABLE 1. ANOVA

	numDF	denDF	F-value	p-value
(Intercept)	1	739.00	45119.99	0.00
Fertilizer	3	112.00	1.64	0.18
Irrigation	3	112.00	5.13	0.00
Time	1	739.00	540.87	0.00
Fertilizer:Irrigation	9	112.00	0.84	0.58

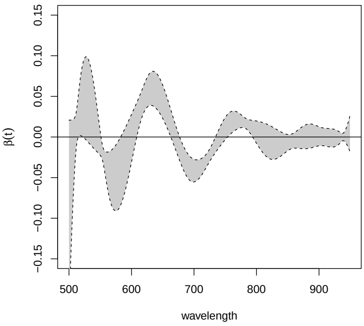


FIGURE 2. Functional parameter with confidence bands

Acknowledgments: A special Thanks to the Organizing Committee of International Workshop on Statistical Modelling for accept this work

References

Ferraty, F., & Vieu, P., (2006). *Nonparametric functional data analysis: theory and practice.* Springer Science & Business Media.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., & Reich, D. (2012). Longitudinal penalized functional regression. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **61**(3), 453 – 469.

Jensen, J.R. (2005). *Introductory digital image processing: a remote sensing perspective. Series in geographic information science..* Pearson Prentice Hall, South California. 526 p.

Patterson, H., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal *Biometrika*, **58**(3), 545 – 554.

Ramsay, J., and Silverman, B., (2005). *Functional Data Analysis.* Springer. New York, Verbeke, G., and Molenberghs, G. (2000). *Linear mixed models for longitudinal data.*. Springer-Verlag New York,

Warner, T., Duane N., and Foody G. (2009). *The SAGE Handboo of Remote Sensing.*. SAGE Publications Ltd , Ed 2nd,.

A two-component regression model for the number of births

Gillian Heller[1], Swaran Naidu[2]

[1] Department of Statistics, Macquarie University, Sydney, Australia
[2] Department of Obstetrics & Gynaecology, Fiji National University, Fiji

E-mail for correspondence: gillian.heller@mq.edu.au

Abstract: A negative binomial regression model for parity (number of births) is developed, in which mean parity is modelled with two components relating to woman's age. The first is a parametric growth curve which operates during the childbearing years. The second is a cohort effect which models changes in birth rates over time, and is a nonparametric term. The model is implemented on Fijian data, and reveals different trends in childbearing between the two main ethnic groups in that country.

Keywords: Negative binomial regression; Number of births; parity; Growth curves; Splines.

1 Introduction

Patterns of parity, the number of times a woman has given birth, typically vary across demographic groupings and over time. In a dataset collected at a Fijian community health facility, examination of parity against woman's age reveals not only the expected positive trend with age during the childbearing years, but also a continuing upward trend post childbearing. The latter is interpreted as a cohort effect due to a tendency to smaller family sizes over the last few decades. It is of interest to characterise the difference in the pattern of parity between the two ethnic groups, iTaukei (indigenous Fijians) and Fijians of Indian Descent (FID). We develop a negative binomial regression model for parity that adjusts for age by separating the childbearing and cohort effects. The childbearing effect is characterised by a parametric growth curve, while the cohort effect is modelled nonparametrically. Maximum likelihood estimation is implemented with a two-stage iterative estimation procedure for the two components of the predictor. This

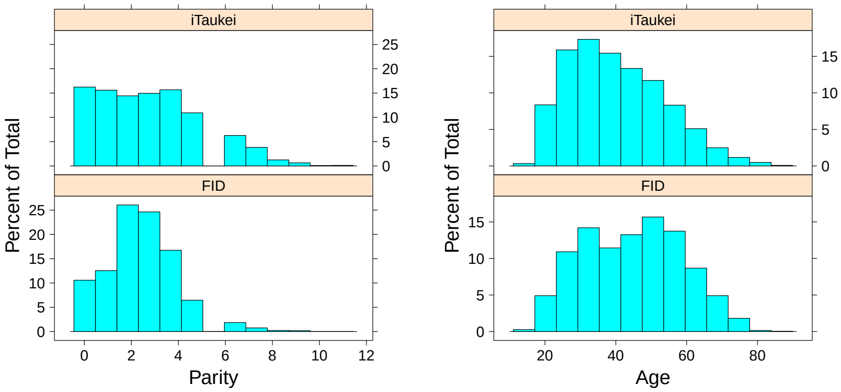


FIGURE 1. Histograms of parity and age by ethnicity

method is semiparametric in that it employs both parametric and nonparametric functions of age. However, it is different from what is usually meant by semiparametric regression models, as the parametric and nonparametric terms are for the same covariate.

2 The data

The data were collected over the period 2013-14 at the Viseisei Sai Health Centre in rural Fiji, on $n = 5{,}136$ women, of whom 48% were iTaukei and 52% FID. Figure 1 shows intriguing differences in the distributions of both parity and age over the two ethnic groups. Figure 2 shows the relationship between parity and age, as estimated by a smoothing spline. Parity increases till about the age of 40 (the childbearing years), then flattens off, then increases again into old age. We consider separating the effect of age on parity into two components: (1) the childbearing years; and (2) a cohort effect, which operates across all ages and is the effect of having borne children in a particular epoch.

3 Statistical model for parity

3.1 Childbearing effect

We employ obvious features of childbearing to characterise its effect in the statistical model: all women start with zero parity; and parity is a nondecreasing function of age, until fertility ceases with the menopause. This suggests modelling expected parity with a growth curve, with lower asymptote zero and upper asymptote expected parity at menopause. A flexible parametric family of growth curves is given by the inverse of the

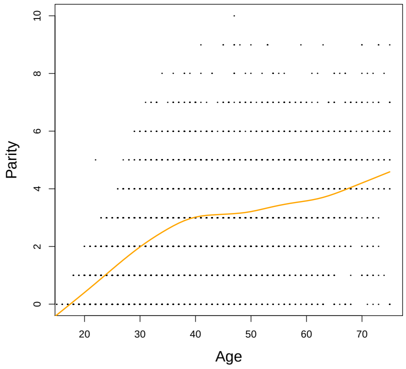


FIGURE 2. Scatterplot of parity vs age, with smoothing spline.

generalized logistic function (Richards 1959), with lower asymptote zero, upper asymptote k, and parameters b and w, which govern the slope and shape:

$$y(x) = \frac{k}{(1+e^{-bx})^w} \ . \tag{1}$$

We utilise (1) as the childbearing component of the parity model.

3.2 Cohort effect

The increasing trend of parity after the childbearing years, which we interpret as a cohort effect, is less amenable to parametric assumptions than the childbearing effect. While it appears from the data to be a positive relationship, we do not wish to make any assumptions about its shape. We therefore include it in the model as a smooth term $s(x)$.

3.3 The regression model

We use a negative binomial regression model for parity (y):

$$y \,|\, x \sim \text{NB}(\mu, \sigma) \ ; \quad \mu = \frac{k}{(1+e^{-bx})^w} + s(x) \tag{2}$$

where x is age, and $b > 0$, $w > 0$, $k > 0$. For interpretability and stability of computation we impose the constraint $s(x) \geq 0$, and model $s(x)$ with exponentiated B-splines. A two-stage iterative procedure for likelihood maximization is used: in the first stage, parameters of the growth curve are estimated for fixed spline; and in the second stage parameters of the spline are estimated for fixed growth curve. The R function `optim` is used for the maximization. We note that in the region of childbearing

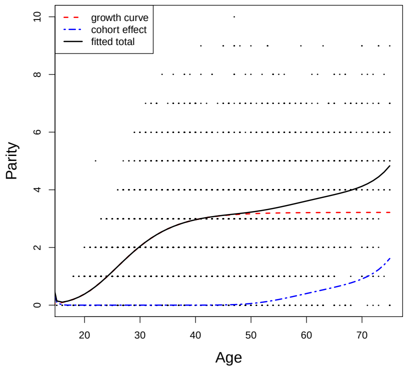


FIGURE 3. Scatterplot of parity vs age, with fitted curves.

growth the model is not identifiable; however we obtain sensible results by initially setting the curve $s(x)$ to zero, and allowing the growth curve to dominate the solution in the region of growth due to childbearing.

4 Results

The model was initially implemented on all women. Fitted curves are shown in Figure 3, and can be seen to match the smoothing spline curve. The growth curve flattens off around the age of 40, at the upper asymptote of $\hat{k} = 3.22$. The cohort effect becomes active around the age of 50, meaning that women who are currently 50 years and over were bearing children at a time when rates of birth were higher than in the current cohort of childbearing women. Randomized quantile residuals (not shown) indicate a good model fit.

Ethnicity effects

Figure 4 shows the components of the fitted model by ethnic group. The FID group has growth curve asymptote $\hat{k} = 2.74$, compared to the iTaukei $\hat{k} = 3.71$. The cohort effect for iTaukei is weak, whereas for FID it rises sharply from the age of 50. From the age of about 60, mean parity for the two ethnic groups is not very different, at a mean of about 4 births per woman. Nonparametric bootstrap confidence intervals were computed for the total parity curves by ethnicity. These show the clear separation of mean parity between the groups at all ages except the oldest (women around 70 years of age and older).

Putting this together we conclude that for the oldest women in our cohort, mean parity for both ethnic groups was around 4 births. In the FID group, mean parity has decreased to 2.74 births in the current cohort of childbearing women, whereas for iTaukei, higher parity persists at 3.71 births. This accords with the trend noted by Seniloli (1992): ".. fertility is levelling off among Fijians and consistently declining among the Indians in Fiji".

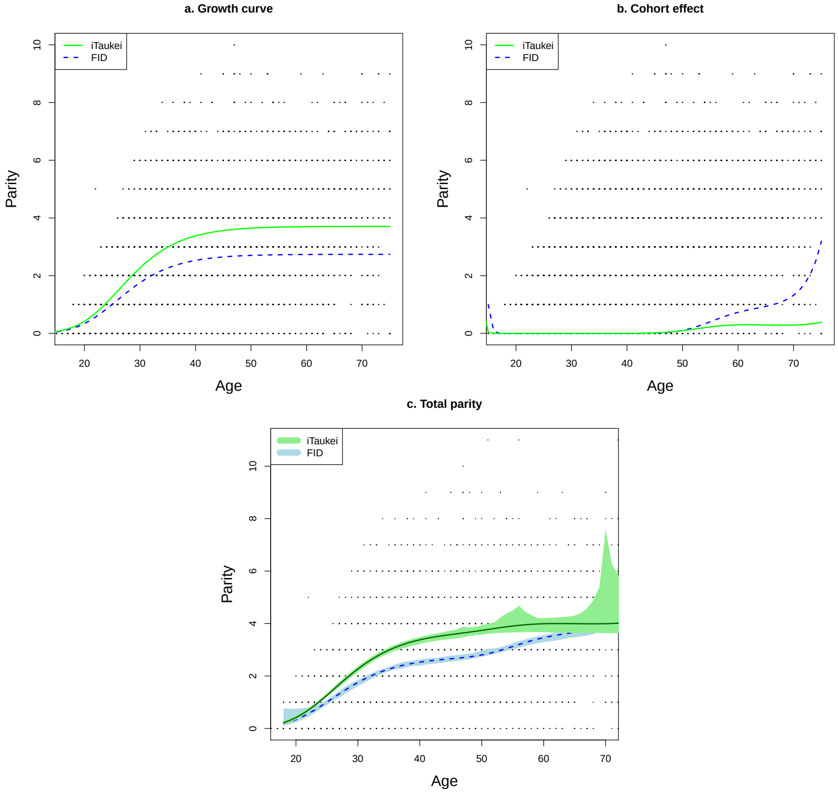


FIGURE 4. Fitted growth curves, cohort effects and total parity by ethnicity. 95% bootstrap confidence intervals are shown for total parity.

Acknowledgments: Special thanks to the Viseisei Sai Health Centre for provision of the data.

References

Richards, F. (1959). A Flexible Growth Function for Empirical Use. *Journal of Experimental Botany*, 10(2), 290-301.

Seniloli, K. (1994). Fertility and family planning in Fiji. *Espace, populations, sociétiés*, 12(2), 237-244.

A flexible approach for modelling a proportion response variable: Loss given default

Abu Hossain[1], Robert Rigby[1], Mikis Stasinopoulos[1], Marco Enea[2]

[1] STORM, London Metropolitan University,UK
[2] University of Palermo, Italy

E-mail for correspondence: a.hossain@londonmet.ac.uk

Abstract: Loss given default (LGD) is a proportion of a credit exposure that is lost if the obligor defaults on a loan. Response variable LGD contains values between 0 and 1 including both 0 and 1, where 0 means that the balance is fully recovered while 1 means total loss of exposure at default. This article addresses two alternative semi parametric approaches for modelling loss given default, which is measured on the interval [0,1]. The class of models are very flexible and can accommodate skewness and bimodal characteristics of LGD data. The dependence of the predictors of each of the parameters (of the proposed model distribution for LGD) on explanatory variables can be additive P- splines, regression trees or neural network models. The proposed models are applied to a loss given default data set and compared with current popular models.

Keywords: GAMLSS; generalised Tobit model; logit distribution.

1 Introduction

Loss given default is the key parameter for a bank's minimum regulatory capital requirement based on Basel II framework. Therefore modelling LGD is pivotal for financial regulators and retailers. However modelling LGD poses substantial challenges due to the bounded nature of LGD data and its unusual distribution, (see Bellotti and Crook(2012)). LGD values often lie on the interval [0,1] and the distribution tends to be bimodal with modes close to the end values.

Previous approaches for modelling (the distribution of) LGD on [0,1] include ordinary least squares, e.g. Qi and Yang (2009), fractional response

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).

regression (FRR), Papke and Wooldridge (1996), transformation models, e.g. Qi and Zhao (2011) and Li *et al.* (2014), the inflated beta model, Ospina and Ferrari (2010), a two step approach combining an ordinal logistic regression model and normal error model, Li *et al.* (2014), and Tobit models obtained by censoring a normal distribution or one or two shifted gamma distributions Li *et al.* (2014). In a very recent paper Hossain *et al.* (2016) proposed inflated `logitSST` and generalised Tobit models for the proportion response variable on the interval (0,1].

The purpose of this paper is to provide two flexible modelling approaches for a proportion response variable measured on the interval from 0 to 1, including both 0 and 1, i.e. range [0,1]. In the first approach a flexible distribution for Z with range $(-\infty, \infty)$ is transformed to Y with range (0,1), using an inverse logit transformation, $Y = 1/(1+e^{-Z})$, which is then inflated by including point probabilities for Y at 0 and 1. The second approach is a generalised Tobit model, in which a flexible distribution for Z on $(-\infty, \infty)$ is censored below 0 and above 1 to provide range $0 \leq Y \leq 1$ with probabilities at 0 and 1.

In practice, for each of the two modelling approaches, any available distribution on $(-\infty, \infty)$ within the `gamlss` package, Stasinopoulos and Rigby (2007), can be used for Z, for example the flexible four parameter skew exponential power (SEP), skew student t (SST), sinh arc-sinh ($SHASHo$) or bi-modal skew symmetric normal ($BSSN$) distribution, Hasan and El-Bassiouni (2016). In the `gamlss` package the dependence of the predictors of each of the parameters of the proposed model distributions for Y on explanatory variables can be linear, non-linear, non-parametric smooth functions, regression trees or neural network models. Note that Qi and Zhao (2011) and Li *et al.* (2014) found that regression tree and neural network models outperformed linear parametric models.

2 Models

2.1 Logit distribution

Any distribution on range $-\infty < Z < \infty$ can be transformed to a restrictive range $0 < Y < 1$ by using an inverse logit transformation $Y = 1/(1+e^{-Z})$. The distribution of Y is called a logit distribution. If Z has a four parameter distribution denoted D is general, i.e. $Z \sim D(\mu, \sigma, \nu, \tau)$, then the distribution of Y is called a logit D distribution denoted $Y \sim logitD(\mu, \sigma, \nu, \tau)$. For example if Z has a bi-modal skew symmetric normal distribution $Z \sim BSSN(\mu, \sigma, \nu, \tau)$ on $(-\infty, \infty)$, then Y has a $logitBSSN$ distribution, $Y \sim logitBSSN(\mu, \sigma, \nu, \tau)$ on $(0, 1)$. The $logitBSSN$ distribution is created using the function `gen.Family()` in `gamlss` which allows any `gamlss` distribution with range $(-\infty, \infty)$, (e.g. $BSSN$), to be transformed to a new `gamlss` distribution, (e.g. $logitBSSN$), with range $(0, 1)$.

2.2 LogitBSSN, inflated at 0 and 1

An inflated logit distribution is suitable for a proportion response variable on $0 \leq Y \leq 1$, that includes both 0 and 1. An inflated logit distribution is a mixture of a logit distribution for $0 < Y < 1$ and a Bernoulli distribution for Y at 0 or 1. The model includes three components: a discrete value 0 with probability p_0, a discrete value 1 with probability p_1 and a logit distribution on the unit interval $(0, 1)$ with probability $(1 - p_0 - p_1)$. For a general four parameter logit distribution, $logitD(\mu, \sigma, \nu, \tau)$, then the inflated logit distribution is denoted $Y \sim Inf.logitD(\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$ with mixed (continuous-discrete) probability (density) function given by

$$f_Y(y|\mu,\sigma,\nu,\tau,\xi_0,\xi_1) = \begin{cases} p_0 & \text{if } y = 0 \\ p_1 & \text{if } y = 1 \\ (1-p_0-p_1) f_W(y|\mu,\sigma,\nu,\tau) & \text{if } 0 < y < 1 \end{cases} \tag{1}$$

for $0 \leq y \leq 1$, where $W \sim logitD(\mu, \sigma, \nu, \tau)$ has a $logitD$ distribution, where $0 < p_0 < 1$, $0 < p_1 < 1$ and $0 < p_0 + p_1 < 1$. The parameters ξ_0 and ξ_1, are related to p_0 and p_1 by $\xi_0 = p_0/p_2$, $\xi_1 = p_1/p_2$. For example if $W \sim logitBSSN(\mu, \sigma, \nu, \tau)$ then Y has a inflated $logitBSSN$ distribution $Y \sim Inf.logitBSSN(\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$ with $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$, $\tau > 0$, $\xi_0 > 0$, and $\xi_1 > 0$.

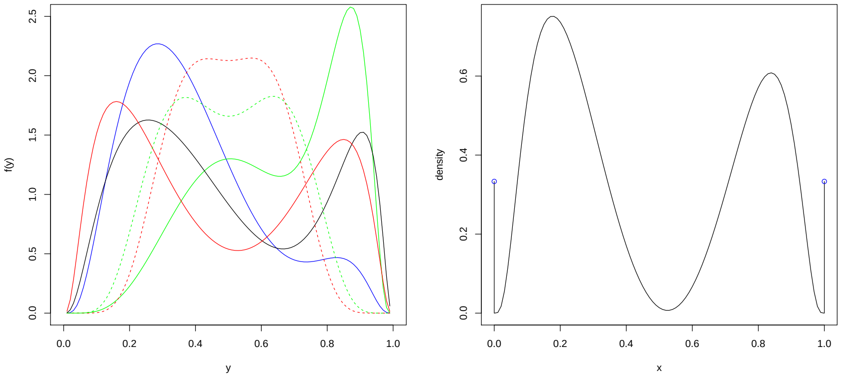


FIGURE 1. PDF of lositBSSN and InflogitBSSN

Model (1) can be fitted using a new function `gamlssinf0to1()`. The log likelihood function for the $Inf.logitBSSN$ model (1) is equal to the sum of the log likelihood functions of the $logitBSSN$ model and the multinomial model with three level (`MN3`). Hence the parameter sets (μ, σ, ν, τ) and (ξ_0, ξ_1) are 'information' orthogonal. Consequently model (1) can be fitted by fitting two models: a logitBSSN model for $0 < y < 1$ and an `MN3`(ξ_0, ξ_1)

for levels defined by $y = 0$, $y = 1$ and $0 < y < 1$. Figure 1 depicts various pdf plots of the logitBSSN and inflated logitBSSN distribution which portrays the model's ability to accommodate various shapes (i.e. skewness, kurtosis and bimodality).

The inflated logit distributions (e.g. $Inf.logitBSSN$) have the advantage of extra flexibility, in that the probabilities of Y at 0 and 1 are modelled independently of the distribution on (0,1), (e.g. $logitBSSN$), but with the cost of introducing extra parameters (ξ_0, ξ_1) into the model. Note that the logit transformation is sensitive to values of Y very close to 0 or 1. To avoid this problem it may be necessary for values $0 < Y < 1$ to be adjusted to $Y' = b + (1 - 2b)Y$, for a predetermined small constant b, prior to model fitting, see Li *et al.* (2014). Alternatively values very close to 0 and 1 can be adjusted to 0 and 1 respectively by $Y' = 0(\text{if } Y < b) + Y(\text{ if } b < Y < 1 - b) + 1(\text{if } Y > 1 - b)$.

2.3 Generalised type I Tobit model

The generalised Tobit model on [0,1] requires censoring below 0 and above 1 of a flexible model distribution on $(-\infty, \infty)$ for its positive probabilities at 0 and 1. Censoring refers to the transformation of observations outside the limiting interval to the border values. Here the values in the model distribution below 0 and above 1 are transformed to 0 and 1 respectively.

Let $Z \sim D(\mu, \sigma, \nu, \tau)$ be a flexible uncensored distribution on $(-\infty, \infty)$. Let $Y \sim Dic(\mu, \sigma, \nu, \tau)$ be the corresponding distribution left censored below 0 and right censored above 1 (called interval censoring, ic) with resulting range $[0, 1]$. Then

$$Y = \begin{cases} 0 & \text{if } Z \leq 0 \\ Z & \text{if } 0 \leq Z \leq 1 \\ 1 & \text{if } Z \geq 1. \end{cases}$$

Hence the (mixed continuous-discrete) probability (density) function of Y is given by

$$f_Y(y) = \begin{cases} P(Z \leq 0) & \text{if } y = 0 \\ f_Z(y) & \text{if } 0 < y < 1 \\ P(Z \geq 1) & \text{if } y = 1 \end{cases}$$

for $0 \leq y \leq 1$. In principle D can be any distribution on $(-\infty, \infty)$, for example the four parameter SEP, SST, $SHASHo$ or $BSSN$ distributions given in Section 1. Interval censoring is achieved using gamlss function gen.cens() in the gamlss package `gamlss.cens`.

In the generalised Tobit models the probabilities of Y at 0 and 1 are directly related to the distribution between 0 and 1 and so are less flexible, but the model is more concise (i.e. parsimonious) in that it has two less parameters. Also the Tobit model is not so sensitive to values of Y very close to 0 or 1.

3 Data

Loss Given Default (LGD) is the proportion of the exposure lost following a default. It is also called the severity of loss. The range of LGD is bounded on $[0, 1]$. The LGD value also tends to follow a bi-modal distribution. The motivating data example is the LGD values collected from one of the leading banks in the USA. The data frame comprises 7713 small business loan defaults between 2000 and 2007. In this analysis the response variable SEVERITY (LGD) is modelled using four covariates: Month-on-Books (MOB), hazard rate (hrate), year of origin (ORIGIN-YR) and year of default (DEFAULT-YR). The four explanatory variables are treated as quantitative variables.

4 Model Selection

The distributions on [0,1] considered for LGD were the beta inflated at 0 and 1 distribution (BEINF), Ospina and Ferrari (2010), together with two proposed models: the inflated logitBSSN and generalised type I Tobit ($BSSN_{ic}$) models and also the standard Tobit model which is based on interval censored normal distribution (NO_{ic}). Each distribution parameters was modelled using additive P-splines in the four explanatory variables. The results are reported in the Table 1. The inflated logitBSSN has by far the lowest AIC and SBC values among the four models. Note that the value of 12180 is added to all values of deviance, AIC and SBC for clear presentation.

TABLE 1. Comparison of Fitted Models

Method	Parameters	df	Deviance	AIC	SBC
logitBSSN	6	133	0	268	1199
GenTobit($BSSN_{ic}$)	4	84	13632	13800	14384
BEINF	4	86	6149	6321	6918
Tobit(NO_{ic})	2	41	18777	18859	19145

5 Conclusion

This paper proposes an inflated logit distribution and a generalised type I Tobit model for loss given default (LGD). Both models use the four parameter bi-modal $BSSN$ distribution (used in order to model the bimodality of the distribution of LGD). Flexible nonparametric P-splines were used to model the parameters of the distribution of the response variable using covariates. The dependence of each of the parameter of the two proposed models on explanatory variable can be replaced with linear, regression trees or neural network models. The proposed models were compared with the beta inflated and Tobit models. Based on the AIC and SBC criterion, the study concluded that the inflated logitBSSN provided the best fit to the loss given default data.

References

Bellotti, T.,Crook, J. A. (2012). Loss Given Default Models Incorporating Macroeconomic Variables for Credit Cards. *International Journal of Forecasting*, **28(1)**, 171 – 182.

Hasan, M.Y. and El-Bassiouni, M. Y. (2016). Bimodal skew symmetric normal distribution. *Communications in Statistics- Theory and Methods*, **45(5)**, 1527 – 1541.

Hossain, A., Rigby, R. A., Stasinopoulos, D. M. and Enea,M. (2015). Centile estimation for proportion response variable. *Statistics in Medicine*, **35(6)**, 801 – 960.

Li, P., Qi, M., Zhang, X., and Zhao, X. (2014). *Further Investigation of Parametric Loss Given Default Modelling*, Office of the Comptroller of the Currency, Economics Working Paper. **2014-2**,

Ospina, R. and Ferrari, S. L. P. (2010). Inflated beta distributions. *Statistical Papers*, **23**, 111 – 126.

Papke, L. E. and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, **11(6)**, 619 – 632.

Qi, M. and Yang, X. (2009). Loss given default of high loan-to-value residential mortgages. *Journal of Banking and Finance*, **33**, 788 – 799.

Qi, M. and Zhao, X. (2011). Comparison of modelling methods for loss given default. *Journal of Banking and Finance,*, **35(11)**, 2842 – 2855.

Rigby, R. A. and Stasinopoulos, D. M. (2005).Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics*, **54**, 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)**, 1 – 46.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, **26(1)**, 24 – 36.

Publication citability : a spatio-temporal modelling approach

Adrien Ickowicz[1]

[1] CSIRO, Hobart, Australia

E-mail for correspondence: adrien.ickowicz@csiro.au

Abstract: "An example of the objective approach is evaluation of the quality of scientific papers produced by the researchers by examination of **citation data**. Both of these methods are routinely used in the United Kingdom, and both have advantages and disadvantages." (OECD Report, 1997 - The evaluation of scientific research: selected experience. *Workshop on the Evaluation of Basic Research*, Paris, France).

Keywords: self-exciting process;citation modelling; thinned process

1 Introduction

A common idea is that the dynamic of a particular science area is observable through the publications and the citations related to the field. A research scientist will see his contribution be evaluated according to his publications, and his citations. This approach relies on the idea that the more an article is cited, the more it has contributed to the field (see the numerous algorithms dedicated to the ranking of papers based on the citation network solely - Zhou et al. 2015). While simple this hypothesis is the current working model because measurements/observations are easy under this framework. The reality is probably more complex. Several factors are to be considered if one wants to understand why an article is being cited, or why a particular science domain is rapidly expanding. Among those, we may mention the number of publications in the field, which can be linked to the number of journals and the number of researchers embracing the field. The amount of funding available is important too, and very often related to the previously mentioned factors.

It can be very difficult to assess the part played by these aforementioned factors in a science domain / research scientist success. In this article,

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016.

we introduce a new approach to understand these article / citation data and model them. Our framework is the network structure of the articles published in a science domain, where each vertex represents an article, and the edges between them can be many things: citations, common authors or institutions...
We propose an approach where publications related to the field are events of a point process which intensity can be modelled using a Hawkes process intensity. Then, we assume that for each article, its citations are thinned event from the publication process. Because the citation network is a dynamical structure, we ask two questions:

- How many vertices will appear during the next time period?
- Where (in the network) will these vertices be located?

The latter question is related to the edges of the networks, which are created according to the thinned process described above. It is a quite complex question though, as to answer it we have to model the creation of edges between nodes. We believe our model helps answering those.

2 Data

We work on a dataset of 492 articles selected on the "Web of Knowledge" using the keywords "approximate Bayesian computation", "ABc", and "Likelihood free". We only kept the articles published on or before 2013. The choice of these keywords has been made as this area of research is quite recent, and then can be qualified as an emerging area of research. For each of the articles, we recorded the title, the keywords, the authors, the date of publication, the journal of publication and the references.

2.1 Point process aspect - Network nodes

First, we assume that publications (or publication times - nodes of the network) are events of a Hawkes process with exponential decay. This representation has been chosen as it conveys two important ideas. Each new publication in the domain increases the chances of another publication in the same domain in the near future; the longer time since the last publication, the less likely a new publication event will occur.

2.2 Thinned process - Network edges

Then, we assume that the citation (edges of the network) times of a particular paper are following a thinned point process, where the parent process events are the publication times for all the papers in the same domain.
Citation events depend on a number of features, not being necessarily observable. For instance, an author is likely to cite its previous work. Or works

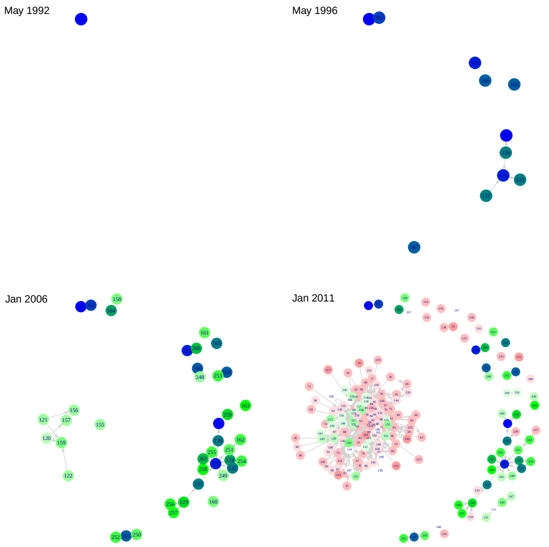


FIGURE 1. Evolution of the "ABc" publication network. Four dates are displayed: May 1992, September 1996, January 2006 and January 2011. We are very much in the presence of a spatio-temporal process, where the spatial dimension is embedded in the network manifold.

from colleagues part of the same institution. Or previous articles published in the same journal. And of course, articles in the same domain are likely to be cited. For the sake of completeness of the literature review. Our aim is to include these elements in the thinning model.
Other graph or networks can be defined, by changing the nature of the edge between two vertices:

- Authorship network. That is, two articles are connected if the have at least one author in common.
- Institutional network. Two articles are connected if the have at least one institution in common.
- Two articles are connected if they are published in the same journal.

These networks are also likely to be complementary. That is, the citation network is a simple linear transformation of the three networks listed above. We analysed the networks of the top 20 most cited articles in the data we collected. Out of these networks, only one connection (citation) cannot be explained by either a common author, a common institution or a common journal. It is then important for these features to be included in the thinned process model.

3 Modelling

3.1 Publications

Our model has two sets of observations, hence two components. In the first component, the events of interest are the publication times of articles related to the field of interest. This model can be used as a proxy to the dynamic of the field. It is modelled using a Hawkes process, following the assumptions that the intensity of the process increase whenever a publication in the domain is made. Its intensity $\lambda(t)$ is defined such that,

$$\lambda(t) \quad = \quad \mu(t) + \int_T \kappa(t-s)dH(s), \tag{1}$$

where $\mu(t)$ is a deterministic baseline intensity, and $\kappa(t)$ is a kernel function expressing the influence (usually positive) of past events on the current value of the intensity process (see Hawkes, 1971). In this article, we will more specifically be using fully parametric intensity functions of the form,

$$\lambda(t) \quad = \quad a + \sum_{t_k<t} \beta e^{-\delta(t-t_k)} \tag{2}$$

which describe a Hawkes process also known as exponentially decaying self-exciting point process (see Dassios and Zhao, 2013). In this formula, a describes the baseline (or long-term) intensity, δ the rate of the exponential decay, and β the influence of an event on the intensity.

3.2 Citations

In the second component, we assume that the citation process is a sub-process of the publication process (a thinned process with a given thinning probability). Let N_t be the number of new articles (nodes) at time t, $I_t = \{1, ..., \sum_{\tau<t} N_\tau\}$ be the set of existing articles, and $J_t = \{1, ..., N_t\}$ be the set of appearing articles at time t. Then our observations are $c_{i,j,t}$ for $i \in I_t, j \in J_t, t \in \{1, ..., T\}$ where $c_{i,j,t} = 1$ if article j cites article i. For every observation $c_{i,j,t}$ we have a set of predictors $X_{i,j}$ which we use to model $p_{i,j}(t) = p(c_{i,j,t} = 1|X_{i,j})$. We define this probability as the thinning probability applied to the publication process defined in the previous section. We include in $p_{i,j}(t)$ all the information related to article i, such as its author(s), institution(s), journal it has been published in. We expect that probability to be time-varying, in particular we assume that the time since publication and the time since last citation are important factors. Its mathematical expression is:

$$\text{logit}(p_{i,j}(t)) \quad = \quad X_{i,j}\beta + \delta_1 \min_k (t-t_k)^+ + \delta_2 (t-t_i)^+ \tag{3}$$

where t_k is the k-th citation event time, t_i is the publication (birth) time, X are covariates (journal name, number of references, ...).

For a given article i, the citation process is then an inhomogeneous Poisson process with intensity $\mu_i(t) = \sum_j p_{i,j}(t)\lambda_j(t)$, where $\lambda_j(t)$ is the intensity of the process for the articles which share the same features as article j.

4 Prediction

Based on this modelling option, we can answer the questions asked int the introduction using the prediction properties of the models.
How many vertices will appear over the next period? By itself, this question is technically straightforward to answer, using the Hawkes process model to simulate events in the next period, allowing us to then predict $\hat{N}_t$.
Where (in the network) will these vertices be located? To predict the presence of an edge between two vertices, we calculate the probabilities $\hat{p}_{i,j}(t), i \in I_t, j \in J_t$. For a given new node j^*, we have a set of possible edges with probability $p_{i,j^*}(t)$ given by Eq. 3. It is a partial answer though, as it does not tell us how many edges are connected to this new vertex. To answer that, we propose two approaches. In the first one, we use a ROC curve to identify the best threshold, and create an edge for probabilities above that threshold. For a given new node j^*, we will then a a set of edges, and an estimated location. In the second approach, we use the estimated probabilities to simulate the creation of edges between the new node and the old ones. For each simulation, we will have an estimated location for j^*. With enough simulations we can draw a map of likely locations for j^*.
The accuracy of the answer to the second question depends on the level of details we can provide in the prediction of the first question. For example, predicting that 15 new articles will be published over the next 4 months provides less details than predicting 6 articles in Wiley, 5 articles in Oxford Press and 4 articles in Springer. But then, the uncertainty in these figures will be higher. A trade-off is then necessary.

5 Results

5.1 Fitting

Using this modelling approach, we can identify features that favour citation events. For instance, the following features favoured citations in the analysed dataset:
- *Common journal, themes or authors, published in proceedings, writing a review article.*

The following features, on the other hand, disfavoured citations:
- *Time since last citation, writing a long article, publishing later in the calendar year.*

5.2 Prediction

If we use the simplest model, that is we do not add features to the intensity estimation of the Hawkes process, the prediction for the number of article published in 2012 is 144 with a 95% confidence interval $[94, 195]$. The actual number of article published in 2012 was 128. And using the apporaches described above, we have the estimated locations for a new node j^* in Figure 2.

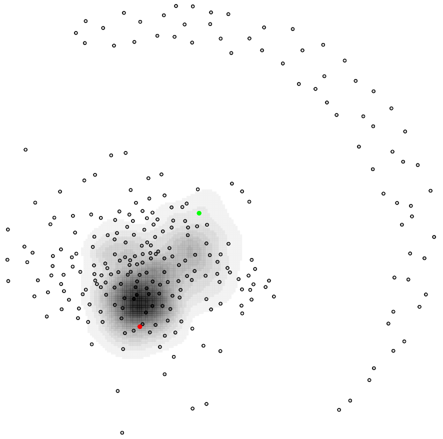


FIGURE 2. Map of potential new publication (node) location. Grey levels for map of likely location, green dot for the true location, red dot for the ROC-estimated location.

5.3 Conclusion

We have presented a spatio-temporal approach to model citations and publications, and tried two answer two questions. The second one though is only partially answered, as we only could provide spatial estimation for a single new node, and not a set of them.

References

Dassios, A. and Zhao, H. (1994). Exact simulation of Hawkes process with exponentially decaying intensity. *Electronic Communications in Probability*, **18**, 1 – 13.

Hawkes, A. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58**, 83 – 90.

Zhou, J., Zheng, A., Fan, Y., and Di, Z. (2015). Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, 1 – 12.

A non-homogeneous dynamic Bayesian network model with partially sequentially coupled network parameters

Mahdi Shafiee Kamalabad[1], Marco Grzegorczyk [1]

[1] Johann Bernoulli Institute (JBI), Rijksuniversiteit Groningen, The Netherlands

E-mail for correspondence: m.shafiee.kamalabad@rug.nl

Abstract: We propose a novel non-homogeneous dynamic Bayesian network (NH-DBN) model with partially sequentially coupled network parameters. The idea is to segment a time series of network data using a multiple changepoint process, and to model the data in each segment by linear Bayesian regression models. Our new model is an extension of a recently proposed Bayesian network model with sequentially coupled network parameters. The earlier model, which we refer to as the fully sequentially coupled NH-DBN model, assumes that all segments are coupled with the same coupling strength between segments. Our new partially coupled NH-DBN model infers for each segment whether it is coupled to (or uncoupled from) the preceding one. Our new model can thus be seen as a consensus model between (i) an uncoupled NH-DBN model without any network parameter coupling and (ii) a fully sequentially coupled NH-DBN model.

Keywords: Non-homogenous dynamic Bayesian networks; Partially sequentially coupled network parameters

1 Introduction

In systems biology non-homogeneous dynamic Bayesian networks (NH-DBNs) have become popular tools for modelling cellular regulatory processes. The idea is to use a multiple changepoint process to divide the observed temporal network data into disjunct time segments, and to model the data within each segment by linear Bayesian regression models. For most cellular processes it is reasonable to assume that the network structure (i.e. the regulator sets or regressors of each network node) do not change over time, while the network parameters (regression coefficients) are time-dependent. Thus, except for identical regulator sets there is no information-

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).

sharing among segments. To allow for information-sharing with respect to the network parameters a NH-DBN with sequentially coupled network parameters was proposed in Grzegorczyk and Husmeier (2012). The idea is to allow for information sharing between neighbouring segments by using the posterior expectation of the network parameters as prior expectation for the consecutive segment. Node-specific coupling parameters regulate the variance of the parameter priors and so the strength of coupling. A shortcoming of this approach is that all neighbouring segments are coupled with the same coupling parameter. Low coupling parameters yield peaked (informative) and high coupling parameters yield vague (uninformative) parameter priors for the subsequent segments. Thus, for networks with substantially varying parameters, information coupling can be counter-productive, as uncoupling can only be reached by making the network parameter priors vague. In this paper we address this shortcoming. We extend the model from Grzegorczyk and Husmeier (2012) by introducing an option to uncouple. This gives the new partially sequentially coupled NH-DBN model. In the new model we infer for each individual segment whether it is coupled to (or uncoupled from) the preceding one. Our model regularizes between the two extreme cases: An uncoupled NH-DBN with independent parameters (= every segment is uncoupled) and a NH-DBN with sequentially coupled parameters (= every segment is coupled to the preceding one).

2 Methodological details

In this representation we focus on one single network node y which takes the role of the regulatee (response) in a segment-wise linear Bayesian regression model. Given a set of k regulators or regressor variables $\pi = \{X_1, \ldots, X_k\}$, we assume that the temporal data can be divided into H segments with different regression coefficients. Let $\mathbf{y}_h$ be the vector of the response values and $\mathbf{X}_{\pi,h}$ be the design matrix for segment h, where each $\mathbf{X}_{\pi,h}$ includes a first column of 1's for the intercept. For $h = 1, \ldots, H$ we have:

$$\mathbf{y}_h \sim \mathcal{N}(\mathbf{X}_{\pi,h}\mathbf{w}_h, \sigma^2\mathbf{I}) \tag{1}$$

where σ^2 is the noise variance parameter, with $\sigma^{-2} \sim GAM(0.01, 0.01)$, and $\mathbf{w}_h$ $(h = 1, \ldots, H)$ are the $(k+1)$-dimensional segment-specific regression coefficient vectors, on which we impose the novel priors:

$$P(\mathbf{w}_h) = \mathcal{N}(v_h \cdot \mathbf{m}_{h-1}, \lambda^{v_h}\delta^{1-v_h}\sigma^2\mathbf{I}) \tag{2}$$

We set $v_1 = 0$, $\mathbf{m}_0 = \mathbf{0}$, $\lambda^{-1} \sim GAM(3,3)$ and $\delta^{-1} \sim GAM(2, 0.2)$, as in Grzegorczyk and Husmeier (2012). The newly introduced indicator variables $v_h \in \{0,1\}$ indicate whether segment $h \geq 2$ is coupled to segment $h-1$ $(v_h = 1)$ or not, and $\mathbf{m}_h$ $(h \geq 1)$ is the posterior expectation of $\mathbf{w}_h$:

$$\mathbf{m}_h = \left(\lambda^{-v_h}\delta^{-(1-v_h)}\mathbf{I} + \mathbf{X}_{\pi,h}^T\mathbf{X}_{\pi,h}\right)^{-1}\left(\lambda^{-v_h}\delta^{-(1-v_h)}\mathbf{m}_{h-1} + \mathbf{X}_{\pi,h}^T\mathbf{y}_h\right)$$

We assume the new indicator variables $v_2, \ldots, v_H$ to be Bernoulli distributed, $v_h \sim BER(p)$, where $p \in [0, 1]$, can either be set fixed (e.g. $p = 0.5$) or be assumed to be Beta distributed, $p \sim BETA(a, b)$.

- $p = 0$ yields $v_h = 0$ for all h, so that independent priors, $P(\mathbf{w}_h) = \mathcal{N}(\mathbf{0}, \delta\sigma^2\mathbf{I})$, are used for the segments. This refers to an **uncoupled NH-DBN** without information sharing between the parameters $\mathbf{w}_h$.
- $p = 1$ yields $v_h = 1$ $(h \geq 2)$, and thus gives the priors from Grzegorczyk and Husmeier (2012), $P(\mathbf{w}_h) = \mathcal{N}(\mathbf{m}_{h-1}, \lambda\sigma^2\mathbf{I})$ for $h \geq 2$. The model is then the **(fully sequentially) coupled NH-DBN**, proposed in Grzegorczyk and Husmeier (2012).
- Our new **partially (sequentially) coupled NH-DBN** infers the variables v_h $(h \geq 2)$ from the data and therefore tries to find the right trade-off between the uncoupled NH-DBN and the coupled NH-DBN.

For a network with n nodes we apply the segment-wise linear regression model to each node y_i $(i = 1, \ldots, n)$ separately, and the potential regulator sets of y_i are all subsets of the other $n - 1$ nodes, symbolically $\pi_i \subset \{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n\}$. The system of parent sets $G := \{\pi_1, \ldots, \pi_n\}$ describes a network G: there is an edge $y_j \rightarrow y_i$ if and only if $y_j \in \pi_i$.

3 Simulation study

We generate synthetic data for a yeast network with 5 genes: $GAL80$, $GAL4$, $SWIS$, $CBF1$, and $ASH1$. We consider each gene to be a response variable y and generate data using the segment-wise linear model from Eq. (1) with the regulator sets: $\pi_{ASH} = \{SWIS\}$, $\pi_{SWIS} = \{GAL4\}$, $\pi_{GAL80} = \{GAL4, SWIS\}$, $\pi_{GAL4} = \{GAL80, CBF1\}$, and $\pi_{CBF1} = \{ASH, SWIS\}$. In a dynamic Bayesian network all interactions are subject to a time delay, so that the segment-specific design matrices are built from the values of the regulators at the preceding time points. We perform two simulations studies.

Simulation study 1: We assume that there are $H = 4$ segments, divided by three changepoints, and that each segment has 10 time points. We sample the segment-specific regression coefficients w_h in Eq. (1) as follows: For segment $h = 1$ we draw samples from a standard multivariate Gaussian distribution and we re-normalise the sampled vectors to Euclidean norm 1; this gives w_1. For the subsequent segments we sample again from standard multivariate Gaussian distributions and we re-normalise the sampled vectors to Euclidean norm $\kappa = 0.1$; this yields $w_h^\star$. If segment h is coupled $(v_h = 1)$ we add $w_h^\star$ to the vector of the preceding segment $h - 1$, $w_h = w_{h-1} + w_h^\star$, so that we obtain very similar parameters. If segment h is uncoupled $(v_h = 0)$, we compute: $w_h = (-1) \cdot w_{h-1} + w_h^\star$, so that we obtain dissimilar parameters. We distinguish three scenarios (S1)-(S3):

- **(S1) uncoupled data** $v_h = 0$ for all h,
- **(S2) coupled data** $v_h = 1$ for $h \geq 2$
- **(S3) partially coupled data** $(v_1, \ldots, v_4) = (0, 1, 0, 1)$

For (S1)-(S3) we consider 5 noise levels $\sigma = 0.1, 0.2, 0.4, 0.8, 1.6$. This gives 15 combinations and for each we generate 25 independent data instantiations. In this study we assume the three changepoints locations (i.e. the segmentation) to be known and fixed.

Simulation study 2: We assume that there are $H = 2$ segments, divided by a changepoint, and that each segments has 10 time points. For each gene y of the yeast network we draw an unbiased coin to decide whether the second segment is coupled or not; i.e. we draw $v_2 \in \{0, 1\}$. For $v_2 = 0$ y's regulation changes drastically, $w_2 = (-1) \cdot w_1 + w_2^\star$, for $v_2 = 0$ its regulation stays similar, $w_2 = w_1 + w_2^\star$, where w_1 and $w_2^\star$ are random samples from standard multivariate Gaussian distributions, re-normalised to Euclidean norm 1 and 0.1, respectively. This refers to a partially coupled scenario. In each data set each gene has a probability of 0.5 to be coupled ($v_2 = 1$) and a probability of 0.5 to be uncoupled ($v_2 = 0$). We consider 6 noise levels $\sigma = 0.1, 0.2, 0.4, 0.8, 1.6, 3.2$ and for each σ we generate 25 independent data instantiations. Different from the first study, we assume the changepoint location (i.e. the segmentation) to be unknown, so that the changepoint(s) have be inferred from the data.

4 Simulation Details

We use a multiple changepoint process to segment the data into H segments, but unlike Grzegorczyk and Husmeier (2012) we assume that there is a network wide changepoint set which applies to the complete network G; i.e. all nodes $y_1, \ldots, y_n$ share the same segmentation.

For inference we extend the partially collapsed Gibbs sampler from Grzegorczyk and Husmeier (2012), which samples the hyperparameters δ, λ, σ, the regulator sets π, and the changepoint set (if unknown) from the data. For our model we include an additional Metropolis-Hastings move to change the values of the variables v_h. The new move randomly selects one v_h $(h \geq 2)$ and proposes to set it to $1 - v_h$. The acceptance probability can then be computed with the Metropolis-Hastings criterion, where the Hastings-ratio is equal to 1. The goal of our study is to infer the 8 interactions of the yeast network. We perform Markov Chain Monte Carlo (MCMC) simulations on each data set to generate samples of networks $G_t = \{\pi_{1,t}, \ldots, \pi_{n,t}\}_{t=1,\ldots,T}$. We average accross those networks to obtain for each individual edge $j \rightarrow i$ $(j, i \in \{1, \ldots n\} : j \neq i)$ a marginal posterior probability $\hat{e}_{j,i} = \frac{1}{T}\sum_{t=1}^{T} I_{j\rightarrow i}(G_t)$, where $I_{j\rightarrow i}(G_t) = 1$ if $j \in \pi_i$, and $I_{j\rightarrow i}(G_t) = 0$ otherwise. As $\hat{e}_{j,i} \in [0, 1]$ and the true interactions are

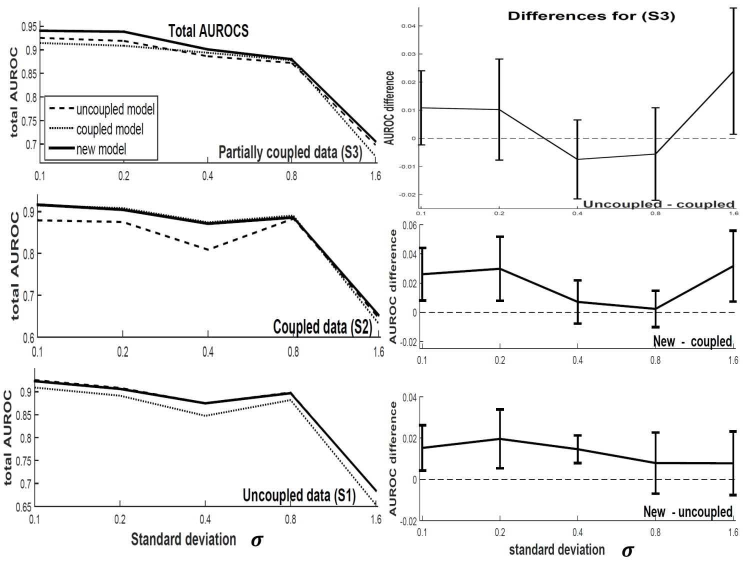


FIGURE 1. **Results of simulation study no 1: Known segmentation.** Network reconstruction accuracy for simulated yeast network data. There are $H = 4$ segments with 10 observations each, and the changepoint locations are assumed to be known. Three scenarios (coupled, uncoupled and partially coupled data) are distinguished. *Left panel*: The average total AUROC values for the three scenarios (S1)-(S3). *Right panel*: AUROC differences between the models for partially coupled data (S3), with errorbars representing t-test confidence intervals.

known, $e_{i,j} \in \{0, 1\}$, we can quantify the network reconstruction accuracy of the three NH-DBN models for each individual data set in terms of three AUROC values.

5 Results

The results of the two studies are shown in FIGURES 1-2. The results of the first study, shown in FIGURE 1, suggest that the new partially coupled NH-DBN model is never inferior to the other two models. For uncoupled data (S1) it performs as well as the uncoupled NH-DBN, and for coupled data (S2) it performs as well as the coupled NH-DBN. For partially coupled data (S3) and moderate noise levels σ, the t-test confidence intervals for the AUROC differences in the right panel of FIGURE 1 show that the partially coupled NH-DBN is significantly superior to the two competing NH-DBNs. The results of the second study are shown in FIGURE 2. For partially coupled data with an unknown changepoint the partially coupled model yields significantly better AUROC values than the other two approaches, see bottom left and top right panel of FIGURE 2. The uncoupled model appears to perform slightly better than the fully coupled NH-DBN for

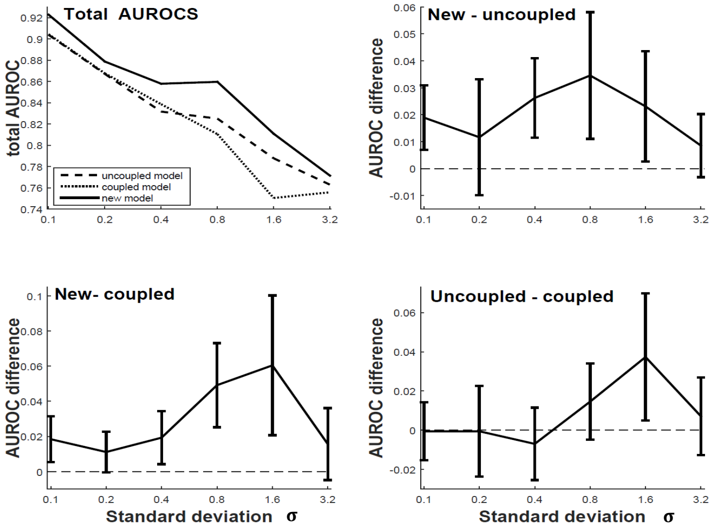


FIGURE 2. **Results of simulation study no. 2: Unknown segmentation.** Network reconstruction accuracy for simulated yeast network data. There are $H = 2$ segments with 10 observations each, and the changepoint(s) are unknown and have to be inferred from the data. The data are partially coupled. *Top left panel*: The average total AUROC values of the three NH-DBN models. *Lower left and right panels*: Pairwise AUROC differences between the three NH-DBN models, with errorbars representing t-test confidence intervals.

$\sigma \geq 0.8$, see lower right panel. However, the differences are significant only for $\sigma = 1.6$.

6 Conclusion

Our results suggest that the partially sequentially coupled NH-DBN, proposed here, is a promising consensus model between the standard uncoupled NH-DBN and the fully sequentially coupled NH-DBN. The new model appears to infer correctly from the data whether network parameters are coupled or not. The new partially coupled NH-DBN was never inferior to the gold-standard NH-DBN model, and for partially coupled data the new model performed significantly better than the two competing NH-DBNs.

References

Grzegorczyk, M., and Husmeier, D. (2012). *A non-homogeneous dynamic Bayesian network with sequentially coupled interaction parameters for applications in systems and synthetic biology.* Statistical Applications in Genetics and Molecular Biology, 11(7), online article.

Median bias reduction of maximum likelihood estimates

Euloge C. Kenne Pagui[1], Alessandra Salvan[1], Nicola Sartori[1]

[1] Department of Statistical Sciences, University of Padova

E-mail for correspondence: kenne@stat.unipd.it

Abstract: For a scalar component of interest of a multidimensional parameter, we propose a median modification of the profile score equation whose solution respects equivariance under reparameterizations. As Firth's (1993) implicit method for bias reduction, the new estimator does not depend on the maximum likelihood estimator and is effective in preventing infinite estimates. We also extend the approach to a multidimensional parameter of interest. An application and a simulation to a generalized linear model for binary data compare the proposed method with maximum likelihood and implicit bias reduction.

Keywords: Cornish-Fisher; Infinite estimates; Median unbiased; Modified score.

1 Introduction

Most available corrections of the maximum likelihood estimator or of the score estimating equation are aimed at first-order bias adjustment. See Kosmidis (2014) for an up to date review. Although bias correction of the maximum likelihood estimator depends on the chosen parameterization, implicit methods for bias correction following Firth (1993) exhibit some relevant advantages. In particular, the modified estimating equation does not depend explicitely on the maximum likelihood estimator and has been found to overcome infinite estimate problems that may arise with positive probability mainly, but not only, in models for discrete or categorical data. For a scalar parameter of interest, we propose a median modification of the profile score equation whose solution i) is second-order median unbiased; ii) respects equivariance under reparameterizations; iii) does not depend on the overall maximum likelihood estimator and is effective in preventing infinite estimates. The modification is obtained by considering the median as a centering index for the profile score and defining a new estimating function by subtracting from the profile score its approximate median. The approach is extended to a vector parameter, maintaining componentwise equivariance and second-order median centering. An application and a simulation in a generalized linear model for binary data compare the proposed method with maximum likelihood and Firth's (1993) bias reduction. The new method proves to be remarkably accurate in achieving median centering of the estimator.

2 Modified score

For data y, consider a regular model with probability mass function $p_Y(y;\theta)$ and parameter θ with real components $(\theta_1,\dots,\theta_p)$. Let $\ell(\theta)$ be the log likelihood and $U(\theta)=\partial\ell(\theta)/\partial\theta$, the score function. The maximum likelihood estimator $\hat\theta$ is a solution of $U(\theta)=0$. We assume that the covariance matrix of $U(\theta)$, $i(\theta)$, i.e. Fisher information, and third-order cumulants of $U(\theta)$ are finite and of order $O(n)$, where n is the sample size or, more generally, an index of information in the data.
Let $\theta=(\psi,\lambda)$, with ψ a scalar parameter of interest. We denote by $\ell_P(\psi)=\ell(\psi,\hat\lambda_\psi)$ the profile log likelihood for ψ, where $\hat\lambda_\psi$ is the maximum likelihood estimate of λ for a given value of ψ. The profile score is $U_P(\psi)=\partial\ell_P(\psi)/\partial\psi$. Using Cornish-Fisher expansion (see e.g. Pace and Salvan, 1997, Section 10.6), the following asymptotic expansion holds for the median of the profile score in the continuous case

$$Me_\theta\left\{U_P(\psi)\right\}=\kappa_{1\psi}-\frac{1}{6}\frac{\kappa_{3\psi}}{\kappa_{2\psi}}+O(n^{-1})\,. \tag{1}$$

In (1), $\kappa_{j\psi}$, $j=1,2,3$, are the first three cumulants of $U_P(\psi)$, possibly replaced by suitable expansions.
A modified profile score can thus be defined by equating $U_P(\psi)$ to the leading term of its median, giving

$$\tilde U_P(\psi)=U_P(\psi)-\kappa_{1\psi}+\frac{1}{6}\frac{\kappa_{3\psi}}{\kappa_{2\psi}}\,. \tag{2}$$

In (2), the modification term is of order $O(1)$ and only the leading terms of asymototic expansions for $\kappa_{j\psi}$, $j=1,2,3$, are needed to ensure that $\tilde U_P(\psi)$ has median zero with error of order $O(n^{-1})$. The needed expansions can be obtained using results in McCullagh and Tibshirani (1990) and Barndorff-Nielsen and Cox (1989, Chapter 7) and are evaluated at $(\psi,\hat\lambda_\psi)$.
Let $\tilde\psi$ be the estimator defined as solution of $\tilde U_P(\psi)=0$. The modified profile score has median zero with error of order $O(n^{-1})$, i.e.

$$P_\theta\left\{\tilde U_P(\psi)\le 0\right\}=\frac{1}{2}+O(n^{-1})\,. \tag{3}$$

If $\tilde{U}_P(\psi)$ is monotone decreasing in ψ, the events $\tilde{U}_P(\psi) \leq 0$ and $\tilde{\psi} \leq \psi$ are equivalent so that, from (3), $\tilde{\psi}$ will be median unbiased with error of order $O(n^{-1})$, i.e. second-order median unbiased. On the other hand, the asymptotic distribution of $\tilde{\psi}$ is the same as that of $\hat{\psi}$.
Under interest-respecting reparameterizations $\tilde{U}_P(\psi)$ transforms as a covariant tensor of order one, so that $\tilde{\psi}$ behaves equivariantly, as does $\hat{\psi}$.
Modification (2) can be used also in the discrete case, ignoring the oscillatory terms in the Cornish-Fisher expansion.
A limitation of (2) is that it allows estimation of a scalar component of θ at a time and requires constrained maximum likelihood estimates of the remaining parameters. An extension for joint estimation of θ is proposed by Kenne Pagui et al. (2016). Let indices $a, b \ldots$ take values in $\{1, \ldots, p\} \setminus \{r\}$ with summation understood when they are repeated. Moreover, let U_r be a generic component of $U(\theta)$, i_{rs} be a generic entry of $i(\theta)$ and ν^{ab} be a generic entry of the inverse of the matrix with entries i_{ab}. The modified score vector $\tilde{U}(\theta)$ has components

$$\tilde{U}_r = U_r - \gamma_{ra} U_a - \kappa_{1r} + \frac{1}{6} \frac{\kappa_{3r}}{\kappa_{2r}} \,, \qquad r = 1, \ldots, p \,, \tag{4}$$

where $\gamma_a^r = i_{rb} \nu^{ab}$ and κ_{jr}, $j = 1, 2, 3$, are as in (2) referred to $\psi = \theta_r$. Then, the joint estimator $\tilde{\theta}$ is defined as solution of $\tilde{U}(\theta) = 0$.
Let $\tilde{\theta}_r$ be the r-th component of $\tilde{\theta}$ and denote here by $\tilde{\theta}_r^P$ the solution of $\tilde{U}_P(\theta_r) = 0$, with $\tilde{U}_P(\cdot)$ given by (2). Kenne Pagui et al. (2016) show that, in a regular model,

$$\tilde{\theta}_r - \tilde{\theta}_r^P = O_p(n^{-3/2}) \,,$$

$r = 1, \ldots, p$, thus achieving componentwise second-order median unbiasedness.

3 An application to binary regression

We consider the endometrial cancer grade dataset analyzed e.g. in Agresti (2015, Section 5.7.1). The goal of the study was to evaluate the relationship between the histology of the endometrium of 79 patients and three risk factors: neovasculation (NV), pulsatility index of arteria uterina (PI) and endometrium height (EH). Logistic regression maximum likelihood estimation leads to infinite maximum likelihood estimate of the effect of NV due to the quasicomplete separation problem. Let us consider first the coefficient of NV as the parameter of interest while the remaining parameters are treated as nuisance. The estimate from (2) is equal to 3.883, while Firth's (1993) bias reduced estimate is equal to 2.929. The corresponding standard errors are 2.407 and 1.551, respectively. The joint estimate $\tilde{\beta}$ of the four components of the parameter β, obtained using (4), are reported in Table 1, together with $\hat{\beta}$ and the implicit bias reduced estimate $\hat{\beta}^*$. A

TABLE 1. Endometrial cancer study: logistic regression estimates (s.e.).

	intercept	NV	PI	EH
$\hat\beta$	4.305 (1.637)	$+\infty$ ($+\infty$)	-0.042 (0.044)	-2.903 (0.846)
$\hat\beta^*$	3.775 (1.489)	2.929 (1.551)	-0.035 (0.040)	-2.604 (0.776)
$\tilde\beta$	3.969 (1.552)	3.869 (2.298)	-0.039 (0.042)	-2.708 (0.803)

simulation study has been performed in order to evaluate the properties of estimators of β in terms of percentage probability of underestimation (PU%), median absolute error (MAE), bias (B), root mean squared error (RMSE) and coverage of Wald-type confidence intervals (Coverage). Table 2 shows the results obtained with 10,000 replications, covariates fixed at the observed values and $\beta = (1.5, 2, 0, -2)$. The new method proves to be remarkably accurate in achieving median centering of all components of the estimator, as indicated by PU%. We also note that in 684 samples out of 10,000 the maximum likelihood estimate of the coefficient of NV is infinite, while both $\hat\beta^*$ and $\tilde\beta$ are always finite. Coverage for the three methods is rather similar, although coverage probabilities for maximum likelihood should be judged with caution since samples with infinite estimates are excluded. Similar results have been observed in unreported simulations with a probit model.

TABLE 2. Endometrial cancer study. Simulation of estimates of the regression coefficients with logistic link: PU%, percentage of underestimation; MAE, median absolute error; B, bias; RMSE, root mean squared error. For maximum likelihood, B, RMSE and coverage are conditional upon finiteness of the estimates.

	PU%	MAE	B	RMSE	Coverage (%)
$\hat\beta$	45.1	0.97	0.29	1.60	95.8
	43.0	0.66	0.12	0.90	97.4
	51.0	0.03	0.00	0.04	95.0
	56.0	0.57	-0.26	1.02	96.0
$\hat\beta^*$	52.6	0.86	0.00	1.38	96.6
	53.0	0.56	0.02	0.90	97.4
	49.6	0.02	0.00	0.04	96.3
	44.4	0.52	0.01	0.83	94.8
$\tilde\beta$	50.1	0.90	0.09	1.46	96.4
	49.7	0.59	0.15	1.07	97.5
	50.7	0.02	0.00	0.04	96.1
	49.6	0.52	-0.10	0.89	95.8

References

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* New York: Wiley.

Barndorff-Nielsen, O.B.N. and Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics.* London: Chapman & Hall.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27 – 38.

Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2016). Median bias reduction of maximum likelihood estimates. *http://arxiv.org/abs/1604.04768.*

Kosmidis, I. (2014). Bias in parametric estimation: reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics*, **6**, 185 – 196.

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society Series B*, **52**, 325 – 344.

Pace, L., and Salvan, A. (1997). *Principles of Statistical Inference from a neo-Fisherian Perspective.* Singapore: World Scientific.

Penalized likelihood inference in meta-regression

Ioannis Kosmidis[1], Annamaria Guolo[2], Cristiano Varin[3]

[1] Department of Statistical Science, University College London, London, United Kingdom
[2] Department of Statistical Sciences, University of Padova, Padova, Italy
[3] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy

E-mail for correspondence: i.kosmidis@ucl.ac.uk

Abstract: Random-effects models are frequently used to synthesise information from different studies in meta-analysis. While likelihood-based inference is attractive both in terms of limiting properties and in terms of implementation, its application in random-effects meta-analysis may result in misleading conclusions, especially when the number of studies is small to moderate. The current paper shows how methodology that reduces the asymptotic bias of the maximum likelihood estimator of the variance component can also substantially improve inference about the mean effect size. The results are derived for the more general framework of random-effects meta-regression, which allows the mean effect size to vary with study-specific covariates.

Keywords: Bias reduction; Heterogeneity; Meta-analysis; Penalized likelihood; Random effect

1 Meta-analysis and meta-regression

Meta-analysis is a widely applicable approach to combine information from different studies about a common effect of interest. One major topic of debate in meta-analysis is how to best deal with the heterogeneity across studies. A large body of applications has resorted to using the formulation described in DerSimonian & Laird (1986), which accounts for the between-study heterogeneity via a random-effects specification.

Suppose there are K studies about a common effect of interest, each of them providing pairs of summary measures $(y_i, \hat{\sigma}_i^2)$, where y_i is the study-specific estimate of the effect, and $\hat{\sigma}_i^2$ is the associated estimation variance

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016.

$(i = 1, \ldots, K)$. In some situations, the pairs $(y_i, \hat{\sigma}_i^2)$ may be accompanied by study-specific covariates $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$, which describe the heterogeneity across studies. In the meta-analysis literature, it is usually assumed that the within-study variances $\hat{\sigma}_i^2$ are estimated well enough to be considered as known and equal to the values reported in each study. Under this assumption, the random-effects meta-regression model postulates that i) $y_1, \ldots, y_K$ are realizations of random variables $Y_1, \ldots, Y_K$, respectively, which are independent conditionally on independent random effects $U_1, \ldots, U_K$; ii) the conditional distribution of Y_i given $U_i = u_i$ is $N(u_i + x_i^{\mathrm{T}}\beta, \hat{\sigma}_i^2)$, where β is an unknown p-vector of effects.

We use the random-effects meta-regression model as a working model for theoretical development, and in §4, we illustrate the good performance of the derived procedures under deviations from this working model.

Typically, $x_{i1} = 1$ and the random effect U_i is assumed to be distributed according to $N(0, \psi)$ $(i = 1, \ldots, K)$, where ψ accounts for the between-study heterogeneity. In matrix notation, and conditionally on $(U_1, \ldots, U_K)^{\mathrm{T}} = u$, the random-effects meta-regression model is

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

where $Y = (Y_1, \ldots, Y_K)^{\mathrm{T}}$, X is the model matrix of dimension $K \times p$ with x_i^{T} in its ith row, and $\epsilon = (\epsilon_1, \ldots, \epsilon_K)^{\mathrm{T}}$ is a vector of independent errors each with a $N(0, \hat{\sigma}_i^2)$ distribution. Under this specification, the marginal distribution of Y is multivariate normal with mean $X\beta$ and variance $\hat{\Sigma} + \psi I_K$, where I_K is the $K \times K$ identity matrix and $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2)$. The random-effects meta-analysis model is a meta-regression model where X is a column of ones.

The parameter β is naturally estimated by weighted least squares as $\hat{\beta}(\psi) = \{X^{\mathrm{T}}W(\psi)X\}^{-1}X^{\mathrm{T}}W(\psi)Y$, with $W(\psi) = (\hat{\Sigma} + \psi I_K)^{-1}$. Then, inference about β can be based on that, under model (1), $\hat{\beta}(\psi)$ has an asymptotic normal distribution with mean β and variance $X^{\mathrm{T}}W(\psi)X$. In this case, the reliability of the associated inferential procedures critically depends on the availability of an accurate estimate of the between-study variance ψ. A popular choice is the DerSimonian & Laird (1986) estimator $\hat{\psi}_{\mathrm{DL}} = \max\{0, (Q - n + p)/A\}$, where $Q = (y - X\hat{\beta}_{\mathrm{F}})^{\mathrm{T}}\hat{\Sigma}^{-1}(y - X\hat{\beta}_{\mathrm{F}})$ and $\hat{\beta}_{\mathrm{F}} = \hat{\beta}(0)$ and $A = \mathrm{tr}(\hat{\Sigma}^{-1}) - \mathrm{tr}\{(X^{\mathrm{T}}\hat{\Sigma}^{-1}X)^{-1}X^{\mathrm{T}}\hat{\Sigma}^{-2}X\}$. Viechtbauer (2005) presents evidence on the loss of efficiency of $\hat{\psi}_{\mathrm{DL}}$, which can impact inference (see also, Guolo, 2012).

Inference about β can alternatively be based on the likelihood function. The log-likelihood function for $\theta = (\beta^{\mathrm{T}}, \psi)^{\mathrm{T}}$ in model (1) is

$$\ell(\theta) = -\frac{1}{2}\log|W(\psi)| - \frac{1}{2}R(\beta)^{\mathrm{T}}W(\psi)R(\beta), \tag{2}$$

where $|W(\psi)|$ denotes the determinant of $W(\psi)$ and $R(\beta) = y - X\beta$. A calculation of the gradient $s(\theta)$ of $\ell(\theta)$ shows that the maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}} = (\hat{\beta}_{\mathrm{ML}}^{\mathrm{T}}, \hat{\psi}_{\mathrm{ML}})^{\mathrm{T}}$ for θ results from solving the equations

$$\begin{cases} s_{(\beta)}(\theta) = X^{\mathrm{T}} W(\psi) R(\beta) = 0_p \,, \\ s_{(\psi)}(\theta) = R^{\mathrm{T}}(\beta) W(\psi)^2 R(\beta) - \mathrm{tr}\,[W(\psi)\}] = 0 \,, \end{cases} \tag{3}$$

where 0_p denotes a p-dimensional vector of zeros, and $s_{(\beta)}(\theta) = \nabla_\beta \ell(\theta)$ and $s_{(\psi)}(\theta) = \partial \ell(\theta)/\partial \psi$, so that $\hat{\beta}_{ML} = \hat{\beta}(\hat{\psi}_{ML})$. As observed in Guolo (2012) and Zeng and Lin (2015), inferential procedures that rely on first-order approximations of the log-likelihood (e.g., procedures based on likelihood-ratio and Wald statistics) perform poorly when the number of studies K is small to moderate.

2 Bias-reducing penalized likelihoods

Using the results in Kosmidis and Firth (2009, 2010), straightforward calculation gives that the first term in the expansion of the bias function of the maximum likelihood estimator is $b(\theta) = \{0_p^{\mathrm{T}}, b_{(\psi)}(\psi)\}^{\mathrm{T}}$, where $b_{(\psi)}(\psi) = -\mathrm{tr}\{W(\psi)H(\psi)\}/\mathrm{tr}\{W(\psi)^2\}$, with $H(\psi) = X\{X^{\mathrm{T}} W(\psi) X\}^{-1} X^{\mathrm{T}} W(\psi)$. An estimator that corrects for the bias of $\hat{\theta}_{\mathrm{ML}}$ results from solving the adjusted score equations $s^*(\theta) = s(\theta) - F(\theta) b(\theta) = 0_{p+1}$ (Firth, 1993; Kosmidis and Firth, 2009). After some algebra, the adjusted score functions for β and ψ are found to be $s^*_{(\beta)}(\theta) = s_{(\beta)}(\theta)$ and

$$s^*_{(\psi)}(\theta) = R^{\mathrm{T}}(\beta) W(\psi)^2 R(\beta) - \mathrm{tr}\,[W(\psi)\{I_K - H(\psi)\}] = 0 \,,$$

respectively. The expression for the differential of the log-determinant gives that the adjusted score functions $s^*_{(\beta)}(\theta)$ and $s^*_{(\psi)}(\theta)$ can also be obtained as derivatives of the penalized log-likelihood function

$$\ell^*(\theta) = \ell(\theta) - \frac{1}{2} \log \left| F_{(\beta\beta)}(\psi) \right| \,, \tag{4}$$

where $\ell(\theta)$ is as in (2), $F_{(\beta\beta)}(\psi) = X^{\mathrm{T}} W(\psi) X$ is the β-block of the information matrix $F(\theta)$, and $|\cdot|$ denotes determinant. So, the solution of the adjusted score equations is the maximum penalized likelihood estimator $\hat{\theta}_{\mathrm{MPL}}$. For $\beta = \hat{\beta}(\psi)$, expression (4) reduces to both the logarithm of the approximate conditional likelihood of Cox and Reid (1987) for inference about ψ, when β is treated as a nuisance component, and to the restricted log-likelihood function of Harville (1977). Hence, maximising the bias-reducing penalized log-likelihood (4) is equivalent to calculating the maximum restricted likelihood estimator for ψ. The latter estimator has originally been constructed to reduce underestimation of variance components in finite samples as a consequence of failing to account for the degrees of freedom that are involved in the estimation of the fixed effects β.

2.1 Penalized likelihood inference

The profile penalized likelihood function can be used to construct confidence intervals and regions, and carry out hypothesis tests for β. If $\beta = (\gamma^{\mathrm{T}}, \lambda^{\mathrm{T}})^{\mathrm{T}}$, and $\hat{\lambda}_{\mathrm{MPL},\gamma}$ and $\hat{\psi}_{\mathrm{MPL},\gamma}$ are the estimators of λ and ψ, respectively, from maximising (4) for fixed γ, then the penalized deviance $2\{\ell^*(\hat{\gamma}_{\mathrm{MPL}}, \hat{\lambda}_{\mathrm{MPL}}, \hat{\psi}_{\mathrm{MPL}}) - \ell^*(\gamma, \hat{\lambda}_{\mathrm{MPL},\gamma}, \hat{\psi}_{\mathrm{MPL},\gamma})\}$ has the usual limiting χ^2_q distribution, where $q = \dim(\gamma)$. To derive this limiting result, note that the adjustment to the scores in (3) is additive and $O(1)$, so the extra terms depending on it and its derivatives in the asymptotic expansion of the penalized likelihood disappear as information increases.

3 Inference on standardized mean differences

The profile penalized likelihood is developed under the validity of the random-effects meta-analysis model. This assumption may be unrealistic, especially in settings where the estimation variances are directly related to the associated summary measure. Here, we examine the performance of penalized likelihood inference and of other popular methods under an alternative specification of the data generating process, where the study-specific effects and their variances are calculated by simulating individual-within-study data. Specifically, we assume that the ith study consists of two arms with n_i individuals each, and that $n_1, \ldots, n_K$ are independent uniform draws from the integers $\{30, 31, \ldots, 100\}$. Then, conditionally on a random effect $\alpha_i \sim N(0, \phi)$, we assume that the observation $z_{i,rj}$ for the jth individual in the rth arm is the realisation of a $N(\mu + I_r(\delta + \alpha_i)\sigma, \sigma^2)$ random variable, where $I_1 = 0$ and $I_2 = 1$. Note that the difference between the marginal variances of the arms increases with ϕ. The true effects are set to $\mu = 0$, $\sigma = 1$ and $\delta = -2$. The study-specific effect of interest is δ, estimated using the standardized mean difference $y_i = J_i(\bar{z}_{i,2} - \bar{z}_{i,1})/s_i$, where s_i^2 is the pooled variance from the two arms of the ith study, and $J_i = 1 - 3/\{8(n_i - 1) - 1\}$ is the Hedges correction (see, e.g., Borenstein et al., 2009, Chapter 4). The corresponding estimated variance for y_i is $\hat{\sigma}_i^2 = 2J_i/n_i + J_i y_i^2/(4n_i)$, which is a quadratic function of y_i. The between-study variance ϕ ranges from 0 to 2.5 and the number of studies K from 5 to 50. For each considered combination of ϕ and K, 10 000 data sets are simulated using the same initial state for the random number generator.

Figure 1 shows the empirical coverage of confidence intervals based on i) the profile penalized likelihood, ii) the inversion of Skovgaard's statistic, which is designed to produce second-order accurate p-values for tests on the mean effect size (see, Guolo 2012), iii) the DerSimonian & Laird estimator $\hat{\beta}(\psi_{\mathrm{DL}})$ and its estimated variance $\sum_{i=1}^{K} 1/(\hat{\sigma}_i + \hat{\psi}_{\mathrm{DL}})$, and, iv) the recent double-resampling proposal in Zeng and Lin (2015).

The profile penalized likelihood interval has comparable performance to the one based on Skovgaard's statistic, with the latter having empirical cover-

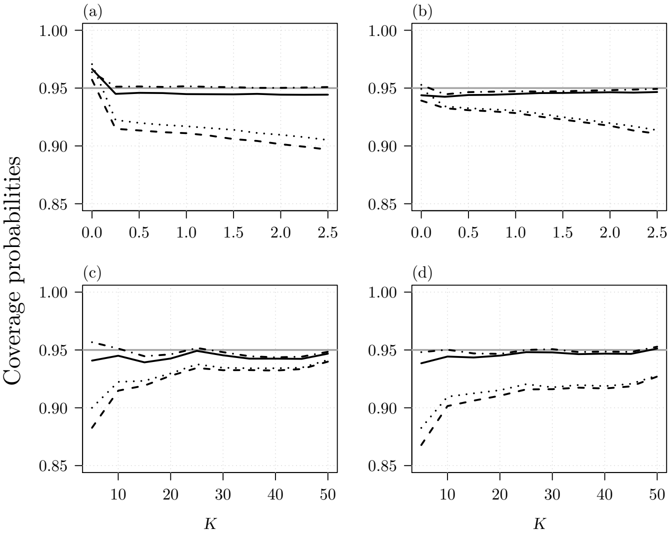


FIGURE 1. Empirical coverage probabilities of two-sided confidence intervals for increasing values of ϕ, when (a) $K = 10$ and (b) $K = 35$, and for increasing values of K when (c) $\phi = 0.25$ and (d) $\phi = 2$. The curves correspond to the proposed profile penalized likelihood method (solid), the DerSimonian & Laird method (dashed), the Zeng & Lin double resampling method (dotted), and the Skovgaard's statistic (dotted-dashed). The grey line is the 95% nominal level.

age that is slightly closer to the nominal level for a wider range of values for ϕ. In general, though, the numerical inversion of Skovgaard's statistic can be unstable due to the discontinuity of the statistic around the maximum likelihood estimator. In contrast, the calculation of profile penalized likelihood intervals is not prone to such instabilities. Intervals based on the DerSimonian & Laird estimator and double resampling perform poorly.

4 Case study

Ambulatory hysteroscopy is a useful instrument to diagnose intrauterine pathologies. Cooper et al. (2010) perform a meta-analysis about the efficacy of different types of local anesthesia used to control pain during hysteroscopy. The available data refer to the use of paracervical anesthesia and consist of the standardized mean differences of pain scores measured at the time of hysteroscopy from five randomized controlled trials. The DerSimonian & Laird estimate of ψ is $\hat{\psi}_{\mathrm{DL}} = 1.08$, while the maximum likelihood estimate is $\hat{\psi}_{\mathrm{ML}} = 2.31$, which is appreciably larger. The maximum penalized likelihood estimate is even larger; $\hat{\psi}_{\mathrm{MPL}} = 2.93$.

The DerSimonian & Laird method strongly supports the importance of the local anesthesia efficacy with a p-value of 0.007, as does the double resampling approach with a p-value of 0.018. The opposite conclusion is obtained using the penalized deviance, which results in a p-value of 0.137. The Skovgaard statistic gives the same conclusion, with a p-value of 0.158.

5 Associated material

Kosmidis et al. (2015) provides a more detailed account of this work, with reproducible case studies and extensive simulation studies.

References

Borenstein, M. H., L. Higgins, and J. Rothstein (2009). *Introduction to meta-analysis.* Chichester, England: Wiley

Cooper, N. A. M., Khan, K. S. & Clark T. J. (2010). Local anaesthesia for pain control during outpatient hysteroscopy: systematic review and meta-analysis. *Brit. Med. J.* **340**, 1130.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc. B* **49**, 1 – 18.

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7**, 177 – 88.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27 – 38.

Guolo, A. (2012). Higher-order likelihood inference in meta-analysis and meta-regression. *Stat. Med.* **31**, 313 – 27.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Stat. Ass.* **72**, 320 – 38.

Kosmidis, I. & Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika* **96**, 793 – 804.

Kosmidis, I. & Firth, D. (2010). A generic algorithm for reducing bias in parametric estimation. *Electron. J. Stat.* **4**, 1097 – 112.

Kosmidis, I., A. Guolo, and C. Varin (2015). Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *arXiv:1509.00650.*

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J. Educ. Behav. Stat.* **30**, 261 – 93.

Zeng, D. & Lin, D. Y. (2015). On random-effects meta-analysis. *Biometrika,* **108**, 281 – 94.

Reduced-bias estimation in generalised linear mixed effects models

Sophia Kyriakou[1], Ioannis Kosmidis[1]

[1] Department of Statistical Science, University College London, London, United Kingdom

E-mail for correspondence: sophia.kyriakou.14@ucl.ac.uk

Abstract: Firth (1993) and Kosmidis and Firth (2009) show that estimates with smaller asymptotic bias than the maximum likelihood estimate can be obtained by suitably adjusting the score function. This approach, though, is not directly applicable in cases where the likelihood is intractable, such as for example, generalised linear mixed models. The current work proposes the Iterated Bootstrap with Likelihood Adjustment, that can reduce bias of the maximum likelihood estimator regardless of the tractability of the likelihood. Simulation studies are used to illustrate the effectiveness of the proposed bias-reduction method. The results also show an improvement in inference about the regression coefficients. We conclude with an application on the salamander mating dataset.

Keywords: Adjusted score equations; Asymptotic bias correction; Bootstrap; Intractable likelihood.

1 Introduction

The popularity of maximum likelihood (ML) in fitting regular statistical models is partly because, under standard regularity conditions, the ML estimator is asymptotically unbiased and fully efficient. However, for finite samples the ML estimator can be severely biased. Such severe bias is observed, for example, in the estimation of the variance components in generalised linear mixed models (GLMMs) and can affect the performance of standard inferential procedures, like Wald tests. While there is a large repository of methods to reduce bias, the intractability of the likelihood in the case of GLMMs prevents the direct use of some of the most popular ones, like the adjusted scores functions approach (Firth, 1993) and asymptotic bias corrections (Efron, 1975).

The current work proposes computational methods for the reduction of bias of the ML estimator that are applicable regardless of the tractability of the likelihood. Our methods systematically correct the mechanism that produces the ML estimates by introducing a small bias in the score function. We adapt and extend the framework in Firth (1993) and Kosmidis and Firth (2009) to the GLMM context. This extension relies on the use of Laplace approximation of the log-likelihood and Monte Carlo approximation of the bias function.

2 Model formulation

Generalised linear mixed models are widely used for analysing non-normal clustered data. The key characteristic of such models is the use of random effects to capture the between-cluster heterogeneity. McCullogh et al. (2008, Chapter 7) provide a thorough description of these models.

A GLMM is specified by (i) the linear predictor, (ii) the link function, (iii) the conditional distribution for the response variable given the random effects, and (iv) the random effects distribution. The linear predictor is $X\beta + Z\alpha$, where X is the $N \times p$ matrix of fixed-effects terms associated with the p regressors, β is the corresponding $p \times 1$ vector of the fixed-effects regression coefficients, Z is the $N \times q$ design matrix for the q random effects and α is the $q \times 1$ vector of the random effects. The conditional mean μ_i of the response is modeled as $g(\mu_i) = X_i^{\mathrm{T}}\beta + Z_i^{\mathrm{T}}\alpha$, where $g(\cdot)$ is the link function. Given an unobserved vector of random effects, the response vector Y is assumed to consist of conditionally independent elements, each following a distribution from the exponential family. To complete the specification of the model we assign a distribution to the random effects, which are commonly assumed to follow a multivariate normal distribution with zero mean.

In general, the integrals involved in the likelihood of GLMMs are intractable and numerical integration techniques can be used to evaluate them (McCullogh et al., 2008, Chapter 7). Maximising the approximated log-likelihood with respect to the model parameters yields the maximum approximate likelihood (MAL) estimates.

As an illustration of the bias of MAL estimates for GLMM parameters, we performed a simulation study where we considered a generalised linear model for binary responses with logistic link function and a random intercept. If the data is arranged in a series of q clusters of observations (Y_{ij}, X_{ij}), where i identifies the cluster, and $j \in \{1, \ldots, n_i\}$ identifies subjects within clusters, the model can be expressed as

$$\begin{aligned} Y_{ij}|\alpha &\sim \text{indep. Bernoulli}(\pi_{ij}) \\ \alpha_i &\sim \text{i.i.d. } N(0, \sigma_\alpha^2) \\ \operatorname{logit}(\pi_{ij}) &= \beta_0 + \beta_1 X_{ij} + \alpha_i \, . \end{aligned} \tag{1}$$

Table 1 shows the estimated bias, mean squared error and Monte Carlo error of the MAL estimates based on 1000 simulated samples, with true parameter values $(\beta_0, \beta_1, \sigma_\alpha) := (0, 5, \sqrt{1.5})$, $X_{ij} = j/n$, and using the Laplace approximation for evaluating the log-likelihood. There are $q = 8$ levels of the random effect and $n \in \{15, 50, 100\}$ observations per level of the random effect. The bias of β_1 is large, especially for small n, and the random effect parameter σ_α is underestimated, but as n increases bias decreases in absolute value. The mean squared error of the parameter estimates decreases with increasing sample size. We also evaluate the size of the Wald test for the null hypothesis $\beta_1 = 5$ using the MAL estimates. The Type I error is estimated to be equal to 0.028, 0.053, 0.055 at a 5% nominal level for n = 15, 50, 100, respectively, indicating bad performance of the Wald test when n is small.

TABLE 1. Bias, mean squared error (MSE) and Monte Carlo error (MCE) of the maximum approximate likelihood estimates for the parameters of model (1).

	Bias			MSE			MCE		
n	15	50	100	15	50	100	15	50	100
β_0	-0.041	0.002	0.004	0.512	0.276	0.246	0.023	0.017	0.016
β_1	0.313	0.097	0.063	2.824	0.680	0.318	0.049	0.026	0.018
σ_α	-0.193	-0.123	-0.109	0.424	0.172	0.140	0.020	0.013	0.011

3 Bias reduction in models with intractable likelihood

Firth (1993) showed that an estimator with $o(N^{-1})$ bias, where N is the sample size or some other measure of how information accumulates for the parameters of the model, results by the solution of the adjusted score equations $S^*(\theta) = S(\theta) - I(\theta)b(\theta) = 0$. In the latter equations, θ is the p-dimensional parameter of the model, $S(\theta)$ is the score function, $I(\theta)$ is the observed information matrix, and $b(\theta)$ is the first-order term in the expansion of the bias of the ML estimator. More generally, the theory in Firth (1993) and Kosmidis and Firth (2009) guarantees that estimators with $o(N^{-1})$ bias result by the solution of

$$S^*(\theta) = S(\theta) - I(\theta)B(\theta) + v(\theta) = 0\,, \qquad (2)$$

where $B(\theta) = \mathbb{E}_\theta(\hat{\theta} - \theta)$ is the bias of the ML estimator $\hat{\theta}$ and $v(\theta) = O_p(N^{-1/2})$. For models with intractable likelihood, (2) cannot be solved directly, because all quantities involved are generally intractable.

In this paper we will consider variations of (2) that are tractable and lead to estimators with $o(N^{-1})$ bias. Theorem 1 shows that Laplace approximation can be used to construct such variations.

Theorem 1: Let $\tilde{\theta}^*$ be the solution of

$$\tilde{S}^*(\theta) = \tilde{S}(\theta) - \tilde{I}(\theta)B(\theta) = 0\,, \tag{3}$$

where $\tilde{S}(\theta)$ and $\tilde{I}(\theta)$ are the gradient and negative Hessian matrix of Laplace approximation of the log-likelihood. Then, $\tilde{\theta}^*$ has $o(N^{-1})$ bias.
Proof: Using the results in Tierney and Kadane (1986) it can be shown that $S(\theta) - \tilde{S}(\theta) = O(N^{-1})$ and $I(\theta) - \tilde{I}(\theta) = O(N^{-1})$. Given also that $B(\theta)$ is of order $O(N^{-1})$, $\tilde{S}^*(\theta) - [S(\theta) - I(\theta)B(\theta)]$ has smaller order than $O_p(N^{-1/2})$. Hence, from (2), the solution of (3) has $o(N^{-1})$ bias. □

4 Iterated Bootstrap with Likelihood Adjustment

A natural way to estimate the generally unknown bias $B(\theta)$ in (3) is by Monte Carlo. Let $\hat{B}_T(\theta)$ be the Monte Carlo estimator of the bias at θ based on T samples generated under the model at θ. Substituting $\hat{B}_T(\theta)$ for $B(\theta)$ in (3), the adjusted score equations are

$$\tilde{S}^*_T(\theta) = \tilde{S}(\theta) - \tilde{I}(\theta)\hat{B}_T(\theta) = 0\,. \tag{4}$$

A direct approach for evaluating the solution of (4), $\tilde{\theta}^*_T$, is through a quasi Newton-Raphson iteration of the form

$$\theta^{(j+1)} = (2\theta^{(j)} - \bar{\theta}^{(j)}_T) + \{\tilde{I}(\theta^{(j)})\}^{-1}\tilde{S}(\theta^{(j)})\,, \tag{5}$$

where $\theta^{(j)}$ is the candidate value for $\tilde{\theta}^*_T$ at the jth iteration, and $\bar{\theta}^{(j)}_T$ is the average of the MAL estimators calculated for each of T simulated samples from the model at $\theta^{(j)}$. Starting from the MAL estimate, a single iteration gives a bootstrap corrected estimate, so iteration (5) can be seen as a generalisation of the bootstrap for bias correction (Efron and Tibshirani, 1993, Ch. 10). For this reason we refer to the proposed bias reduction method as Iterated Bootstrap with Likelihood Adjustment (IBLA). A stopping criterion for the iterations is $|\tilde{S}^*_T(\theta^{(j+1)})| < \epsilon$, for some small $\epsilon > 0$. We generally recommend using the same state for the random number generator in each iteration in order to achieve a smooth estimator of the bias function.
We now revisit the simulation study performed in Section 2. Table 2 reports the IBLA parameter estimates of model (1), where the simulation size T used for the calculation of $\hat{B}_T(\theta)$ was set to 200, 500, 1000 for $n =$ 15, 50, 100, respectively, and $\epsilon = 0.05$. The results show that IBLA reduces the bias of MAL, especially for small and moderate values of n. A reduction in the mean squared error is also observed. The empirical size of the Wald test when the statistic is based on the IBLA estimates is calculated to be 0.051, 0.051, 0.055 at a 5% nominal level for $n = 15, 50, 100$, respectively. Comparing with the results in Section 2, use of the bias-reduced estimates seems to also deliver a marked improvement in inference compared to the MAL estimates.

Figure 1 shows the IBLA iterations for one of the simulated samples, with the MAL estimates used as initial values. In each iteration a new set of T bootstrap samples is generated from the model at the parameter estimates from the previous iteration, and (5) is used to obtain updated estimates. For this particular sample the estimates stabilize in less than 10 iterations.

TABLE 2. Bias, mean squared error (MSE) and Monte Carlo error (MCE) of the IBLA estimates for the parameters of model (1).

	Bias			MSE			MCE		
n	15	50	100	15	50	100	15	50	100
β_0	-0.015	0.006	0.012	0.430	0.270	0.244	0.021	0.016	0.016
β_1	-0.230	0.001	0.006	1.970	0.617	0.305	0.044	0.025	0.017
σ_α	-0.182	-0.003	0.021	0.327	0.190	0.181	0.017	0.014	0.013

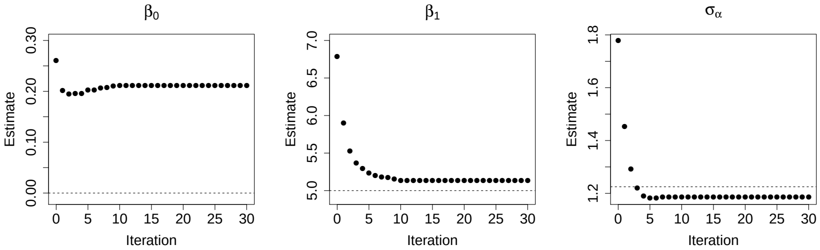


FIGURE 1. Candidate values for the IBLA parameter estimates of model (1) computed using algorithm (5), based on one simulated sample. The dashed lines represent the true parameter values.

5 Numerical example

We use IBLA on the salamander mating data (McCullagh and Nelder, 1989, Ch. 14.5). Three experiments were carried out in order to investigate whether there are barriers to inter-breeding in the salamanders from two populations, Rough Butt (R) and Whiteside (W). In our analysis we use the data from the first experiment, which involved 20 female and 20 male salamanders. Each female salamander mated with six male salamanders; three from its population and another three from the other population under a crossed design. In total, there are 120 observations.

We considered a GLMM with a logistic link function that takes into account heterogeneity between the female and male salamanders used in the experiment. The linear predictor of the model is

$$\text{logit}\{P(Y_{ij}=1|\alpha_i,b_j)\} = \beta_0+\beta_1 f_i+\beta_2 m_j+\beta_3 f_i m_j+\alpha_i+b_j\,, \quad (6)$$

where $Y_{ij} = 1$ if the mating is successful, and 0 otherwise. The fixed effects are the population type of the female ($f_i = 1$ if the ith female is a W, 0 otherwise), the population type of the male ($m_j = 1$ if the jth male is a W, 0 otherwise), and their interaction. We assume α_i and b_j are independently distributed as $N(0, \sigma_\alpha^2)$ and $N(0, \sigma_b^2)$, respectively.
Table 3 reports the IBLA estimates of the model parameters. For comparison we also include the estimates obtained by MAL and the (corrected) penalised quasi-likelihood (C)PQL, as reported in Noh and Lee (2007). Compared to the MAL estimates, IBLA (with $T = 1000$) shrinks the regression coefficients towards 0, and results in the largest estimate for σ_α.

TABLE 3. Estimates for the parameters of model (6).

Method	β_0	β_1	β_2	β_3	σ_α	σ_b
MAL	1.33	-2.94	-0.42	3.18	1.25	0.26
PQL	1.16	-2.57	-0.38	2.81	1.19	0.30
CPQL	1.55	-3.50	-0.50	3.82	1.31	0.63
IBLA	1.23	-2.72	-0.38	2.94	1.36	0.28

References

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics*, **3**, 1189–1217.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Kosmidis, I. and Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, **96**, 793–804.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall, 2nd edition.

McCullogh, C.E., Searle, S.R., and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models.* John Wiley & Sons, 2nd edition.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, **98**, 896–915.

Tierney, L. and Kadane, J.B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, **81**, 82–86.

Smooth composite link models for spatio-temporal dissagregation of epidemiological data

Dae-Jin Lee[1], Diego Ayma[2], María Durbán [2], Jan van de Kassteele[3]

[1] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[2] Department of Statistics, Universidad Carlos III de Madrid, Spain
[3] National Institute for Public Health and the Environment, Bilthoven, The Netherlands

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

Abstract: We propose the extension of the penalized composite link model in both spatial and temporal aggregations levels. We present a smoothing approach to account for aggregated counts both in space and time to estimate smooth incidence maps for the Q-fever outbreak in the Netherlands in 2009.

Keywords: penalized composite link; aggregated counts; mixed models; SAP; space-time modelling

1 Introduction

During the period 2007-2009, the Netherlands faced large outbreaks of Q-fever cases in humans. *Q-fever* is a disease caused by infection with *Coxiella burnetii*, a bacterium that affects humans and other animals. An epidemiological link was established with dairy goat farms, and to a lesser extent with dairy sheep farms, that experienced high abortion rates caused by *C. burnetii.*

The development of spatial statistical methods on routine surveillance data is an important tool for Public Health and veterinary authorities on control measures with respect to Q-fever. To inform decision makers about the relative importance of different infection sources, Van der Hoek *et al.* (2012) developed accurate and high-resolution incidence maps for detection of Q fever hot spots. A 500×500 m grid was imposed over the area of interest and the number of cases and the population number were counted in each cell. They analyzed the Q-fever incidents as an spatial point pattern assuming an inhomogeneous Poisson process where intensity is assumed to vary in space. In order to estimate a smooth incidence map they proposed a penalized splines approach using tensor products of B-splines (Currie *et al*, 2006). However, usually Public Health authorities do not report the exact georeferenced location of the incidence and instead the aggregated counts are collected and shown in choropleth maps. In this paper, we assume that the observed (raw) data are provided in aggregated form by municipalities and by months and in order to propose a methodology to estimate the space-time latent distribution at disaggregated level. Figure 1 shows an area in the south of the Netherlands with aggregated counts by municipality and monthly counts during 2009.

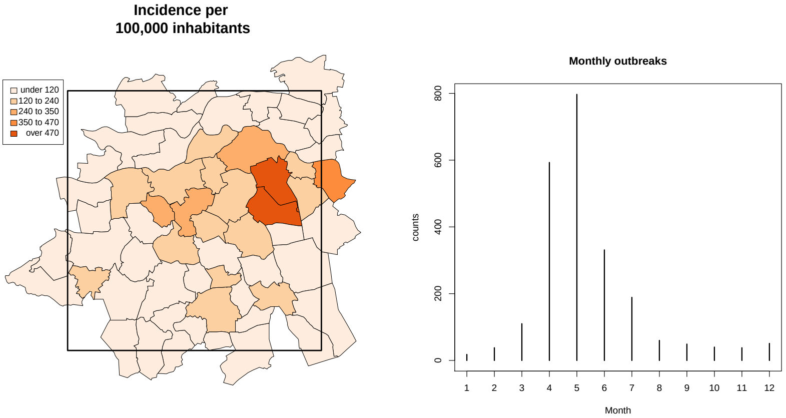


FIGURE 1. Left: Map of study and Q-fever incidents by municipality in 2009 and Right: monthly aggregated incidents of 2009.

2 Space-time Composite Link Model

We extend the so-called *penalized composite link model* (PCLM) approach by Eilers (2007), to deal directly with both spatial and temporal aggregation of the counts and estimate the latent distribution of the incidents across space and time. Ayma et al. (2016) proposed the extension in the spatial case where only spatial aggregation is considered. The model is given by:

$$\boldsymbol{\mu} = \boldsymbol{C}\boldsymbol{\gamma} = \mathbf{C}\exp(\mathbf{B}\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\gamma}$ represents the mean vector of the latent process at a desirable fine resolution, $\boldsymbol{C}$ is the composition matrix that describes how these latent expectations are combined to yield $\boldsymbol{\mu}$ and $\boldsymbol{B}$ is the Tensor product of B-spline bases $\boldsymbol{B} = \boldsymbol{B}_s \otimes \boldsymbol{B}_t$, where $\boldsymbol{B}_s = B_1 \otimes B_2$ is the Tensor product of B-spline bases for longitude and latitude (B_1 and B_2) and $\boldsymbol{B}_t$ the B-spline basis function for the time component. The vector of regression coefficients $\boldsymbol{\theta}$ is penalized by an anisotropic penalty matrix with a smoothing parameter for each dimension, i.e., $\text{Pen}(\lambda_1, \lambda_2, \lambda_t)$.

In order to consider the spatial and temporal aggregation s simultaneously, we build the composition matrix $\boldsymbol{C}$ as the Kronecker product of both spatial and temporal composition matrices, i.e $\boldsymbol{C} = C_s \otimes C_t$, where C_s accounts for the spatial grouping of the counts (i.e. by municipalities) and C_t is the compositional matrix for the temporal aggregation.

In the spatial dimension, we aim to estimate the spatial mortality trend at a fine grid, using health data available at coarse geographical units, i.e., the area-to-point (ATP) case, reducing the visual bias associated with the interpretation of choropleth maps caused by the variation in shape and size of the units. Hence, we impose a regular grid of points into the map and for C_s, we consider $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ as the coordinates of the points of the fine grid, which fall inside of the geographical units $\boldsymbol{v}_i$. Thus, the elements of the associated (spatial) composition matrix C_t become:

$$c_{ij} = \begin{cases} 1 & \text{if } (x_{1j}, x_{2j}) \text{ belongs to unit } \boldsymbol{v}_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

for $i = 1, ..., n$, and $j = 1, ..., m$. For the temporal counts disaggregation, suppose m coarse intervals are given and the smooth time trend is to be estimated on r times narrower intervals, then the temporal composition matrix C_t is a $m \times mr$ matrix with elements equal to zero, except that $c_{ij} = 1$ if $r(i-1) < j \leq ri$. Hence, for monthly disaggregation, when data are agreggated in quarters ($m = 4$ and $r = 3$), i.e.:

$$C_t = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}. \tag{3}$$

2.1 Model estimation

To estimate the model in (1) , we reformulated it as a mixed model $\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma}$, $\boldsymbol{\gamma} = \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})$, such that the estimation of the fixed and random effects is done as the iterative solution of the system of equations:

$$\begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{X}} & \breve{\mathbf{X}}'\mathbf{W}\breve{\mathbf{Z}} \\ \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{X}} & \mathbf{G}^{-1} + \breve{\mathbf{Z}}'\mathbf{W}\breve{\mathbf{Z}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \breve{\mathbf{X}}'\mathbf{W}\boldsymbol{z} \\ \breve{\mathbf{Z}}'\mathbf{W}\boldsymbol{z} \end{bmatrix}, \tag{4}$$

with "working" design matrices $\breve{\mathbf{X}}$ and $\breve{\mathbf{Z}}$ are defined as $\breve{\mathbf{X}} = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{X}$ and $\breve{\mathbf{Z}} = \mathbf{W}^{-1}\mathbf{C}\boldsymbol{\Gamma}\mathbf{Z}$, respectively, with $\mathbf{W} = \text{diag}(\boldsymbol{\mu})$ and $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$.

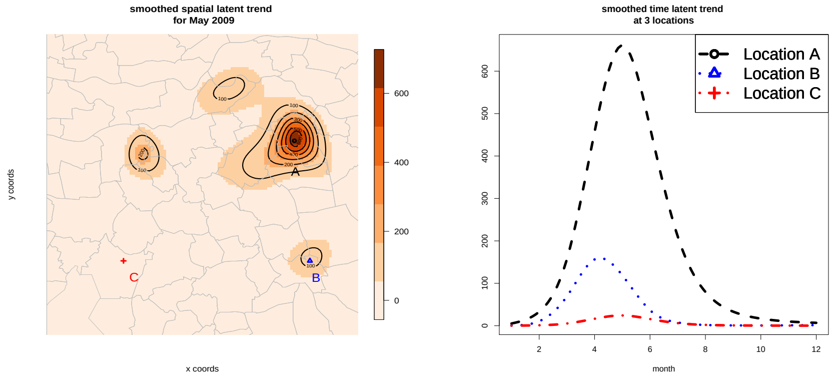


FIGURE 2. Left: Smoothed latent incidence map for May 2009 and Right: Smoothed incidence latent temporal trend in 2009 at three chosen locations.

The random effects $\boldsymbol{\alpha}$ have covariance matrix $\boldsymbol{G}$ which depends on the variance components τ_1, τ_2 and τ_t. The working vector is defined as $\boldsymbol{z} = \breve{\mathbf{X}}\boldsymbol{\beta} + \breve{\mathbf{Z}}\boldsymbol{\alpha} + \mathbf{W}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}\right)$. This yields to a modified version of the standard mixed model estimators:

$$\widehat{\boldsymbol{\beta}} = (\breve{\mathbf{X}}'\mathbf{V}^{-1}\breve{\mathbf{X}})^{-1}\breve{\mathbf{X}}'\mathbf{V}^{-1}\boldsymbol{z}, \text{ and } \widehat{\boldsymbol{\alpha}} = \mathbf{G}\breve{\mathbf{Z}}'\mathbf{V}^{-1}(\boldsymbol{z} - \breve{\mathbf{X}}\widehat{\boldsymbol{\beta}}). \tag{5}$$

Figure 2 shows the smoothed map for Q-fever latent incidences in May 2009 and the smoothed latent time trends at 3 chosen locations (indicated in the map as A, B and C).

3 Conclusions

We presented a methodology for estimation of latent intensity functions in both spatial and temporal counts using Penalized composite link models. The approach takes the computational advantages of recent methods for variance components estimation (Rodríguez-Álvarez *et al*, 2015) combined with array methods (Currie *et al*, 2006). We consider the Q-fever disease incidents in The Netherlands during 2009. Further extensions include a simulation study and model performance comparisons of the disaggregated latent smooth intensity map by means of the PCLM with the space-time P-spline model where a spatial point pattern is considered. The proposed approach might help the authorities to document the development of epidemics and therefore determine Public Health policies.

Acknowledgments: This research was supported by the Spanish Ministry of Economy and Competitiveness MINECO grants MTM2011-28285-

C02-02 and MTM2014-52184. The research of Dae-Jin Lee was also supported by the Basque Government through the BERC 2014-2017 program and the Basque Government Industry Department under the ELKARTEK Program and by the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323.

References

Ayma, D, Durbán, M., Lee, D.-J. and Eilers, P.H.C. Penalized composite link models for aggregated spatial count data: a mixed model approach. *Under review.*

Currie, I. D., Durbán, M. and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc. B*, **68**, 1-22.

Eilers, P.H.C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, 7(3):239–254.

Lee, D.-J., and Durbán, M. (2011). *P*-spline ANOVA-Type interaction models for spatio-temporal smoothing. *Statistical Modelling*, Vol. 11, Issue 1, Pages 49-69.

Rodríguez-Álvarez, M.X., Lee, D.-J., Kneib, T. Durbán, M. and Eilers, P.H.C. (2015). Fast algorithm for smoothing parameter selection in multidimensional *P*-splines. *Statistics and Computing* 25:5 (2015) 941- 957.

Van der Hoek, W., Van de Kassteele, J., Bom, B., de Bruin, A., Dijkstra, F., Schimmer, B., Vellema, P., ter Schegget, R. and Schneeberger, P. M. (2012). Smooth incidence maps give valuable insight into Q fever outbreaks in The Netherlands. *Geospatial Health* 7(1), pp. 127–134.

Penalized Fixed Effects Model for Clustered and Longitudinal Data

Ying Lu[1], Marc Scott[1], Heng Peng[2]

[1] New York University, NY, NY, USA
[2] Hong Kong Baptist Unversity, Hong Kong, China

E-mail for correspondence: ying.lu@nyu.edu

Abstract: Linear mixed effects models are often used to fit data with clustered structure. The cluster-specific effects can be modelled either via random or fixed effects. The latter is more attractive when inference for cluster-specific effects is of interest. However, the fixed effect approach is criticized for producing inefficient covariate effect estimates and being unable to include predictors whose values remain constant within cluster or individual. In this article, we propose a penalized fixed effects estimation approach that produces a more parsimonious set of fixed effect estimates and as a result allows for estimating the effects of cluster-level predictors. We will illustrate this approach using the National Longitudinal Survey of Youth. Theoretical and numerical results regarding the effects for covariates and the penalized cluster-specific fixed effects will be discussed.

Keywords: Model selection; category combination; penalized least squares

1 Introduction

In this paper, we consider fitting the following model,

$$Y_{ij} = X_{ij}\beta + u_j + \epsilon_{ij} \qquad \text{for} \quad i = 1, \ldots, n_j, \quad \text{and} \quad j = 1, \ldots, J \quad (1)$$

where j indexes the J clusters and there are n_j observations within each cluster indexed by i. X_{ij} is a vector of observed variables, and u_j is the so-called cluster specific effect. Y_{ij} is the outcome variable of interest and ϵ_{ij} is the error term. Traditionally, depending upon the inference of interest, there are two approaches to estimating model (1): treating u_j as random variable with a parametric distribution such as $N(\mu_{u_0}, \sigma_u^2)$, or treating u_j as parameters to be estimated (sometimes called econometric "fixed effects" modelling). When β are the quantities of interest, the random effect approach is likely preferred as it produces a more efficient estimator of β

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).

than the fixed effect approach does. However, the consistency of this estimator depends on the assumption that u and X are independent, which can often be violated if the level of a predictor depends on the context (i.e., the cluster). When both the inference about β and u_j are of interest, the u_j are often directly estimated using an indicator variable approach. When the number of clusters is large, this fixed effect approach requires estimating a much larger number of parameters than the random effect approach, reducing the precision of estimates. Moreover, if the predictors X contain a subset of cluster-specific variables, Z_j, such that observations within the same cluster j do not vary on predictors Z_j, one runs into an identification problem when simultaneously estimating $u_j, j = 1, \ldots, J$ and the coefficients for Z_j. To ameliorate the problems of the fixed effect approach and to provide a consistent estimator of β when the independence assumption fails, we propose a penalized fixed effect model that is based on an assumption that the number of unique values of u_j is much smaller than the number of clusters (i.e., there are natural groups of larger clusters). The proposed approach aims to simultaneously detect the underlying structure of u_j to combine the clusters with the same underlying u_j values and estimate parameters β and unique values of u_j.

2 The Proposed Method

Rewrite the fixed effect model in the matrix form

$$Y \quad = \quad X\beta + Z\gamma + Du + \epsilon \tag{2}$$

where matrix D represents a dummy variable design matrix of the clusters, in which row $D_{i\cdot}$ takes value 1 in position j when in group j, and 0 elsewhere, representing cluster membership, and $u = (u_1, \ldots, u_J)$. Note that due to collinearity, γ and u can not be immediately identified under this model without further constraints.

- Obtain initial values of β, γ and u using a ridge regression or random effects specification. Order the data according to the ridge regression estimates of u_j, $\tilde{u}_j$, to obtain the linear transformation matrix $\tilde{D}$ such that $Du = \tilde{D}\xi$, and $\xi = (u_{(1)} - u_{(2)}, \ldots, u_{(J-1)} - u_{(J)}, u_{(J)})'$ such that $\tilde{u}_{(1)} < \tilde{u}_{(2)} < \ldots < \tilde{u}_{(J)}$. We estimate β and ξ using penalized least squares
$$Q(\beta, \gamma, \xi) = \frac{1}{2}(Y - X\beta - Z\gamma - \tilde{D}\xi)'(Y - X\beta - Z\gamma - \tilde{D}\xi) + J\sum_{j=1}^{J-1} p_\lambda(|\xi_j|)$$
- Obtain penalized least squares estimates of β, γ and ξ: following Fan and Li (2001), the penalty $p_\lambda(|\xi_j|)$ can be locally approximated by the quadratic function
$$p_\lambda(|\xi_j|) = p_\lambda(|\xi_j^{(0)}|) + \frac{1}{2}p'_\lambda(|\xi_j^{(0)}|)/|\xi_j^{(0)}|(|\xi_j|^2 - |\xi_j^{(0)}|^2)$$

when the initial value $\xi^0 \neq 0$. Let $B = (X, \tilde{D})$; $\gamma = (\beta, \xi)$, the estimator of γ can be obtained by solving the following equation iteratively,

$$\left[B^\tau B + n\Omega_\lambda(\gamma^{(0)})\right] \gamma = B^\tau Y,$$

where $\Omega_\lambda(\gamma^{(0)}) = diag\left[\mathbf{0}_p, \left(\frac{p'_\lambda(|\xi_1^{(0)}|)}{|\xi_1^{(0)}|}, \ldots, \frac{p'_\lambda(|\xi_{J-1}^{(0)}|)}{|\xi_{J-1}^{(0)}|}\right)\right]$. During the iteration, if $\hat{\xi}_j$ is less than an arbitrarily chosen small value, for example 0.001 times its estimated standard error, it will be shrunk to zero, and the corresponding two adjacent groups j and $j+1$ will be combined into the same group. Correspondingly, $\tilde{D}$ will be redefined based on the new grouping structure.

- Upon convergence, u is recovered through its back transformation $u = \tilde{D}^{-1}\xi$.

- The standard errors of the proposed estimators can be obtained using the sandwich variance estimator (Kauermann and Carroll, 2001). Denote $\hat{\gamma}_1$, the nonvanishing component of $\hat{\gamma}$. The covariance of $\hat{\gamma}_1$ is given as follows,

$$\hat{\sigma}^2 \left[B_1^\tau B_1 + n\hat{\Omega}_\lambda\right]^{-1} B_1^\tau B_1 \left[B_1^\tau B_1 + n\hat{\Omega}_\lambda\right]^{-1},$$

where $\hat{\sigma}^2$ is given as

$$\hat{\sigma}^2 = \frac{SSE_\lambda}{df_\lambda} = \frac{\|Y - B_1^\tau \hat{\gamma}_1\|^2}{tr\{B_1[B_1^\tau B_1 + n\Omega_\lambda(\hat{\gamma}_1)]^{-1}B_1^\tau\}},$$

where B_1 is the sub-matrix of B corresponding to $\hat{\gamma}_1$

- The tuning parameter λ can be chosen using cross-validation.

3 Theoretical and Numerical Studies

Theoretical results regarding consistency of the estimator for β and the sparsity of the clusters are established in a longer paper. Moreover, the estimators can be shown to have the Oracle property.

We conduct two simulation studies to assess the performance of the proposed penalized fixed effect model (PFE) in comparison with the classic fixed effect dummy variable regression (FE) and random effect model (RE). We also compare the results with those based on the ordinary least squares (OLS) ignoring grouping structure and a fixed effect dummy variable regression under the oracle situation when the true group labels are known (OFE). In both simulation studies, the data are simulated according to the following model,

$$Y_{ij} = X_{ij}\beta + Z_j\gamma + u_j + \epsilon_{ij}, \quad i = 1, \ldots, I, j = 1, \ldots J \tag{3}$$

where X_{ij} is a $n \times p_1$ balanced design matrix, while Z is a $n \times p_2$ cluster-specific covariate matrix, with $n = I \times J$. $\epsilon_{ij} \overset{iid}{\sim} N(0,1)$.

Simulation One: First, we compare the performance of the various models and estimation approaches varying numbers of groups and group sizes when none of the covariates X and Z are correlated with the group effects u_j. To keep the comparison simple and allow the inclusion of a fixed effects model, we set $p_1 = 2$ and $p_2 = 0$; X_1 and X_2 are drawn independently from uniform distribution on $[-2, 2]$, and they are generated independently of u. The associated regression coefficients $\beta = (1, 2)$. We let the number of groups and the group size vary ($k = 12, 48$, $m = 2, 5, 10, 20, 40$). We further assume that u_j only take three distinct values, $(-1, 0, 1)$, in equal proportions.

According to the simulation set up, there should be three meta-groups after recombination. To assess the performance of group recombination, for each member in the recombined group, we calculate the percent of its fellow group members that are correctly assigned to the same group. Table 1 summarizes the results in terms of % correct using the penalized fixed effect approach using three different methods–AIC, BIC and General Cross Validation(GCV)–for selecting the tuning parameter for the penalty function methods. We can see that even at very low group size ($m = 2$), about 60% of observations are correctly assigned. As the group size grows, the penalized fixed effect method can effectively combine the groups with minimal error. The three tuning parameters have similar performance, but it seems that GCV performs better at lower sample size, while AIC leads to higher percentage of correctly combined groups at higher sample size. The PFE approach also produces unbiased estimators for β and comparable standard errors when compared to the nominal standard errors over simulation repetitions (results not shown).

	k=12			k=48		
m	BIC	AIC	GCV	BIC	AIC	GCV
2	61.01	60.20	61.36	58.62	60.87	60.32
5	71.70	70.06	74.52	71.91	76.98	74.03
10	84.80	83.38	88.08	85.94	89.98	87.99
20	95.79	95.58	98.14	96.13	97.39	97.66
40	99.46	99.73	99.83	99.86	99.96	100.00

TABLE 1. Percent of clusters correctly grouped under Penalized Fixed Effect Model

Simulation Two: In this subsection, we simulate the data mimicking a more realistic longitudinal setup, where there are $k = 24$ individuals, each observed at $m = 10$ different time points. In this context, we assume that there are two time-variant covariates X, $p_1 = 2$, and three time-variant covariates Z, $p_2 = 3$. We set the true $\beta = (1, 2)$ and the true $\gamma = (1, 0.5, 1)$.

Each of the X and Z variables are i.i.d. $N(0,1)$. The individuals are assumed to be in three equal sizes groups, with distinct u values, $(-1,0,1)$. Furthermore, we consider two situations of endogeneity: case one, $cor(X_1,u) = 0.8)$; case two $cor(X_1,u) = cor(Z_1,u) = 0.8$. Table 2 summarizes the bias in parameter estimates for each case using different modelling approaches. To highlight the impact of imposed endogeneity on parameter estimates, the coefficients of the endogenous covariates are highlighted in bold font. We find that when some time-varying covariate X_1 is specified as endogenous to the group effects u, both OLS and random effect models produced biased estimates of β_1. In contrast, the bias in estimating β is ameliorated under the PFE approach. On the other hand, when some time-invariant covariate Z_1 is also specified as endogenous to the group effects u, all approaches produce biased estimator of γ. In terms of group recovery, when only X is endogenous with respect to u, the performance is comparable to the independent case as in simulation one, but not when both X and Z are endogenous.

	β_1	β_2	γ_1	γ_2	γ_3	% correct
	Case One					
OLS	**0.354**	0.007	-0.006	-0.004	-0.001	–
RE	**0.237**	0.009	-0.005	-0.004	0.005	–
PFE-aic	**0.135**	0.006	0.005	-0.001	-0.004	71.44
PFE-gcv	**0.049**	0.006	0.001	0.004	-0.000	69.22
PFE-bic	**0.072**	0.006	-0.003	-0.004	0.001	73.93
OFE	**0.012**	0.006	0.008	0.000	0.001	–
	Case Two					
OLS	**0.290**	0.012	**0.162**	0.013	0.019	–
RE	**0.174**	0.013	**0.270**	0.010	0.017	–
PFE-aic	**0.118**	0.014	**0.323**	0.016	0.021	51.56
PFE-gcv	**0.030**	0.015	**0.380**	0.018	0.008	50.77
PFE-bic	**0.056**	0.015	**0.362**	0.023	0.014	55.23
OFE	**0.001**	0.012	**-0.018**	0.019	0.017	–

TABLE 2. Comparison of bias in parameter estimation under different approaches. The last column shows the percent of individuals are correctly combined

4 Data Analysis

In this section, we illustrate one application of the proposed methodology in the context of panel data. The dataset used is a subset of the National Longitudinal Survey of Youth (1980–2000) which consists of 424 subjects who have been observed at least twice during this period. Following Doughty (2006), we are interested in evaluating an array of predictors on individual earnings. Table 3 shows the parameter estimates of the predictors under three alternative specifications. One notable feature of this table is that the coefficients of the random effect model vary greatly from those of the fixed effect model. We further conducted a Durbin-Watson-Hausman test and obtained statistical significance, which suggests that there is endogeneity from the unobserved variables, and the random effect model would yield biased parameter estimates. On the other hand the standard errors of the parameters in the fixed effect model are considerably large due to the $J-1=423$ additional parameters. The number of distinct u_j reduces to 54 under the PFE approach, and this leads to a significant reduction of the standard error of $\hat{\beta}$.

					Combination	
	Random Effect		Fixed Effect		Category	
Variable	$\hat{\beta}$	SE	$\hat{\beta}$	SE	$\hat{\beta}$	SE
Age	-0.213	0.109	-0.643	0.394	-0.606	0.061
Education (yrs)	1.326	0.127	1.745	0.663	1.747	0.066
To be married	0.603	0.770	-0.449	1.220	-0.258	0.514
Single	-1.272	0.653	0.381	1.426	0.123	0.323
Experience (yrs)	0.638	0.269	1.033	0.602	1.013	0.192
Experience SQ	-0.003	0.013	-0.001	0.016	-0.001	0.010
Tenure	0.219	0.143	0.100	0.162	0.110	0.108
Tenure SQ	-0.004	0.011	-0.005	0.012	-0.007	0.008
Private sector	-6.393	1.792	-5.642	1.898	-5.886	1.520
Public sector	-7.847	1.935	-5.353	2.139	-5.470	1.588
Hours work per week	-0.061	0.021	-0.095	0.025	-0.093	0.016
Union member	-0.527	1.864	-2.072	2.025	-1.773	1.520
Union wage limit	1.424	1.785	2.064	1.875	1.835	1.494
$\hat{\sigma}^2$	40.47		40.50		33.64	
model size	14		437		67	

TABLE 3. Parameter comparison between three different approaches

References

Doughty, C. (2006). Introduction to Econometrics, 4th ed, Oxford Press.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of American Statistical Association. 96: 1348–1360.

Kauermann, G. and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimation. Journal of the American Statistical Association 96: 1387–1396.

Statistical Inference for Single-Case Design: Application to Post-stroke Rehabilitation

Ying Lu[1], Marc Scott[1], Preeti Raghavan[1]

[1] New York University, NY, NY, USA

E-mail for correspondence: ying.lu@nyu.edu

Abstract: We propose a practical yet novel solution to a longstanding statistical testing problem regarding single subject design. In the clinical setting, we evaluate whether a new patient behaves the same as one from a healthy population. This question cannot be answered using the traditional single subject design when only test subject information is used, nor can it be satisfactorily resolved by comparing a single-subject's data with the mean value of a healthy population without proper assessment of the impact of between and within subject variability. We use a training set of healthy subjects and a Bayesian framework to generate a template null distribution of the test statistic of interest. The performance of the proposed test such as false positive rate and power can be also readily evaluated. Notably, refitting of models with new subjects is unnecessary, and the single subject trial designs may differ from those of the healthy population, making this approach feasible in a tele-medicine situation.

Keywords: Hypothesis testing; multilevel models; Bayesian hierarchical models.

1 Introduction

Making an inference regarding a single subject is an important goal in clinical and applied settings and in health and behavioral research. Our example is a sensitive test for assessing hand function which has implications for a wide range of neurological conditions, such as stroke, multiple sclerosis and Parkinson's disease. Using a non-invasive device to measure muscular activity, interest focuses on the logarithm of the peak Load Force Rate (PLFR), a measure of predictive control during a grasping task, which, according to a study by Lu et al. (2015), increases linearly with the object's weight among healthy subjects. We have available a small training data set of healthy subjects, and a set of stroke patients in a clinic whose hand function needs to be assessed on an individual basis for diagnosis and treatment planning.

2 Model and Current Approaches

A linear hierarchical model (Laird & Ware, 1982) can be used to model this type of data. Since the training data and the test subject use different experimental designs and potentially belong to populations with different parameters, we outline the model for each group separately.

$$\begin{aligned} Y_{ijt} &= \alpha_i + \beta^{pop} W_{ij} + u_{ij} + \epsilon_{ijt}, && \text{(2.1) model for training data} \\ Y_{i'j't'} &= \alpha_{i'} + \beta_{i'}^{test} W_{i'j'} + u_{i'j'} + \epsilon_{i'j't'}, && \text{(2.2) model for test subject} \end{aligned}$$

where Y is the log-PLFR, subscript i and i' are subject ID, j and j' are weight ID and t and t' are trial ID, for the healthy training and patient test set separately. Note the design matrix W and the number of trails are not necessarily the same for the two datasets. According to Lu et al. (2015), although the level of log-PLFR, α_i, can vary across individuals (so we assume $\alpha_i \overset{i.i.d.}{\sim} N(a, \sigma_\alpha^2)$)), the scaling factor β^{pop} is best modeled as a fixed parameter across the healthy population. We are interested in testing:

H0: The patient i' has normal predictive control $\beta_{i'}^{test} = \beta^{pop}$.

Ha: The patient i' has suboptimal predictive control $\beta_{i'}^{test} < \beta^{pop}$.

To test the hypothesis, one option is to jointly model the two equations and test the difference in β via interaction terms using a Wald test under the MLE framework. However due to low sample size, this approach tends to suffer exaggerated type I error (Lu, Scott & Raghavan, 2016).
Alternatively one can utilize Bayesian inferential tools. For example, we can consider assigning the following priors and hyper-priors:

$$p(\beta_{(k)}, \sigma^2_{\epsilon_{(k)}}) \sim \frac{1}{\sigma^2_{\epsilon_{(k)}}},\ p(\alpha_i) \sim N(a_{(k)}, \sigma^2_{\alpha_{(k)}}),\ \sigma^2_{\alpha_{(k)}} \sim \text{Inv-Gamma}(\eta_{(k)}, \nu_{(k)})$$

where the subscript k denotes whether the parameters are for healthy training set ($k = 1$) and for patient test subject ($k = 2$). Given limited sample size, to borrow strength we choose to use non-informative priors whenever possible and constrain the prior parameters to be the same between the two equations. However, if additional information regarding the test subjects is available, different prior/hyper-priors can be used. Although inference based on a Bayesian p-value controls type I error at the desired level, for each single subject to be tested, one has to recalculate the entire model using an MCMC algorithm which is time consuming and impractical for real-time clinical assessment.

3 Proposed Framework For Single Subject Design

In this section we propose an inferential framework for single subject design that can be conveniently implemented in clinical settings. We first propose a *natural estimator* of the scaling factor β^{test} that is based on a simple concept of change in log PLFR per unit change in weight. Assuming a equally spaced weight design.

$$\bar{\beta}_{i'} = \frac{1}{\sum_{k=1}^{J-1} |\mathcal{I}_k|} \sum_{k=1}^{J-1} \sum_{(j,j') \in \mathcal{I}_k} \sum_{t=1}^{T} \frac{(y_{i'jt} - y_{i'j't})}{T(w_j - w_{j'})} \tag{3}$$

where $\mathcal{I}_k = \{(j, j') : j - j' = k\}$, w_j or $w_{j'}$ are weights associated with the corresponding indices, and T is the number of trials. For two weights, this is the naive estimator of the difference in PLFR per unit change in weight, averaged over trials.

Unlike directly comparing the naive estimator in equation (3) with a predetermined benchmark value and making a visual judgement about the status of the test subject, the use of Maximum Likelihood and Bayesian modeling allow us to compare the test subject with the training data set taking into account the within-subject and between-subject variability, and statistical tests are available to assist decision making. However, in order to make inference regarding a new subject, one needs to refit the entire multilevel model, which is not convenient in the clinical setting. Moreover since most of the parametric modelling approach depends on a large sample, the behavior of the aforementioned methods in hypothesis testing for a single subject is unknown. The Bayesian approach handles the non-asymptotic setting more elegantly, but is inherently more difficult to fit without specialized knowledge of statistical programming languages such as STAN, BUGS or JAGS. To address these concerns, we propose a novel approach that allows clinicians to make an informed decision about the test subject's status as compared to reference subjects in training data.

We start with the naive estimator $\bar{\beta}^{test}$ based on (3), which is available in the clinical setting. The goal is to provide the clinician with a template distribution of the possible values that we expect to observe given the weight design and the number of repeated measures used by the test subject. This template distribution is to be developed in a laboratory where the scientists and statisticians collaborate to design experiments and collect data based on a carefully controlled set of training subjects, for example, a random sample of healthy subjects. and based on it, the clinician can easily test the hypotheses described previously, regarding patients' anticipatory control (H0: $\beta^{test} = \beta^{pop}$; Ha: $\beta^{test} < \beta^{pop}$).

The probability of observing any values β^{test} as extreme as the naive estimate $\bar{\beta}^{test}$ had the test subject behaved the same way as the reference population can be easily generated using the template distribution. This probability has the interpretation of a classic p-value in a statistical inference problem (P(observation as extreme, given the model) under the null). The clinician can choose a desired level of the test, say 0.05 and reject the null hypothesis whenever the p-value is less than 0.05. An equivalent alternative is to compare the naive estimate $\bar{\beta}^{test}$ directly with the critical value $C_{0.05}(\beta^{test})$ derived from the template distribution. If $\bar{\beta}^{test} < C_{0.05}(\beta^{test})$

then reject the null hypothesis. Moreover, the performance (power) of such decisions can be evaluated ahead of time.
We construct *a template distribution* for inference based on the proposed natural estimator using the following algorithm.

1. We fit a Bayesian hierarchical model (2.1) using the training data set alone to obtain the posterior distribution of the parameters ($\Theta = \{a, \beta^{pop}, \sigma_\alpha^2, \sigma_u^2, \sigma_\epsilon^2\}$.

2. Given test-subject design W^{new}, we assume, under the null, that all parameters Θ in model (2.2) are the same. We then generate a set of posterior predictive outcomes $\tilde{y} \sim \text{MVN}(\mu^{new}, \Omega^{new})$ for the test subject, where $\mu^{new} = a + \beta^{pop} W^{new}$ and Ω^{new} has (possibly different) compound symmetry structure based on the new design.

3. We draw a large sample of pseudo-subjects from posterior $p(\tilde{y}|\Theta, W^{new})$ and compute $\bar{\tilde{\beta}}$ using equation (3). Its density approximates the null distribution of $\bar{\beta}$.

Given the linearity and normality assumptions, we can, in addition, easily generate posterior distributions under alternative hypothetical values of β^{test}. Since the generation of such template distributions depends only on the training data and the design of the test subject, rejection regions defined with respect to $\bar{\beta}$ can be given to the clinician prior to evaluating a new patient. Similarly, a clinician could be informed of the power associated with different magnitudes of deviation from the healthy population's rate.

4 Results

We conducted simulation studies (using clinically determined model parameters from the training dataset) comparing the proposed approach to several alternatives(Lu, Scott & Raghavan, 2016). In Table 1, $\delta = \beta^{pop} - \beta^{test}$ specifies different alternative situations. We use a training set design described subsequently and a hypothetical test subject design as follows: lifting three weights of 250g, 500g and 750g, with five trials each. The error rates of the proposed inference method are listed in Table one; type I error rate when $\delta = 0$ and type II error rate when $\delta \neq 0$. They are compared with the oracle situation where copies of $\bar{\beta}$ are directly simulated from the true distribution. The level of test is set to be 10%. We can see that our test is highly compatible to the oracle test with a slightly lower type I error than targeted. We also studied the power of the test under various test subject designs (varying the number of weights, the distance between different weights, and the number of trials). As Figure 1 shows, the distance between different pairs of the weights has much greater impact on the power than the number of different weights to be lifted.
We applied the proposed method to examine a group of patients with stroke from a single-subject design perspective. We are interested in understanding each patient's status as compared to a small group of healthy subjects

in terms of their ability to predict the fingertip forces to object weight as measured by the scaling factor for PLFR, in two different experiments. The data was collected using protocols approved by the Institutional Review Boards of Mount Sinai School of Medicine and New York University School of Medicine. All participants provided written informed consent as approved by the IRB. The data is described next.

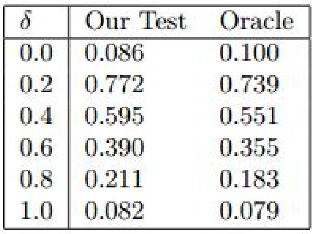
The training data set consists of data from 10 healthy subjects, each lifting 10 weights, ranging from 250 grams to 750 grams, 50 grams apart. The order of the weights is randomized to avoid an ordering effects. Each subject lifts each weight 6 times after one practice trial to learn the weight of the object. There are 22 test subject patients with stroke who are at various stages of post-stroke recovery. Each of the patients participate in two experiments to assess their fingertip force coordination with the affected hand. In experiment one, the patients lift weights with affected hand, at 550 and 800 grams each. In two, they lift weights with the affected hand following a practice lift with unaffected hand, 350 and 600 grams. There are 4 trials after a practice lift. To avoid an ordering effect, the experimental conditions and weights within each condition are randomized.

Some patients suffer from sensory impairment in the affected hand, hence they may not be able to learn the weight of the object through practice using the affected hand alone only (experimental condition one), instead, such information may be learned by practicing with unaffected hand first (experimental condition two) (Raghavan, Krakauer, & Gordon, 2006). Ultimately, clinicians and researchers need this information to decide whether the patients should practice the grasping task with the affected hand alone, or incorporate the unaffected hand into practice protocols.

Using the data from test subjects, we first fit a linear hierarchical model to test the effects of experimental condition two as compared to condition one. At level 10%, we found that the results of experimental condition two were significantly better in terms of the scaling factors. However these results from the entire group are not particularly useful when the clinician needs to make a decision and a recommendation of a practice protocol for any single patient during the course of their rehabilitation.

Using the proposed method, we can assess each subject separately under each experimental condition. Since most stroke rehabilitation treatment protocols are non-invasive and low-risk, we choose to tolerate a higher false positive rate, and set the level of the test to be at 10%. The assessment results are therefore controlled at an expected 10% false positive rate; the p-value and the (post hoc) power of each assessment is also estimated and reported. This table reflects all the information available to the clinician.

Based on the test results, setting $\alpha = 0.10$, we find that out of 18 subjects who completed both experimental conditions, six subjects switched status from “ABNORMAL” under condition one to ”NORMAL” under condition two, one subject remained “ABNORMAL” in both conditions. Nine subjects remained “NORMAL” and two switched from “NORMAL”



δ	Our Test	Oracle
0.0	0.086	0.100
0.2	0.772	0.739
0.4	0.595	0.551
0.6	0.390	0.355
0.8	0.211	0.183
1.0	0.082	0.079

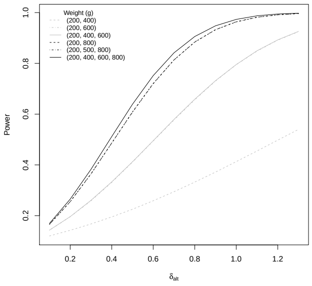


FIGURE 1. Table (left): The error rates of the proposed test and the oracle test ; Figure (right): The power of the proposed test for different design scenarios.

to "ABNORMAL". Clinicians can thus use these results to design customized training protocols for each patient. Moreover, among those who receive an initial "NORMAL" assessment in experimental condition one, the information of post-hoc power and p-value can further inform clinicians about how effective the test is in detecting "ABNORMAL" status given the observed effect size, and the minimal false positive rate they have to accept if they choose to switch a "NORMAL" patient to the "ABNORMAL" status. Combining the information provided by the observed scaling factor, the power of the test, the p-value of the test and other patient-specific conditions, a clinician can make an informed choice to assign a particular patient to an appropriate training paradigm.

References

Lu, Y., Bilaloglu, S., Aluru, V., Raghavan, P. (2015). Quantifying Feedforward Control: A Linear Scaling Model for Fingertip Forces and Object Weight, Journal of Neurophysiology,Epub 2015. doi: 10.1152/jn.00065.2015

Lu, Y., Scott, M. and Raghavan, P. (2016). A Statistical Framework for Single Subject Design with an Application in Post-stroke Rehabilitation. arXiv:1602.03855 [stat.AP]

Gelman, A., and Carlin, J.B. (2013). *Bayesian Data Analysis.* Chapman & Hall/CRC Texts in Statistical Science, 3rd edition.

Laird, N.M., and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963974.

Raghavan, P., Krakauer, J. W., and Gordon, A. M. (2006). Impaired anticipatory control of fingertip forces in patients with a pure motor or sensorimotor lacunar syndrome. Brain, 129:14151425.

Rstan (2015). RStan: the R interface to Stan, Version 2.8.0. http://mcstan.org/rstan.html

A flexible multi-state model for aneurysm data

Robson J. M. Machado[1], Ardo Van den Hout[1], Giampiero Marra[1]

[1] University College London, Department of Statistical Science, United Kingdom

E-mail for correspondence: robson.machado.14@ucl.ac.uk

Abstract: In this work a flexible method is presented to model multi-state processes with interval-censored observation times. P-splines are used to model the progression of aortic diameter in elderly men patients. The method uses a large fixed number of knots to describe the multi-state process. Penalised likelihood is used to estimate the parameters of the model. The application to the aortic aneurysm progression (aneur) data provides insightful information about the rates of diameter aortic progression over time.

Keywords: Multi-state models; P-splines.

1 Introduction

Abdominal aortic aneurysm is common in elderly men in the UK. The aneur data consist of measurements of grades of aortic aneurysms, measured by ultrasound of the diameter of the aorta. The states represent successive degrees of aneurysm severity, as indicated by the aortic diameter. States are defined as follows: healthy (1), less than 30 mm; mild (2), $30-44$ mm; moderate (3), $45-54$ mm; severe (4), 55 mm and above. Modelling transition process is important because screening policies are defined w.r.t. the risk of moving across states. Severe aneurysms are repaired by surgery. Multi-state models can be used to estimate risk of diameter progression. They are specified by transition hazards. A four-state progressive model for diameter aortic progression is illustrated in Figure 1.

The aneur data include transitions moving from a higher state to a lower one, due to misclassification of state. Jackson et al. (2013) analysed the aneur data using a hidden multi-state Markov model. This class of 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).

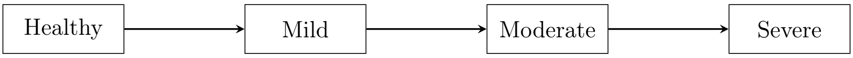


FIGURE 1. Four-state model for aneurysm progression

can estimate transition rates and probabilities of state misclassification. Hazard transitions were constant over time.

For this application, the aneur data are defined for the history of disease. This means that the recorded state is the highest state observed. All individuals are aged 60 at the beginning of the study. For the analysis, age is shifted by minus 59 years. The hazard transitions are specified by P-splines basis functions to allow for more flexibility (Eilers and Marxs, 1996). Estimation is undertaken using maximum penalised likelihood.

2 Continuous-time multi-state models

Let $\{X(t),\ t > 0\}$ be a continuous-time Markov process which takes values in the discrete state space $\mathcal{S} = \{1, \ldots, m\}$. If $X(t)$ is time-homogeneous transition probabilities are given by

$$p_{rs}(t) = P(X(t) = s|X(0) = r), \tag{1}$$

and the transition probability matrix is defined by $\mathbf{P} = (p_{rs})$, for $r, s \in S$. The transition hazards are given by

$$q_{rs}(t) = \lim_{\Delta t \to 0} \frac{P(X(t + \Delta t) = s|X(t) = r)}{\Delta t}, \tag{2}$$

for $r \neq s$, and can be used to derive the transition probabilities. Let $q_{rr}(t) = -\sum_{s \neq r} q_{rs}(t)$ for all $r \in \mathcal{S}$ and define the transition intensity matrix by $\mathbf{Q} = (q_{rs})$. Subject to the initial condition $\mathbf{P}(0) = \mathbf{I}_m$, it is known that $\mathbf{P}(t) = \exp(t\mathbf{Q})$; see Kalbfleisch and Lawless (1985). Given time-dependent hazards, transition probabilities for the likelihood function are derived using a piecewise-constant approximation.

The hazard model for transition from state r to state s is

$$q_{rs}(t) = q_{rs.0}(t) \exp(\boldsymbol{\beta}_{rs}^{\mathrm{T}} \mathbf{z}), \tag{3}$$

where $\mathbf{z}$ is a covariate vector, $\boldsymbol{\beta}_{rs}$ is a parameter vector and $q_{rs.0}(t)$ is the baseline hazard function. Let K be the number of P-splines bases to model transition from state r to state s. Then,

$$q_{rs.0}(t) = \exp(\sum_{k=1}^{K} \alpha_{rs.k} B_k(t)). \tag{4}$$

3 Penalised log-likelihood function

Suppose there are J transitions to be smoothed and K_j knots for each transition $j = 1, \ldots, J$. Writing $\boldsymbol{\theta}$ for the full set of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and let $L(\boldsymbol{\theta})$ be the usual log-likelihood for model (3) with baseline hazard functions parametrised by (4) for interval-censored data. This is an extension of the log-likelihood in Kalbfleisch and Lawless (1985). Then the penalised log-likelihood function is

$$L_p(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \sum_{j=1}^{J} \lambda_j \boldsymbol{\alpha}_j^\top \mathbf{D}_j^\top \mathbf{D}_j \boldsymbol{\alpha}_j, \tag{5}$$

where $\mathbf{D}_j$ is the matrix representation of the difference operator for $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jK_j})^\top$.
Estimation is undertaken using the function optim in R. The criteria to select the optimal set of smoothing parameters λ_j is the Akaike Information Criterion (AIC) which is defined in this context by

$$\text{AIC} = -2 \cdot L_p + 2 \cdot df, \tag{6}$$

where df is the overall model degrees of freedom (Gray, 1992).

4 Application

For the aneur data, consider the four-state progressive model illustrated in Figure 1. The model for transition from state r to state s is

$$q_{rs}(t) = \exp(\sum_{k=1}^{K_j} \alpha_{rs.k} B_k(t)), \tag{7}$$

where $j = 1, 2, 3$. For transition from state 1 to state 2 there are $K_1 = 18$ cubic P-splines bases. For the other transitions, simpler models are defined with $K_2 = K_3 = 3$ quadratic P-splines bases. The log-likelihood is penalised for the first transition parameters. Table 1 shows that the optimal smoothing parameter is $\lambda = 10^{-3}$. The fitted hazard from state 1 to state 2 is illustrated in Figure 2. It shows that the risk of moving from state 1 to state 2 increases 15 years after the beginning of the study.

TABLE 1. Values of AIC and df for several values of λ

λ	10^{-8}	10^{-3}	10^{-2}	10^{-1}	2×10^{-1}	5×10^{-1}
AIC	1739.254	1736.989	1736.992	1738.684	1739.518	1740.493
df	22.95	20.30	19.36	18.43	18.14	17.78

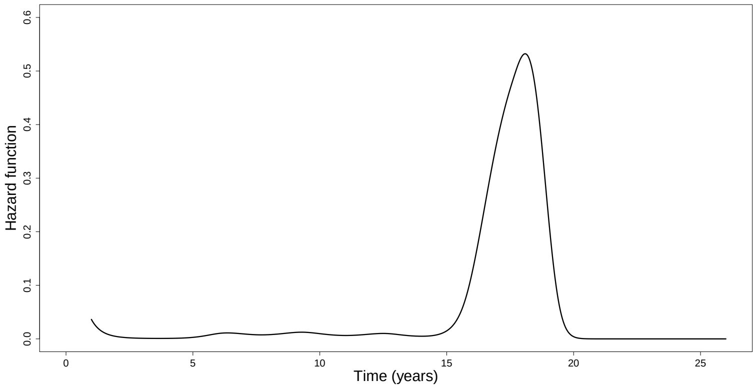


FIGURE 2. Fitted hazard transition from state 1 to state 2

4.1 Prediction

Interpretation of the estimated model is straightforward using transition probabilities. They can be calculated by using a piecewise-constant approximation. Let $\mathbf{P}(t_1, t_2)$ denote the transition probability matrix for any time interval $(t_1, t_2]$. Given the grid $t_1, \ldots, t_n$, the transition probability matrix for the interval $(t_1, t_n]$ is defined by

$$\mathbf{P}(t_1, t_n) = \mathbf{P}(t_1, t_2) \times \ldots \times \mathbf{P}(t_{n-1}, t_n), \tag{8}$$

where the matrices at the right-hand side are derived using transition intensity matrices $\mathbf{Q}(t_1), \ldots, \mathbf{Q}(t_{n-1})$. The transition probabilities for 25 years after the beginning of the study is estimated at

$$\widehat{\mathbf{P}}\,(t_1 = 1, t_2 = 26) = \begin{pmatrix} 0.253 & 0.165 & 0.241 & 0.341 \\ 0.000 & 0.088 & 0.147 & 0.765 \\ 0.000 & 0.000 & 0.001 & 0.999 \\ 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}, \tag{9}$$

where t denotes age transformed by subtracting 59 years.

The interpretation of this matrix is as follows. An individual aged 60 in the healthy state has a probability of 0.341 for developing severe aneurysm 25 years from now and a probability of 0.253 for being still disease-free. For the same time interval, an individual with mild aneurysm has a probability of 0.765 for moving to severe state and with moderate aneurysm a probability of 0.999 for developing severe aneurysm.

Figure 3 illustrates estimated probability transitions from state 1 to 2, from state 1 to 3 and from state 1 to 4. As indicated in Figure 2, the risk of moving from state 1 to state 2 has a steep increase for individuals aged 75. To a lesser extent, the risks of moving from state 1 to states 3 and also increase at around the same time.

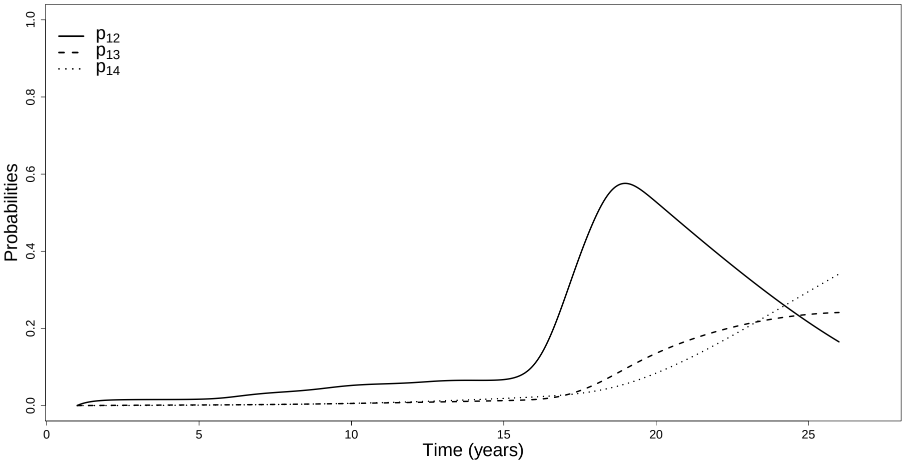


FIGURE 3. Transition probability for 1 → 2, 1 → 3 and 1 → 4 conditional on being in state 1 at age 60

5 Comments

As a next step, we aim to set a large number of P-splines bases to model each transition. A grid search can be used to select the three smoothing parameters. However, this method is computationally expensive. A more efficient alternative is to employ an automatic smoothing parameter selection (Wood, 2006). We aim to implement this method for the presentation in the summer.

Acknowledgments: Special Thanks to Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq. 249308/2013-4.

References

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

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, **87(420)**, 942 − 951.

Jackson, C. H., Sharples, L. D., Thompson, S. G., Duffy, S. W., Couto, E. (2003). Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society, Series D (The Statistician)*, **52(2)**, 193 − 209.

Kalbfleisch, J.D., Lawless, J.F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, **80**, 863–871.

Wood, S. (2006). *Generalized additive models: an introduction with R.* CRC press.

Lasso model selection in multi-dimensional contingency tables?

Gilbert MacKenzie[1], Susana Conde[2]

[1] ENSAI, France and University of Limerick, Ireland
[2] Manchester University, UK. and The University of Glasgow, UK

E-mail for correspondence: `gilbert.mackenzie@ul.ie`

Abstract: We develop a Smooth Lasso for sparse, high dimensional, contingency tables and compare its performance with the usual Lasso and with the now classical backwards elimination algorithm. In simulation, the usual Lasso had great difficulty identifying the correct model. Irrespective of the sample size, it did not succeed in identifying the correct model in the simulation study! By comparison the smooth Lasso performed better improving with increasing sample size. The backwards elimination algorithm also performed well and was better than the Smooth Lasso at small sample sizes. Another potential difficulty is that Lasso methods do not respect the marginal constraints on hierarchy and so lead to non-hierarchical models which are unscientific. Furthermore, even when one can demonstrate, classically, that some effects in the model are inestimable, the Lasso methods provide penalized estimates. These problems call Lasso methods into question.

Keywords: False estimation, Lasso, Model selection, Non-hierarchical models, Smooth Lasso

1 Introduction

Sparse contingency tables arise often in genetic, bioinformatic and database applications. Then the target is to estimate the dependence structure between the variables modelled via the interaction terms in a log-linear model. High dimensionality will force attention on identifying important low-order interactions - a technical advance since most model selection work relies only on main effects.
Penalized likelihood attaches a penalty function of the parameters to the likelihood in order to achieve some purpose such as smoothing (Eilers and Marx, 1996), or sparsity (Freidman, 2008). Using the LASSO (L_1-norm

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).

penalty), some of the parameters go to zero allowing a more parsimonious model to be found. Dahinden (2007) extended the LASSO (Tibrishani, 1996) to contingency tables and log-linear models. However, in the Lasso the penalty is a non-differentiable function of the parameters thus necessitating specialized optimization algorithms.
We present the smooth LASSO, a penalized likelihood, which does not require specialized optimization algorithms such as the method of coordinate descent. It uses a convex, parametric, analytic penalty function that asymptotically approximates the LASSO: minimization is accomplished using standard Newton-Raphson algorithms and standard errors are available.

2 Model Formulation

2.1 Log-linear modelling

Assume $X_1, \ldots, X_v$ correlated binary variables (off=0, on=1) and these form a v-dimensional contingency table with $q = 2^v$ cells. Let Y_i be the random variable indicating the frequency in the ith cell, $i = 1, \ldots, q$ and let $\mu_i = E(Y_i)$. We consider a log-linear regression model: $\log(\mu) = A^{\mathrm{T}}\theta$ where A is a $(q \times p)$ design matrix of fixed constants with typical element a_{ij}, and θ is a vector with p dimensions measuring the influence of the effects (constant, main effects and interactions) on the response vector of counts Y. We use Yates' design matrix coding scheme whence the columns of A are orthogonal. Finally, let $n = \sum_{i=1}^{q} Y_i$ denote the total number of observations. Estimation is via the log-likelihood, which may be taken in Poisson form: $\ell(\theta \,|\, y) \propto \sum_{i=1}^{q}\{y_i(a_i^T\theta) - \exp(a_i^T\theta)\}$, as the maximum likelihood estimators are the same in multinomial and independent Poisson schemes provided $\sum_{i=1}^{q} \mu_i = n$ (Birch 1963). The log-likelihood may be maximized numerically using iterative proportional fitting or by generating the design matrix A and using the `nlm` procedure in the R software package.

2.2 A Smooth LASSO

The penalized log-likelihood is:

$$\ell_\lambda(\theta) = \ell(\theta) - \mathrm{pen}_\lambda \tag{1}$$

where pen_λ, is the penalty term, $\lambda > 0$. For the LASSO $\mathrm{pen}_\lambda = \lambda \sum_{j=2}^{p} |\theta_j|$ omitting the intercept term and for the Smooth LASSO $\mathrm{pen}_\lambda = \lambda \sum_{j=2}^{p} Q_\omega(\theta_j)$ where $Q_\omega(\theta_j) = \omega \log\left[\cosh\left(\frac{\theta_j}{\omega}\right)\right]$ for a constant ω that regulates the approximation of the function to that of the absolute value function (Salje et al, 2005). Note that $Q_\omega(\theta_j) \in \mathcal{C}^\infty$, the set of functions that are infinitely differentiable, and is convex. Following we define the maximum penalised likelihood estimator (MPLE) as

$$\hat{\theta} := \arg\max_{\theta \in \Theta} \{\ell(\theta) - \mathrm{pen}_\lambda(\theta)\} \,. \tag{2}$$

We should more properly write $\hat{\theta}_\lambda$, rather than $\hat{\theta}$, but the dependence on λ will be understood in what follows. For a large λ, all the estimates go to 0 and for $\lambda = 0$, there is no constraint, whence $\hat{\theta}_{\lambda=0}$ is equivalent to the usual maximum likelihood estimator (MLE).

3 Non-hierarchical model

We digress to make an important methodological point by comparing Yates' and Binary design matrix coding schemes in a non- hierarchical model using a well known example. Agresti (2002) gave the following 2^3 table $y' = (19, 11, 0, 6, 132, 52, 9, 97)$ of counts classified by: A = defendant's race (0. white, 1. black), B = victim's race (0. white, 1. black) and C = death penalty (0. yes, 1. no). The contingency table is written in vector notation in which the leftmost subscript varies fastest. Table 1 shows the result of fitting the non-hierarchical model A, AB, AC with with Yates' ($\hat{\gamma}$) and Binary ($\hat{\beta}$) design matrices. We have the same data, the same model, but the likelihoods differ and the effects have different interpretations in the two models. This simple example shows that we should restrict model selection to hierarchical models.

TABLE 1. Comparison of Yates' and binary coding schemes when fitting a non-hierarchical model comprising A, AB, AC.

		Estimated	Quantities			
Parameters	$\hat{\gamma}$	$\hat{\beta}$	$se_{\hat{\gamma}}$	$se_{\hat{\beta}}$	$z_{\hat{\gamma}}$	$z_{\hat{\beta}}$
θ_0	3.520	3.689	0.067	0.079	52.714	46.670
θ_A	0.018	−1.825	0.055	0.274	0.332	−6.666
θ_{AB}	0.630	0.492	0.067	0.160	9.442	3.074
θ_{AC}	0.031	2.171	0.055	0.256	0.554	8.480

$$\ell(\hat{\gamma}) = -161.7495, \qquad \ell(\hat{\beta}) = -150.65$$

Even when fitting hierarchical models, only effects in the generating set of the fitted model are invariant to the choice of design matrix. The application of Wald tests to other effects is mistaken. The likelihoods, however, are invariant. These findings apply to *all* statistical models with interaction terms.

TABLE 2. Simulation: Percentage of correct models identified by three methods.

		Estimation	Methods
Sample size	BE	Lasso	SL-95
50	62.3	0*	0.1
100	51.6	0	11.8
500	33.0	0	50.1
1000	29.2	0	51.6

* The Lasso persistently over fits effects.

4 Lasso Model Selection

4.1 Simulation

We conducted a small simulation study designed to study the percentage of correct models identified by three algorithms: Backwards Elimination, the usual Lasso and the Smooth Lasso. For the purposes of illustration we simulated a 2^5 contingency table when the main effects model was true. The number of replications was $m = 1000$ and we started with the all 2-way interactions design-matrix. For the backwards elimination method we used a R function written Conde (2011), for the usual Lasso we used the `glmnet` R package and for the Smooth Lasso we used another R function, which called `nlm`. The tuning parameter λ was estimated by 10 fold cross-validation in the Lasso functions. The sample sizes studied were: $n = 50, 100, 500, 1000$. For the Smooth Lasso one must pick a level of statistical significance, as with ordinary regression methods (Conde & MacKenzie, 2010). Thus SL-95 corresponds to the 5% level. It will be noticed that the 5% level produces poor results when the sample size is small, but improves with increasing sample size, while the classical Backward Elimination algorithm performs better for smaller sample sizes.

4.2 Obesity Data Analysis

We now present the results of analysing a set of obesity data comprising 8 binary comorbidities measured on $n = 5550$ patients. The resulting contingency table has 2^8 cells of which 45.3% are zero cells. We compare the three algorithms described above using the same fitting methods. Table 3 presents the *generating sets* defining the final models together with their AICs. Several interesting features emerge.

First the fitted Lasso-based solution comprised non-hierarchical models. Each non-hierarchical model was then augmented by adding in effects to produce a minimum hierarchical model. The models were re-estimated (Table 3). Unfortunately, this idea does not always work - often, in sparse

TABLE 3. Generating sets of models found by Backwards Elimination, LASSO and Smooth LASSO.

	BE	LASSO*	SL-95*
Model	[c1c6, c1c8, c2c3, c2c4, c2c5, c3c4, c5c6, c1c4, c4c5c7, c4c6c8, c6c7c8]	[c1c2c4, c1c2c7, c1c3c7, c1c3c8, c1c5c6, c1c5c7, c1c5c8, c1c6c7, c1c6c8, c1c7c8, c2c3c5, c2c3c6, c2c3c7, c2c4c7, c2c4c8, c2c6c8, c3c4c6, c3c4c7, c3c4c8, c3c5c6, c3c6c8, c4c5c6, c4c5c7, c4c5c8, c4c6c7, c4c6c8, c4c7c8, c5c6c8,c5c7c8, c6c7c8]	[c1c6, c1c7, c1c8, c2c4, c2c5,c3c4, c3c6, c4c5, c4c8, c6c7,c6c8]
AIC	722.687	749.831	1254.699

*Minimal hierarchical model that includes the effects in the support. For the smooth LASSO, $\omega = 1$.

tables, one finds that minimum hierarchical model contains effects which are non-estimable, whence one is stuck with a Lasso solution which is non-hierarchical. Such solutions are unscientific.

A second problem arises with the Lasso methods investigated. If one preprocesses the table one can identify effects which are inestimable in the classical paradigm (using a theorem due to the first author). On first noticing this we hoped that if the Lasso was going to produce a sparse model it would somehow identify the inestimable effects and shrink these to zero. However this is not the case and we have many examples of the Lasso and Smooth Lasso solutions producing penalized estimates of inestimable effects. One might be tempted to regard this as an "advantage", but this seems näive. The solution is inconsistent with the classical theory. One possible explanation is that the penalized likelihoods have a Bayesian interpretation in which the penalty plays the role of a prior. So false estimation of inestimable effects may just correspond to a value assigned by the prior. If so, this is yet another reason for discarding such solutions.

Accepting these caveats, we note that: (a) the BE algorithm always produces a hierarchical model, (b) the BE algorithm is best as judged by the AIC, (c) it is also fastest, (d) the Lasso is not the sparsest model and (e) the smooth LASSO is much more parsimonious than the LASSO. These are consistent findings in our work.

5 Discussion

There is, apparently, a highly impressive literature on Lasso methods. It is, however, predicated on model selection based on main effects models. In the presence of interactions, Lasso methods will often fail to produce scientific models. It has been argued that group Lasso methods provide one answer to this problem, but they require multiple tuning parameters, one for each class of interactions *anticipated* in the final solution. Accordingly, they are prohibitively computationally expensive. Other authors have argued for weak hierarchy (Bien et al, 2013). Their arguments are not compelling and difficult to implement. Moreover, it is well known that the Lasso lacks the *oracle property* and the results in Table 2 confirm this. However, the results suggest that this may not be the case for the Smooth Lasso, a finding which requires further investigation. To our knowledge the problem of false estimation has not previously been reported. All these issues raise serious questions about the usefulness of Lasso methods for model selection.

Acknowledgments: We acknowledge the help and support of Professor Victor Kiri, of FVJK consultants, UK. The second author works on the MIMOMICS project in the Centre of Biostatistics in Manchester University, UK

References

Agresti, A. (2002). Categorical Data Analysis. John Wiley & Sons, Inc., Hoboken, New Jersey:Wiley-Interscience, 2nd ed.

Bien, J., Taylor, J. and Tibrishani, R. (2013). A lasso for hierarchical interactions. The Annals of Statistics 41, 11111141.

Birch, M. W. (1963). Maximum Likelihood in Three-Way Contingency Tables. Journal of the Royal Statistical Society. Series B (Methodological) 25, 220233.

Conde, S. (2011). Interactions: Log-Linear Models in Sparse Contingency Tables. Ph.D. thesis, University of Limerick, Ireland.

Conde, S. & MacKenzie, G. (2011). LASSO Penalised Likelihood in High-Dimensional Contingency Tables. In Proceedings of the 26th International Workshop on Statistical Modelling, Valencia, D. Conesa, A. Forte, A. et al. eds.

Dahinden, C., Parmigiani, G., Emerick, M.C. and Bühlmann, P. (2007). Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics*, **8:476**.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science*, **11(2)** 89-121.

Friedman, J.H. (2008). Fast Sparse Regression and Classification. In: *Proceedings of the 23rd International Workshop on Statistical Modelling*, Utrecht, 27-57. Ed.: Eilers, P.H.C.

Salje, Ekhard K. H., Hayward S.A. and Lee W.T. (2005). Ferroelastic phase transitions:structure and microstructure. Acta Crystallographica Section A 61, 318.

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

NCE: Nonparanormal Causal Effect

Seyed Mahdi Mahmoudi[1], Ernst C. Wit[1]

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

E-mail for correspondence: s.m.mahmoudi@rug.nl

Abstract: We present the general functional form of causal effect in a large subclass of non-Gaussian distributions, called Nonparanormal Causal Effect (NCE). By describing the causal network as a directed acyclic graph it is a possible to estimate a class of Markov equivalent systems that describe the underlying causal interactions consistently, even for non-Gaussian systems. In these systems, causal effects stop being linear and cannot be described anymore by a single coefficient. A statistical analysis of the properties of NCE is given together with empirical results on synthetic and real data, showing that NCE can be effective in estimation nonparanormal observational data.

Keywords: Causal effects, Directed acyclic graph (DAG), Graphical modelling, Nonparanormal distribution, PC-algorithm.

1 Intrudoction

Substantial progress has been made recently on the problem of estimation of the causal structure and the interventional distribution in the Gaussian case (Maathuis et al., 2009). Due to the Gaussian structure, they finds that the causal effect can be described by a set of constants. Harris and Drton (2013) show that the PC-algorithm has high-dimensional consistency properties for a broader class of distributions, when standard Pearson-type empirical correlations are replaced by rank-based measures of correlations in tests of conditional independence, such as Spearmans rank correlation and Kendalls rank. The broader class they consider includes continuous distributions with Gaussian copula, or, in the terminology of Liu et al. (2012), the so-called "nonparanormal distributions." In this paper, we assume the use of the 'Rank PC' (RPC) algorithm Harris and Drton (2013), i.e. the PC-algorithm in the nonparanormal context. Based on the estimated CPDAG, it is our aim to derive the concept of a causal effect of x on y as a collection of functions of x and to find an efficient way to estimate them. In Section

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 we show the structure of a causal effect of a nonparanormal causal effect and define an efficient estimator. In Section 3, we evaluate the performance of our method in a simulation study.We also applied the method to an *Arabidopsis Thaliana* circadian clock network in Section 4.

2 NCE for Graphical Models

If $(X_1, ..., X_p, Y)$ has a multivariate Gaussian distribution, it is very simple to compute the causal effects as

$$E(Y|\mathrm{do}(X_i = x)) = \beta_0 + \beta_i x + \beta_{pa_i}^T pa_i, \tag{1}$$

and, therefore, the intervention effect, or causal effect, becomes

$$\frac{\partial}{\partial x} E[Y|\mathrm{do}(X_i = x)]|_{x=x_i'} = \beta_i \tag{2}$$

Causal effect of X_i on Y with $Y \notin pa_i$ is given by the regression coefficient of X_i in the regression of Y on X_i and pa_i. Note that if $Y \in pa_i$, the causal effect from X_i to Y is, obviously, zero. Our aim is to generalize this to a wider class of nonparanormal distributions. Liu et al. (2012) define the nonparanormal distribution, which is identical to a Gaussian copula distribution. Let $f = (f_i)_{i \in \mathbf{V}}$ be a set of monotone, univariate functions and let $\Sigma \in \mathbb{R}^{\mathbf{V} \times \mathbf{V}}$ be a positive definite covariance matrix. We say a p-dimensional random variable $X = (X_1, ..., X_p)^{\mathbf{T}}$ has a nonparanormal distribution,

$$X \sim \mathrm{NPN}(\mu, \Sigma, f),$$

if $f^{-1}(X) = (f_1^{-1}(X), \ldots, f_p^{-1}(X)) \sim N(\mu, \Sigma)$. If $X \sim \mathrm{NPN}(\mu, \Sigma, f)$, then the univariate marginal distribution for a coordinate, say X_i, can have any distribution F_i, as we can take $f_i = F_i^{-1} \circ \Phi$, where Φ is the standard normal distribution function. Let consider that $(X_1, \ldots, X_{p-1}, Y) \sim \mathrm{NPN}(0, \Sigma, f)$. We will refer to the latent standard normally distributed variables as $Z_i = f_i^{-1}(X_i) = \Phi^{-1} \circ F_i(X_i)$ and $Z = f_y^{-1}(Y) = \Phi^{-1} \circ F_y(Y)$. We are interested in the causal effect of X_i on Y for $i \in (1, \ldots, p-1)$, that from (2), we know that for Gaussian data it is very simple to compute the causal effect, since Gaussianity implies that $E(Y|X_i = x_i; X_{-i} = x_{-i})$ is linear in x_i. Unfortunately, this is no longer true for non-Gaussian random variables. In Propostion 1 we derive the explicit functional form for the causal effect in the entire class of nonparanormal distribution.

Proposition 1 *Let $(X_1, \ldots, X_{p-1}, Y) \sim NPN(0, \Sigma, f)$ and f_i $(i = 1, \ldots, p-1)$ is differentiable and f_y is infinitely differentiable, then the causal effect of X_i on Y in causal graph G is given by*

$$\begin{aligned} CE(Y|X_i = x_i) \quad &= \quad \sum_{k=0}^{\infty} \sum_{r=0}^{\lfloor \frac{k-1}{2} \rfloor} \sum_{s=1}^{k-2r} f_y^{(k)}(z_0) \frac{1}{k!} \binom{k-2r}{s} \binom{k}{2r} s\beta_i(-z_0 + \beta_i z_i)^{s-1} \quad (3) \\ &\times \quad E[(\beta_{pa(i)}^T Z_{pa(i)})^{k-2r-s}](2r-1)\ldots 3.1 \times [(1-\rho^2)]^r (f_i^{-1})'(x_i) \end{aligned}$$

for every $z_0 \in \mathbb{R}$*, where* $f_y^{(k)}$ *is the kth derivative of* f_y*,* $z_i = f_i^{-1}(x_i)$*,* $Z_{pa(i)} = f_{pa(i)}^{-1}(X_{pa(i)})$*,* $(\beta_i, \beta_{pa(i)}) = \Sigma_{p,(i,pa(i))}\Sigma^{-1}_{(i,pa(i)),(i,pa(i))}$ *and* $\rho = \Sigma_{p,(i,pa(i))}\Sigma^{-1}_{(i,pa(i)),(i,pa(i))}\Sigma_{(i,pa(i)),p}$*.*

If we assume that the underlying function f can be appropriately be described by a cubic spline, then in terms of estimation, the terms $f^{(k)}$ can be set to zero for $k \geq 4$. This would reduce the infinite sum in (3) to a sum of merely four terms. These terms, however, still require some estimates of ρ and the various moments of $Z_{\text{pa}(i)}$. By selecting z_0 carefully, it is possible to find good approximations even for lower order Taylor expansions. In particular, we found that by setting $z_0 = 0$, the mean of the latent response $Z = f^{(-1)}(Y)$ the first order Taylor expansion was already quite appropriate to capture non-linear causal effects for a wide ranging collection of distributions. Using (3), the first order Taylor expansion of intervention effect effect for the nonparanormal response is given by

$$\begin{aligned} E(Y|\text{do}(X_i = x_i)) &= \int E(Y|x_i, x_{\text{pa}(i)}) f(x_{\text{pa}(i)})\; d(x_{\text{pa}(i)}) \\ &\approx C + f_y'(0)\beta_i f_i^{-1}(x_i) \qquad (4) \end{aligned}$$

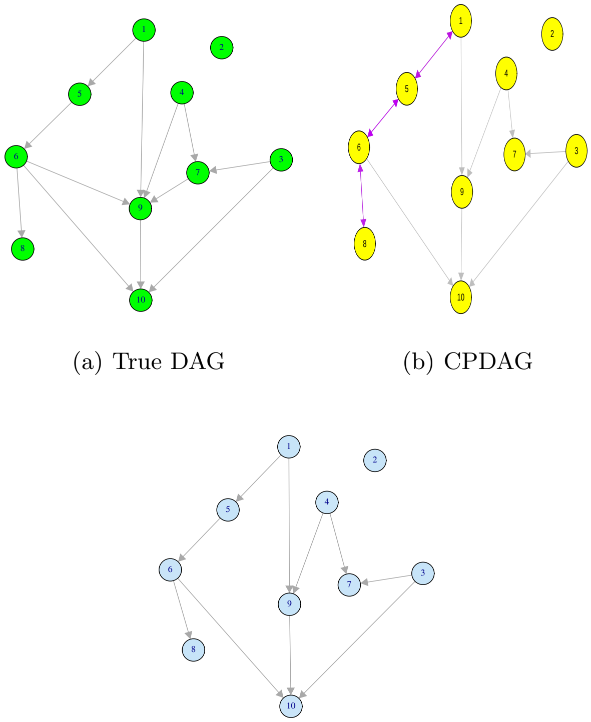
where C is some constant. From (4), we can obtain a simple plug-in estimator for the causal effect,

$$\text{CE}(Y|X_i = x_i) \;=\; f_y'(0)\; \beta_i\; (f_i^{-1})'(x_i), \qquad (5)$$

where β_i is the linear regression coefficient of $f_y^{-1}(Y)$ on $f_i^{-1}(X_i)$ controlling for the parents $f_i^{-1}(X_{\text{pa}(i)})$ of i . The latent values are, obviously, not observed and need to be reconstructed, together with the marginal distributions of Y and X_i.

3 Simulation studies

In this section, we test our estimation method for Gaussian distributions. For Gaussian data, the method should find constant causal effects and can be compared directly with the IDA method Maathuis et al. (2009). We consider two scenario for Guassian data. A small graph on ten vertices and a larger graph on fifty vertices with an expected vertex degree of three. For each $n \in \{100, 1000\}$ and each of the Two types of graphs above, we sample 100 random graphs from both the small and large graph distributions, and then sample n observations from the graph with the normal data distribution. For each resulting combination, we run each of the two considered versions of the RPC-algorithm on a grid of α 's ranging from 10^{-10} to 0.5. For each estimated DAG, we compute the causal effects of each nodes by our estimator and the compare with IDA method. For illustration purposes, Figure1 shows the true graph, the estimated CPDAG and one of three equivalent DAGs for the p = 10 scenario. Table 1 show the results for two scenario which we compere our method with IDA.



(a) True DAG

(b) CPDAG

(c) One of three DAGs consistent with CPDAG

FIGURE 1. Plots generated using the R-package `pcalg`. (a) The true DAG. (b) The estimated CPDAG using the PC-algorithm, based on a a simulated dataset with $n = 1,000$ replicates and significance cut-off $\alpha = 0.01$. (c) One of the three DAGs consistent with CPDAG.

4 Real data analysis

The data consist of transcription profiles for the core clock genes from the leaves of various genetic variants of Arabidopsis thaliana, measured with qRTPCR. We consist of two groups of genes: Morning genes, which including LHY, CCA1, PRR9, and PRR5 whose expression peaks in the morning. Evening genes, include TOC1, ELF4, ELF3, GI, and PRR3 whose expression peaks in the evening. The causal effect network among the genes and functional relation between them are displayed in Figure 2.
There are several directed genes pointing from morning genes to evening genes and vise-versa. Some of the genes play important roles in the circadian clock network. In this work, we analysed the causal effect between genes

TABLE 1. Results of the simulation study for comparison our method(NCE) and IDA for small graph ($p = 10$) and large graph($p = 50$) when the data are Guassian.

		$\alpha = 0.01$		$\alpha = 0.1$	
$\boldsymbol{p}$	$\boldsymbol{n}$	**IDA**	**NCE**	**IDA**	**NCE**
Small graph 10	100	0.101	0.576	0.144	0.554
	1000	0.033	0.385	0.029	0.283
Large graph 50	100	3.732	2.515	2.261	3.759
	1000	1.175	2.100	0.964	1.378

and apply our method with nongaussian assumption. some important genes causal effect are in the Figure 2. The morning gene CCA1 found to repress the evening genes EFL3 and NI. Among the evening genes, EFL4 and TOC1 have more effect on evening and morning genes . The evening gene ELF has positively effect on CCA1 and also it has negatively effect respect on LHY. Moreover, the evening genes ELF3, GI, TOC1 are involved in morning gene PRR and the morning gene LHY has a almost constant effect on the evening genes ELF4, TOC1, EFL4 . In particular ELF4 interacts positively effect with NI and CCA1 and negatively with LHY in Figure 2. Many of the results are consistent with the findings in Grzegorczyk and Husmeier (2011a,b).

5 Conclusion

In this paper, we have derived an explicit formula for describing a causal effect for a flexible class of distributions, the nonparanormal. These distributions are especially useful for observational studies, where normality assumptions are often not warranted. We also present a simple method, NCE, to estimate these causal effects. This is effectively a first order approximation of the general causal effect formula, but it can capture a large range of non-linear shapes. In a simulation study, we have shown that the estimation method works well, particularly away from the tails of the data. We also applied the method to an *Arabidopsis Thaliana* circadian clock network. The estimated causal effects all reveal a tendency for the causal effects to shrink to zero for large values of the cause.

Acknowledgments: The first author is deeply grateful for the travel grant from the Statistical Modelling Society for travelling to IWSM 2016 in Rennes, France.

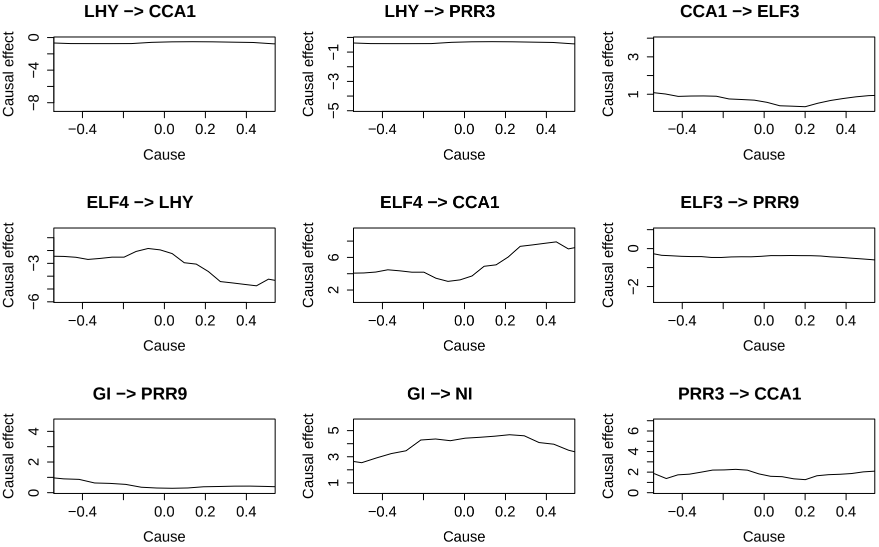


FIGURE 2. Causal effects for contemporaneous circadial gene interaction network in Arabidopsis thaliana. Some genes have functional causal effect and some other have almost linear causal effect.

References

Grzegorczyk, M., Husmeier, D. (2011a). Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes. *Bioinformatics* **27,** 693-699.

Grzegorczyk, M., Husmeier, D. (2011b). Non-homogeneous dynamic Bayesian networks for continuous data. *Machine Learning* **83,** 355-419.

Harris, N. and Drton, M. (2014). PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research* **27,** 3365-3383.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics* **40(4),** 2293-2326.

Maathuis, M. H., Kalisch, M., and Buhlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics* **37,** 3133-3164.

Modelling the shape of emotions

Irene Mariñas[1], Adrian Bowman[1], Vincent Macaulay[1]

[1] School of Mathematics and Statistics, University of Glasgow, UK

E-mail for correspondence: `i.marinas.1@research.gla.ac.uk`

Abstract: This paper addresses the problem of estimating a ridge curve embedded in a three dimensional surface that changes over time. The main challenge is to exploit the details of surface shape, while maintaining computational feasibility. A Gaussian Process approach is adopted to define a lip on a human face and model the change during the performance of an emotion.

Keywords: Gaussian Processes; Shape analysis; 3D curves; Lip morphology

1 Introduction

The statistical analysis of information on shape has been a research topic of considerable interest since the earliest part of the twentieth century, but it has developed substantially in the present century especially thanks to advances in computational tools. Interest in shape analysis of the human face began because of its applications in biology, medicine and psychology. This study is applied to the shape of the lips in a three-dimensional facial image and their variation over the expression of different emotions such as disgust, fear, anger, happiness, et cetera. To record the expressions, a large number of pictures of a person producing the emotion are taken with a stereophotogrammetric camera system, which leads to a set of data in four dimensions (the three spatial dimensions plus time).

2 Gaussian Process model for a 3D lip curve

A Gaussian Process (GP) is a flexible model which does not approximate the system by fitting the parameters of a finite set of basis functions, but rather by explicitly trying to capture the covariance structure of the data. It is a collection of random variables, any finite number of which have a joint Gaussian distribution (multivariate normal distribution) Rasmussen and Williams (2006). Suppose a GP r is defined as:

$$r(s,c) \sim GP\big(m(s,c), k(s,s',c,c')\big), \tag{1}$$

a mixed GP for the continuous index (spatial, i.e. the arc-length of the curve, rescaled to be from 0 to 1), $s \in [0,1]$, and the discrete label (i.e. coordinate), $c \in \{x,y,z\}$. This represents each coordinate as a function of the arc-length: $r(s,x) = x(s)$, $r(s,y) = y(s)$ and $r(s,z) = z(s)$. Let $\mathbf{s} = (s_1 \cdots s_n)^{\mathrm{T}}$ for a choice of n values of s. Also let $\mathbf{x} = (x(s_1) \cdots x(s_n))^{\mathrm{T}}$, $\mathbf{y} = (y(s_1) \cdots y(s_n))^{\mathrm{T}}$ and $\mathbf{z} = (z(s_1) \cdots z(s_n))^{\mathrm{T}}$. Then:

$$\mathbf{r} = \begin{bmatrix}\mathbf{x} & \mathbf{y} & \mathbf{z}\end{bmatrix}^{\mathrm{T}} \sim N_{3n}\left(\mathbf{m}, \mathbf{K}\right), \tag{2}$$

where $\mathbf{m}$ is the mean: $\mathbf{m} = m(\mathbf{s}, c) = (m(s_1,x) \cdots m(s_n,x) m(s_1,y) \cdots m(s_n,y) m(s_1,z) \cdots m(s_n,z))^{\mathrm{T}}$ and $\mathbf{K}$ is the covariance matrix. Separability is assumed: $k(s,s',c,c') = k_s(s,s') \cdot k_c(c,c')$. The space-covariance function used is the Squared-Exponential (SE), i.e. $k_s(s,s') = \sigma^2 \exp\big(-\frac{1}{2\lambda^2}$ $(s-s')^2\big)$, with hyperparameters: σ_f^2, the signal variance and λ, the length-scale. Then $\mathbf{K} = \mathbf{K}_c \otimes \mathbf{K}_s$, where $\mathbf{K}_s$ represents the covariance matrix for the n arc-length inputs, with $(i,j)^{th}$ element equal to $k_s(s_i,s_j)$. For the 3×3 matrix $\mathbf{K}_c$, two hyperparameters were specified: κ_1, the correlation between x and y or z, and κ_2, between y and z. The mean is assumed to be zero and therefore:

$$\mathbf{r} = \begin{bmatrix}\mathbf{x}\\ \mathbf{y}\\ \mathbf{z}\end{bmatrix} \sim N_{3n}\left(\mathbf{0}, \begin{bmatrix}\mathbf{K}_s & \kappa_1\mathbf{K}_s & \kappa_1\mathbf{K}_s\\ \kappa_1\mathbf{K}_s & \mathbf{K}_s & \kappa_2\mathbf{K}_s\\ \kappa_1\mathbf{K}_s & \kappa_2\mathbf{K}_s & \mathbf{K}_s\end{bmatrix}\right), \tag{3}$$

2.1 Likelihood for a lip curve

The distribution in (3) can be factorised as:

$$\begin{aligned}
\mathbf{x} &\sim N_n(\mathbf{0}, \mathbf{K}_s),\\
\mathbf{y} \mid \mathbf{x} &\sim N_n(\kappa_1\mathbf{x}, (1-\kappa_1^2)\mathbf{K}_s),\\
\mathbf{z} \mid \mathbf{x}, \mathbf{y} &\sim N_n\left(\left[\{\kappa_1 - \kappa_1\kappa_2\}\mathbf{x} + \left\{\kappa_2 - \kappa_1^2\right\}\mathbf{y}\right] / \left[1-\kappa_1^2\right],\right.\\
&\qquad \left.\left[1 - \left\{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2\right\} / \left\{1-\kappa_1^2\right\}\right]\mathbf{K}_s\right),
\end{aligned} \tag{4}$$

so that the total log-likelihood of the process is: $l(\theta) = \log p(\mathbf{x}) + \log p(\mathbf{y} \mid \mathbf{x}) + \log p(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$, where $\theta = (\sigma_f, \lambda, \kappa_1, \kappa_2)$ is the the set of hyperparameters. From (4):

$$
\begin{aligned}
\log p(\mathbf{x}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x}. \\
\log p(\mathbf{y} \mid \mathbf{x}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log(1-\kappa_1^2) - \\
&\quad \frac{1}{2(1-\kappa_1^2)}(y-\kappa_1\mathbf{x})^{\mathrm{T}}\mathbf{K}_s^{-1}(\mathbf{y}-\kappa_1\mathbf{x}). \qquad (5) \\
\log p(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log\left(1 - \frac{\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right) \\
&\quad -\frac{(\mathbf{z}-\bar{\mathbf{z}})^{\mathrm{T}}\mathbf{K}_s^{-1}(\mathbf{z}-\bar{\mathbf{z}})}{2\left(1-[\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2]\,/\,[1-\kappa_1^2]\right)},
\end{aligned}
$$

where $\bar{\mathbf{z}}$ denotes the mean of $\mathbf{z} \mid \mathbf{x}, \mathbf{y}$: $\bar{\mathbf{z}} = [(\kappa_1 - \kappa_1\kappa_2)\mathbf{x} + (\kappa_2 - \kappa_1^2)\mathbf{y}]/[1-\kappa_1^2]$.

2.2 Prediction for a lip curve

To make predictions for values of the coordinates at a set of test points $\mathbf{s}^* = (s_1^*, \ldots, s_{n^*}^*)^{\mathrm{T}}$, from the training points $\mathbf{s}$, the distributions of each predicted coordinate can be calculated with the same dependences assumed in Section 2.1. The conditional predictive distributions are:

$$
\begin{aligned}
\mathbf{x}^* \mid \mathbf{x} &\sim N\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}, \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right). \\
\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{x}, \mathbf{y} &\sim N\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y} + \kappa_1[\mathbf{x}^* - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}],\right. \\
&\qquad \left.[1-\kappa_1^2][\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}]\right). \\
\mathbf{z}^* \mid \mathbf{x}^*, \mathbf{x}, \mathbf{y}^*, \mathbf{y}, \mathbf{z} &\sim N\Bigg(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z} + \qquad (6) \\
&\qquad \frac{1}{1-\kappa_1^2}\left[\{\kappa_1 - \kappa_1\kappa_2\}\left\{\mathbf{x}^* - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}\right\} + \right. \\
&\qquad \left.\{\kappa_2 - \kappa_1^2\}\left\{\mathbf{y}^* - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}\right\}\right], \\
&\qquad \left[1 - \frac{\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right]\left[\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\Bigg).
\end{aligned}
$$

$\mathbf{K}_{ss^*}$ denotes the $n \times n^*$ matrix of the covariances evaluated at all pairs of training and test points and $\mathbf{K}_{s^*s}$ is its transpose. $\mathbf{K}_{s^*}$ contains the covariances for the test points.

2.3 Fitting the lip curve model

To study the model, the upper lip of one resting face was estimated. Each coordinate curve has its mean subtracted so that they are centred around zero. The lip curve contains 24 highly correlated points (due to their smoothness). This causes the covariance matrix $\mathbf{K}_s$ (with noise-free data assumed) to be ill-conditioned, tending to make numerical calculation of its inverse unstable. The approach opted for was to add some noise to the model of the observations. This accommodates errors in the observed facial surface and causes the ratio between the largest and the smallest eigenvalue to decrease. Since the lip curves are measured in mm, it was decided that an error of 0.1 mm had little effect on the lip representation while making optimization viable. Optimal values for the hyperparameters were found by maximum likelihood. Figure 1 shows the original data points and 25 predicted points for each coordinate. The optimal hyperparameters $\theta = (\sigma_f, \lambda, \kappa_1, \kappa_2)$ found were: $\theta = (9.4418, 0.1324, 0.0634, 0.4970)$, with respective SE: $1.4059, 0.0123, 0.0978, 0.2179$. The posterior means are displayed with 2 standard deviations bands (shown dotted), which are too narrow to be properly appreciated.

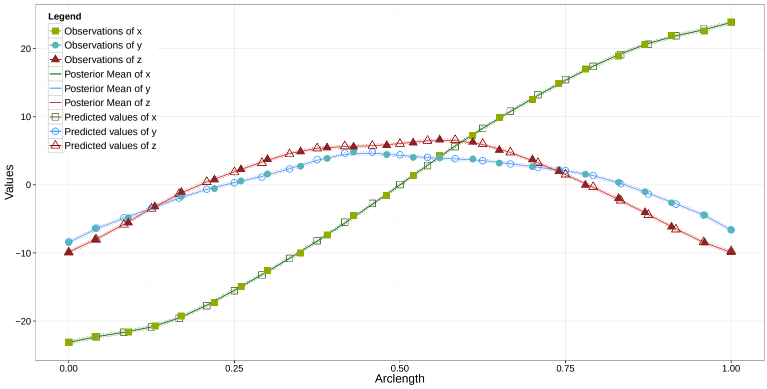


FIGURE 1. Observations, posterior means and predicted values for an upper lip.

3 Gaussian Process model for the evolution of one coordinate

Consider the case where the lip shape varies over the performance of an emotion (Figure 2). A GP can be specified for each coordinate evolving through time. The observed values of y, say, depend on two continuous variables: the space component s and the time component t. The GP can be then defined as

$$y(s,t) \sim GP\big(m(s,t), k(s,s',t,t')\big). \tag{7}$$

Let $\mathbf{s} = (s_1 \cdots s_n)^{\mathrm{T}}$, as in Section 2, $\mathbf{t} = (t_1 \cdots t_T)^{\mathrm{T}}$, for a choice of T values of t, and $\underline{y} = (\mathbf{y}(t_1) \cdots \mathbf{y}(t_T))^{\mathrm{T}}$, where $\mathbf{y}(t_i) = (y(s_1,t_i) \cdots y(s_n,t_i))^{\mathrm{T}}$ represents the points on the curve at time t_i.

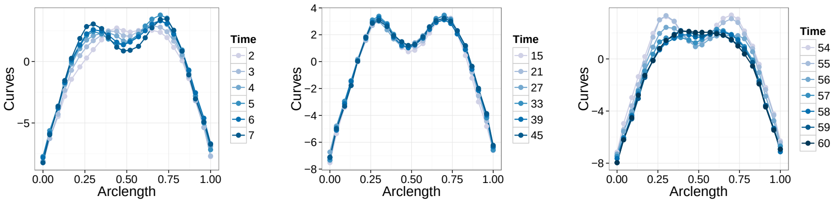


FIGURE 2. Samples of the y coordinate evolving during the performance of disgust.

Separability is assumed: $k(s,s',t,t') = k_s(s,s') \cdot k_t(t,t')$. The space-covariance function is the SE. The process is assumed Markovian and hence the Ornstein-Uhlenbeck (OU) stationary covariance function is used for the time-covariance, i.e. $k_t(t,t') = \exp(- \mid t-t' \mid /\mu)$, with hyperparameter μ, the time scale. The mean is assumed to be zero and therefore the distribution of one curve and its predecessor can be written as:

$$\begin{bmatrix} \mathbf{y}(t) \\ \mathbf{y}(t-1) \end{bmatrix} \sim N_{2n}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_s & \kappa\mathbf{K}_s \\ \kappa\mathbf{K}_s & \mathbf{K}_s \end{bmatrix}\right), \tag{8}$$

where $\kappa = \exp(-1/\mu)$, when the time difference between curves is assumed to be 1.

3.1 Likelihood for the evolution

By the Markov property, given the parameters of the model: $p(\underline{y}) = p(\mathbf{y}(1)) \cdot \prod_{i=2}^{T} p(\mathbf{y}(i) \mid \mathbf{y}(i-1))$. Then the total log-likelihood of the process is: $l(\theta) = \log p\big(\mathbf{y}(1) \mid \theta\big) + \sum_{i=2}^{T} \log p\big(\mathbf{y}(i) \mid \mathbf{y}(i-1), \theta\big)$, where $\theta = (\sigma_f, \lambda, \mu)$, using:

$$\begin{aligned} \mathbf{y}(1) &\sim N_n(\mathbf{0}, \mathbf{K}_s), \\ \mathbf{y}(t) \mid \mathbf{y}(t-1) &\sim N_n(\kappa\mathbf{y}(t-1), (1-\kappa^2)\mathbf{K}_s). \end{aligned} \tag{9}$$

3.2 Prediction for the evolution

Along the sequence of curves that capture the emotion, marginal predictions at time $q \in \mathbf{t}$ can be done at a set of test points $\mathbf{s}^*$. Similarly, predictions and retrodictions for previous or following time points can be estimated, using:

$$\begin{aligned} &\mathbf{y}^*(q) \mid \underline{y} \sim N_n\left(\kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(1), \kappa^{2(1-q)}\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right), && q < 1, \\ &\mathbf{y}^*(q) \mid \underline{y} \sim N_n\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(q), \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right), && q \in \mathbf{t}, \\ &\mathbf{y}^*(q) \mid \underline{y} \sim N_n\left(\kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(T), \kappa^{2(q-T)}\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right), && q > T. \end{aligned} \tag{10}$$

3.3 Fitting the evolution model

A sequence of uppers lip for the emotion *disgust* (61 pictures: $\mathbf{t} = (1 \cdots 61)^{\mathrm{T}}$) was estimated. Each curve had its mean subtracted. Optimal hyperparameters, $\theta = (\sigma_f, \lambda, \mu)$, were found by maximum likelihood: $\theta = (1.7426, 0.0677, 64.4236)$, with respective SE: $0.0751, 0.0009, 5.4307$. Figure 3 shows the original data points and 25 predicted points for time points: -1 (retrodiction), 2 (marginal prediction) and 63 (prediction). Observations shown for time points -1 and 63 are, respectively, from the first and last observed curves. The posterior means are displayed with 2 standard deviations bands (shown dotted).

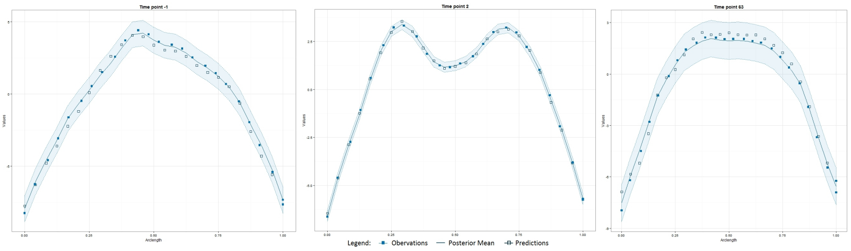


FIGURE 3. Observations, posterior means and predicted values.

4 Conclusions and further lines of investigation

The use of shape information, expressed in a continuous and multivariate scale raises a number of very interesting issues from a methodological perspective. Both models to express the three coordinates as single curves in terms of the arc-length and to express how one coordinate changes over time interpolate the data well. The notion of a shape evolving in time will be extended to a phylogenetic setting, where branching points in the evolution can occur. The aim is to develop statistical methods by which shape information on organisms can be used to reconstruct a phylogenetic tree. This raises a number of interesting questions on ways to process both genetic (discrete) and shape (continuous) information.

Acknowledgments: IM is grateful to the School of Mathematics and Statistics at the University of Glasgow for her research studentship and their support. Special thanks to Oliver Garrod and colleagues from the School of Psychology for kindly contributing their data.

References

Rasmussen, C.E. & Williams, C.K.I (2006). *Gaussian processes for Machine Learning.* Cambridge, Massachusetts: The MIT Press.

Information Retrieval Models: Performance, Evaluation and Comparisons for healthcare Big Data Analytics

Kenan Matawie[1] and Sargon Hasso [2]

[1] Western Sydney University, Australia
[2] Independent Software Architect Consultant, USA

E-mail for correspondence: k.matawie@westernsydney.edu.au

Abstract: We propose analysis, performance and evaluation of different Information retrieval models with a foundational implementation system to the Healthcare Data Anaytics. In this type of systems, patients post questions to patient/caregiver support forums. To reduce repetitiveness due to previously asked questions by other patients with similar conditions, albeit worded differently, the proposed system will offer patients questions that are semantically similar to theirs. The problem is re-formulated as an Information Retrieval (IR) problem and several of the modern implementations of IR models particularly the probabilistic models are available to tackle this problem. Specifically, we utilized Lucene which offers a full-text search library by adding search functionality to our foundational model and system implementation.

Keywords: Statistical Language model; Text Retrieval; Healthcare Informatics.

1 Intruduction

The IEEE International Conference on Healtcare Informatics posed the following Healthcare Data Analytics Challenge ICHI (2015), and the data provided was a subset of a large data sets: in a patient/caregiver support forums, patients submit questions regarding their conditions. Overtime, as these forums grow, so does the repetitive nature of questions asked by different patients who may have similar conditions but ask questions worded differently. Specifically, the challenge specification calls for the following. Given a corpus of question, $Q = q_1, q_2, \ldots, q_n$ from a patient support forum, where each of q_i's representing a question from a patient forum on Type II Diabetes, design and implement a system that for each incoming query,

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).

iq_j, identify a maximum of three most similar questions from the corpus Q. Similarity signifies, in this context, questions that are worded differently but they have the same meaning. Various statistical language models are used with the focus on the effectiveness rather than the efficiency of the information retrieval, which measure the ability to find relevant information accurately. The efficiency, the time taken to return the information, which is the second important principal of the performance requirement for any IR model will not be covered in this paper.

2 Statistical Language Models

The main concern here is what documents (d) satisfy user's information needed (query q) with enough accuracy. Prior to presenting any model, it is important to determine the contribution of the term to the document, which is calculated by using language models based on a given d. Most of the models are the maximum likelihood estimate of the relative counts using the following ranking models: Vector Space Model (VSM) Manning et al. (2008),

$$f(q,d) = \sum_{w \in q \bigcap d} c(w,q)c(w,d)log\frac{M+1}{df(w)}$$

where $c(w,q)$ and $c(w,d)$ denote number of words in a query, and count of words in a document, respectively; M is the total number of documents in the collection; $df(w)$ denote document frequency. Best Match family (BM25) by Jones et al. (2000),

$$f(q,d) = \sum_{w \in q \bigcap d} c(w,q)\frac{(k+1)(c(w,d)}{c(w,d)+k(1-b+b\frac{|d|}{avdl})}log\frac{M+1}{df(w)}$$

where $b \in [0,1]$ is part of the *normalizer* term $1-b+b\frac{|d|}{avdl}$; $avdl$ denotes average document length.

Language models assume that d is used to generate q, and this can also be ranked by the following Bayesian probability function; $P(d|q) \propto P(q|d)P(d)$ Each document is a list of keywords or terms (t) and can be expressed as a product of the probability of the terms in the query generated by the documents: Jelinek-Mercer (JM) Jelinek and Mercer (1980). This model considers the linear structure of the maximum likelihood of the collection model using λ to control the influence of each component,

$$f_{JM}(q,d) = \sum_{w \in q \bigcap d} c(w,q)log[1+\frac{1-\lambda}{\lambda}\frac{c(w,d)}{|d|p(w|C)}]$$

where $\lambda \in [0,1]$; the probability of unseen word in the collection is proportional to the smoothing term $p(w|C)$. Latent Dirichlet Allocation (LDA) Blei (2003),

$$f_{DIR}(q,d) = [\sum_{w \in q \bigcap d} c(w,q) log[1 + \frac{c(w,d)}{\mu p(w|C)}]] + n log \frac{\mu}{\mu + |d|}$$

where $\mu \in [0, \infty)$, n is a constant.

3 An Outline of Our System Implementation Approach

There are several approaches to design such system; however, in this specific instance we have chosen to design and implement the system as an information retrieval (IR) application. An IR application allows users to submit *ad hoc queries* in an attempt to communicate the information need Manning et al. (2008), such as a medical condition related to, say, Type II Diabetes or any other medical condition for that matter. As Manning et al. (2008) state, these types of IR systems commonly have three features we are interested in:

1. They process large documents collection quickly.
2. They allow for more flexible and sophisticated matching operations.
3. They allow to return the best answer to an information need.

With this in mind, we have formulated the Healthcare Data Analytics challenge as an IR problem and have opted to use Lucene, Apache Lucene(2011) release 5.2.1. Lucene is scalable, powerful, and most importantly, open-source Java-based search library that we built into a prototype software system to address this challenge. Our design and implementation is a fully functional prototype that meets all the requirements. IR-based systems allow users to search for documents, information within a document, or the metadata about the documents McCandless (2010). The core implementation, at a minimum, requires an indexing component and a searching component. Documents have to be indexed first and then users are provided with another component that allows them to interact with the system to retrieve information they need. Any type of content can be treated as a document. Following this argument, each question in the Corpus is treated as a document with this logical structure: $< q_{id}, q_{topic-category}, q_{content} >$. From this point forward, we will use the *Question* and *Document* interchangeably to mean the same thing in this paper. In our case, $q_{content}$ holds the actual information users would need. The other elements, i.e. q_{id} and $q_{topic-category}$ are used internally as metadata to uniquely identify each question and optionally store its category. The design elements of each component follow, more or less, the general requirements of any IR-based application and we follow strictly the guidelines discussed by McCandless et al. (2010).

3.1 Indexing Component

This component builds an index for quick access to known fields within a document (question) such as q_{id} and $q_{topic-category}$. It is a process that has the steps listed below. With the exception of document acquisition step, Lucene provides all the other functionalities.

- Document Acquisition: A sample questions corpus was provided for testing. It is one physical document that has many questions in it. For this step, a parser was implemented to break down and separate each question into its individual logical structure for the next step.
- Document Building: Lucene concept of a *Document* which is the smallest logical and indexable unit that acts as a container with fields. This is a transformation step from raw content to analyzable logical content (next step).
- Document Analysis: In this step, all raw text is analyzed and broken down into tokens. A standard Lucene analyzer breaks down text into individual words on white space boundaries including punctuation marks, spaces, tabs, newlines, etc. We have also implemented a custom analyzer that injects synonym of words into the outgoing token stream during indexing or querying. Lucene provides an interface to build such a facility for query expansion and we have built it using WordNet database WordNet(2010). This query expansion facility using a synonym database allows searching for words that have not been entered by users. For example, it will treat 'Type 2', 'Type II', 'Type two' the same.
- Document Indexing: In this step each document is added to the index.

3.2 Searching Component

This component provides the necessary steps to allow users submit questions/queries and render results that match closely to their input. Again we will use input question and query interchangeably. It is a process that has these steps.

1. User Interface: We provide two ways for users to enter their questions/queries. From a file, or from a command line. If entered from a file, multiple queries can be processed at once.
2. Query Building: In its simplest form, users enter questions in a form-free format.
3. Search Query: Here the application searches the index and retrieves the documents that matches the questions the best. Lucene implements several theoretical scoring models to select from such as Vector Space Model (VSM), Okapi BM25, Language Models such as Latent

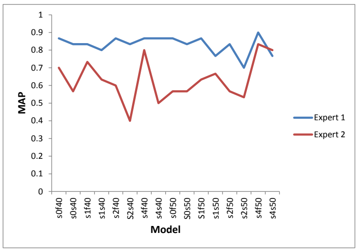


FIGURE 1. *MAP* values for the 16 output cases as judged by two human experts.

Dirichlet Allocation, and Jelinek-Mercer model. Lucene implementors developed the language models smoothing methods described in Zhai and Lafferty (2001). In our application, you can select any model as a scoring function to rank the best matching documents through configuration.

4. Render Results: All top-ranked questions are returned. By default, we only return at most 3 documents; however, this value can be changed through configuration.

4 Methodology and Analysis of the Results

We built the system for maximum runtime configuration and best performance. As we have indicated above, we can choose from four different ranking functions corresponding to VSM, BM25, LDA, and JM language models. We also have implemented synonymic query expansion during indexing and searching. We noticed, without examining the reason at this time, that running a query against an index that was generated with synonyms gave us a different outcome from an index that was generated without synonyms. This resulted in generating 16 different outcome data sets. We used *Mean Average Precision* (MAP) Manning et al. (2008), a single metric, as an evaluation criterion to measure the quality across recall levels among all algorithms, i.e. relevant and non-relevant as judged by two human experts. With the provided sample questions corpus and sample questions/queries, the best performer was the Jelinek-Mercer model with no synonymic query expansion but with synonymic index generation as shown in Fig 1.

5 Conclusion

Healthcare informatics data from the 2015 IEEE International Conference ICHI (2015) was used, this is a sample of large and complex data sets. We have implemented a software system based on IR models. We utilized Lucene search library as the underlying technology to realize this

IR application. The delivered application has an indexing component, and a searching component. The search component implements four ranking models to choose from to rank the retrieved documents. We also implemented synonym query expansion during indexing and searching. Sixteen different outcome data sets were obtained as a result of these model/query expansion behaviors. A best model was selected during our evaluation of the sixteen outcome results using Mean Average Precision (MAP) criterion with the help of human subject matter experts.
This is an active ongoing research, more statistical analysis and graphs will be presented on the performance, evaluation, preference and comparisons of these models in the final version of this paper.

Acknowledgments: The authors would like to thank Dr. Kamiel Maloul and Dr. Jan Slewa for their participation in reviewing the results returned from our system implementation as human subject matter experts and quantifying the relevance.

References

Apache Lucene. (2015) The Apache Software Foundation. http://lucene.apache.

Blei, David M. NG, Andrew Y., and Jordan, Michael I. (2003) Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, **3**, 993 − 1022.

Jelinek, F. and Mercer, R. (1980) Interpolated Estimation of Markov Source Parameters from Sparse Data. In *Proceedings of the Workshop on Pattern Recognition in Practice.*

Jones, S. K., Walker, S. and Robertson, S. E. (2000) A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. In *Information Processing and Management*, pages 779–840.

ICHI (2015), Healthcare Data Analytics Challenge. http:// cs.utdallas. edu/ ichi2015/contributors/ healthcare-data-analytics-challenge/org/.

Manning, C.D., Raghavan, P. and Schutze, H. (2008) *Introduction to Information Retrieval.* Cambridge University Press.

McCandless, M., Hatcher, E. and Gospodneti, O.C. (2010) *Lucene in Action. Manning Publications Co.*, Stamford, CT, 2nd edition.

Princeton University (2010) About WordNet. http://wordnet.princeton.edu/2010. Accessed July 2015.

Zhai, C. and Lafferty, J. (2001) A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of the Conference on Research and Development in Information Retrieval, SIGIR 2001*, NY, USA 334–342.

Gradient test for generalised linear models with random effects

Antonio Hermes Marques da Silva Junior[12], Jochen Einbeck[1], Peter S. Craig[1]

[1] Durham University, Durham, UK
[2] Universidade Federal do Rio Grande do Norte, Natal, Brazil

E-mail for correspondence: a.hermes.marques-da-silva-jun@durham.ac.uk

Abstract: This work develops the gradient test for parameter selection in generalised linear models with random effects. Asymptotically, the test statistic has a χ^2 distribution and the statistic has a compelling feature: it does not require computation of the Fisher information matrix. Performance of the test is verified through Monte Carlo simulations of size and power, and also compared to the likelihood ratio, Wald and Rao tests. The gradient test provides the best results overall when compared to the traditional tests, especially for smaller sample sizes.

Keywords: Generalised linear models; random effects; asymptotic test.

1 Generalised linear models with random effects

Consider a generalised linear model with random effects (GLMwRE) for a data set containing n independent observations of a response variable, denoted $\mathbf{y} = (y_1, \ldots, y_n)^\top$, which by definition has a distribution in the exponential family, and corresponding observations on p explanatory variables, denoted $\mathbf{x}_i^\top = (x_{i1}, \ldots, x_{ip})^\top$ for $i = 1, \ldots, n$. The linear predictor for the i-th observation is $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + z_i$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the vector of regression parameters and z_i is an unobserved random effect. The relationship between y_i and η_i is given by the conditional mean $\mu_i = \mathrm{E}[y_i|z_i]$ and the monotonic and differentiable *link function*, $g(\,\cdot\,)$ such that $\mu_i = g^{-1}(\eta_i)$. The z_i can be considered as sampled from $\mathcal{N}(0, \sigma^2)$, where $\sigma > 0$. An alternative nonparametric approach is to leave the distribution of z_i unspecified. In either case, the distribution of z_i may be approximated by a discrete distribution with finite support. Then the likelihood function $L^*(\boldsymbol{\beta})$ for the GLMwRE and its approximation $L(\boldsymbol{\beta})$ can be written as (Aitkin et al., 2009)

$$\begin{aligned} L^*(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \int f(y_i|\boldsymbol{\beta}, \phi, z_i)\varpi(z_i)dz_i \\ &\approx \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k f(y_i|\boldsymbol{\beta}, \phi, \tilde{z}_k) = \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k f_{ik} = L(\boldsymbol{\beta}), \end{aligned} \tag{1}$$

where $f(\,\cdot\,)$ is the response density, ϕ is the dispersion parameter, $\varpi(\,\cdot\,)$ is the density of the random effect z_i, $\tilde{z}_k$ are mass points and π_k are mass probabilities. From (1) we have an approximate linear predictor for the k-th component of the i-th observation as $g(\mu_{ik}) = \eta_{ik} = \mathbf{x}_i^\top \boldsymbol{\beta} + \tilde{z}_k$ where $\mu_{ik} = \mathrm{E}[y_i|z_i = \tilde{z}_k]$. Let $\ddot{\mathbf{y}}^\top = (\mathbf{y}^\top, \mathbf{y}^\top, \ldots, \mathbf{y}^\top)$ be a vector of nK-dimension of pseudo-observations and the corresponding stacked linear predictor be

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \dddot{\mathbf{X}}\boldsymbol{\beta} + \ddot{\mathbf{z}} \tag{2}$$

where $\boldsymbol{\mu}^\top = (\mu_{11}, \ldots, \mu_{n1}, \ldots, \mu_{1K}, \ldots, \mu_{nK})$, $\boldsymbol{\eta}^\top = (\eta_{11}, \ldots, \eta_{n1}, \ldots, \eta_{1K}, \ldots, \eta_{nK})$, $\ddot{\mathbf{z}}^\top = (\tilde{z}_1, \ldots, \tilde{z}_1, \ldots, \tilde{z}_K, \ldots, \tilde{z}_K)$ is the n times stacked mass point vector, and $\dddot{\mathbf{X}}^\top = (\mathbf{X}^\top, \ldots, \mathbf{X}^\top)$ is the $nK \times p$ pseudo model matrix, where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$. Maximum Likelihood Estimation (MLE) typically proceeds via the EM algorithm. In the non-parametric approach, π_k and z_k are estimated adaptively along with $\boldsymbol{\beta}$ in the M step and this is known as non-parametric maximum likelihood (NPML). Tabulated Gaussian quadrature points are used for π_k and z_k in the case of Gaussian random effects (the latter being scaled by a parameter σ which needs estimation).

2 The gradient test

The problem considered is that of testing a composite hypothesis $\mathcal{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^{(0)}$ against a composite alternative $\mathcal{H}_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^{(0)}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_q)^\top$ is a $q-$dimensional parameter of interest with $q \leqslant p$, $\boldsymbol{\beta}_2 = (\beta_{q+1}, \ldots, \beta_p)^\top$ is a $(p-q)-$dimensional nuisance parameter and $\boldsymbol{\beta}_1^{(0)}$ is a specified vector. This induces the partitioning $\dddot{\mathbf{X}} = (\dddot{\mathbf{X}}_1, \dddot{\mathbf{X}}_2)$. Let

$$\mathbf{U}(\boldsymbol{\beta}) = \partial \log L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \{\mathbf{U}_1^\top(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2), \mathbf{U}_2^\top(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)\}^\top = \{\mathbf{U}_1^\top, \mathbf{U}_2^\top\}^\top$$

be the corresponding partition of the total score function for $\boldsymbol{\beta}$. The unrestricted MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)^\top$ and the restricted MLE of $\boldsymbol{\beta}_2$ under $\mathcal{H}_0$ is written $\tilde{\boldsymbol{\beta}}_2$. Functions evaluated at the point $\tilde{\boldsymbol{\beta}}^\top = (\boldsymbol{\beta}_1^{(0)\top}, \tilde{\boldsymbol{\beta}}_2^\top)$ will be distinguished by the addition of a tilde. The gradient statistic $\xi_{\mathcal{T}}$ for testing $\mathcal{H}_0$ versus $\mathcal{H}_1$ has the simple form $\xi_{\mathcal{T}} = \tilde{\mathbf{U}}_1^\top(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)})$ (Terrell, 2002). In the context and notation set out earlier, one has $\tilde{\mathbf{U}}_1 = \dddot{\mathbf{X}}_1^\top \tilde{\mathbf{D}}(\ddot{\mathbf{y}} - \tilde{\boldsymbol{\mu}})$ and $\mathbf{D}$ is the diagonal matrix with diagonal entries $d_{11}, \ldots, d_{n1}, \ldots, d_{1K}, \ldots, d_{nK}$

given by $d_{ik} = (\phi\omega_{ik}/V_{ik})(d\mu_{ik}/d\eta_{ik})$ where $\omega_{ik} = \pi_k f_{ik}/\sum_{l=1}^{K}\pi_l f_{il}$ and V_{ik} is the variance function applied to μ_{ik}. Therefore, the gradient statistic formula for testing $\mathcal{H}_0$ is

$$\xi_{\mathcal{T}} = (\dddot{\mathbf{y}} - \tilde{\boldsymbol{\mu}})^{\top}\tilde{\mathbf{D}}\dddot{\mathbf{X}}_1(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^{(0)}). \tag{3}$$

Based on Terrell's (2002) results, the distribution of $\xi_{\mathcal{T}}$ tends under $\mathcal{H}_0$ to the $\chi^2(q)$ distribution as n increases. Theoretically, the $\xi_{\mathcal{T}}$, likelihood-ratio (LR) $\xi_{\mathcal{LR}}$, Wald $\xi_{\mathcal{W}}$ and Rao $\xi_{\mathcal{R}}$ statistics are asymptotically equivalent since they all have the same asymptotic distribution under $\mathcal{H}_0$. Nonetheless, for finite samples the size and/or power of the tests may differ. Consequently, we provide numerical simulation results to compare their performance.

3 Simulation experiment

We report results of Monte Carlo simulations assessing properties of $\xi_{\mathcal{T}}$ in finite samples. For this, we establish a model with linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + z_i, \text{ for } i = 1, \ldots, n$$

where x_1, x_3 and x_4 are samples of size n from $\mathcal{U}(0,1)$, $\mathcal{F}(2,5)$ and $t(3)$, respectively. The parameter values are β_0=1, β_1=−1, $\beta_{2i} = (i \mod 3) - 1$ and $\phi = 1$. The random effects z_i are samples from $\mathcal{N}(0, 8^{-2})$ for the Gaussian quadrature fitting and from a discrete distribution which takes K values from $\mathcal{N}(0, 8^{-2})$ and probabilities from $\mathcal{U}(0,1)$ for the NPML fitting. The simulation results are based on Normal with identity link and Poisson and Gamma models with log link function. We took samples of 50, 100, 200 and 400 observations and the number of replications was 10,000 and $K = 3$. Our aim is to test $\mathcal{H}_0 : (\beta_3, \beta_4)^{\top} = (0,0)^{\top}$ versus $\mathcal{H}_1 : (\beta_3, \beta_4)^{\top} \neq (0,0)^{\top}$. Table 1 shows the null rejection rates of each test for two response distributions. Overall, the gradient statistic has rejection rates closer to the nominal levels. We set $n = 400$, $K = 3$ and $\alpha = 5\%$ for the power simulations where we computed the rejection rates under the alternative hypothesis $\beta_3 = \beta_4 = \delta$, for $-4 \leq \delta \leq 4$. Figure 1 shows that the power curves for $\xi_{\mathcal{LR}}$ and $\xi_{\mathcal{T}}$ are practically identical and that $\xi_{\mathcal{W}}$ and $\xi_{\mathcal{R}}$ have rather unusual curves, especially for the NPML model.

4 Concluding remarks

The gradient test shows itself as a useful inferential tool in the context of GLMwRE for several reasons. Firstly, its statistic requires neither the Fisher information matrix nor its inverse, which is an important simplification compared to the Wald and Rao statistics. Secondly according to our

TABLE 1. Null rejection rates (%).

			Gaussian quadrature				NPML			
	n	α	$\xi_{\mathcal{LR}}$	$\xi_{\mathcal{W}}$	$\xi_{\mathcal{R}}$	$\xi_{\mathcal{T}}$	$\xi_{\mathcal{LR}}$	$\xi_{\mathcal{W}}$	$\xi_{\mathcal{R}}$	$\xi_{\mathcal{T}}$
Normal	50	10	13.36	16.10	10.67	11.94	45.97	80.56	3.66	25.00
		5	7.12	9.52	5.24	6.06	33.62	76.57	1.91	16.19
		1	1.78	2.97	0.93	1.12	15.46	68.85	0.48	5.34
	100	10	11.73	12.99	10.51	11.16	24.90	60.19	4.64	17.08
		5	6.08	7.11	5.22	5.59	15.78	53.86	2.58	9.83
		1	1.25	1.72	0.96	1.08	5.18	43.30	0.74	2.50
	200	10	11.45	12.24	10.62	11.16	15.60	38.70	6.53	13.23
		5	5.88	6.49	5.24	5.55	8.58	30.89	3.75	7.09
		1	1.21	1.48	1.02	1.08	2.47	19.79	1.23	1.72
	400	10	10.47	10.95	9.98	10.32	12.78	23.99	9.35	11.96
		5	5.36	5.86	5.04	5.24	6.66	17.20	5.33	6.19
		1	1.15	1.29	0.99	1.09	1.53	8.31	1.70	1.25
Poisson	50	10	10.11	11.92	7.90	10.48	9.36	4.89	16.50	8.91
		5	5.01	6.70	3.86	5.46	4.50	2.24	9.64	4.04
		1	1.13	1.74	0.83	1.40	0.73	0.42	2.82	0.60
	100	10	10.32	12.15	8.56	10.50	9.98	5.31	16.78	9.57
		5	5.20	6.51	4.18	5.43	4.97	2.49	10.06	4.75
		1	1.15	1.65	0.78	1.34	0.88	0.46	3.28	0.92
	200	10	10.45	11.77	8.53	10.72	10.06	5.59	17.77	9.88
		5	4.98	6.20	4.17	5.22	5.05	2.65	10.80	4.88
		1	0.95	1.47	0.74	1.12	0.93	0.52	3.25	1.01
	400	10	9.68	11.15	8.25	9.82	9.50	5.06	16.52	9.68
		5	4.86	5.93	4.23	4.97	4.64	2.13	9.89	4.64
		1	0.97	1.41	0.77	1.04	0.87	0.46	2.83	0.93
Gamma	50	10	13.81	24.13	12.27	15.77	37.47	68.51	7.06	27.24
		5	7.73	16.51	6.79	8.29	27.98	62.20	3.53	17.65
		1	2.07	7.37	1.77	1.89	13.62	51.27	0.93	6.31
	100	10	11.98	18.52	11.62	13.08	22.97	52.04	5.31	18.65
		5	6.39	11.66	6.45	6.49	15.10	43.99	2.89	10.68
		1	1.54	4.49	2.05	1.39	6.06	31.10	0.90	2.84
	200	10	10.78	15.65	10.23	11.24	16.48	38.57	5.14	13.79
		5	5.33	9.33	5.37	5.49	10.45	29.59	2.87	7.59
		1	1.18	2.82	1.64	1.12	3.60	16.98	0.78	1.58
	400	10	10.38	13.21	9.80	10.49	14.10	28.43	5.72	12.22
		5	5.17	7.85	5.17	5.20	8.04	20.40	3.02	6.00
		1	1.08	1.99	1.43	0.91	2.63	10.04	0.77	1.28

simulations, the null rejection rates of the gradient test are much closer to the true nominal levels than the other three tests for the normal response

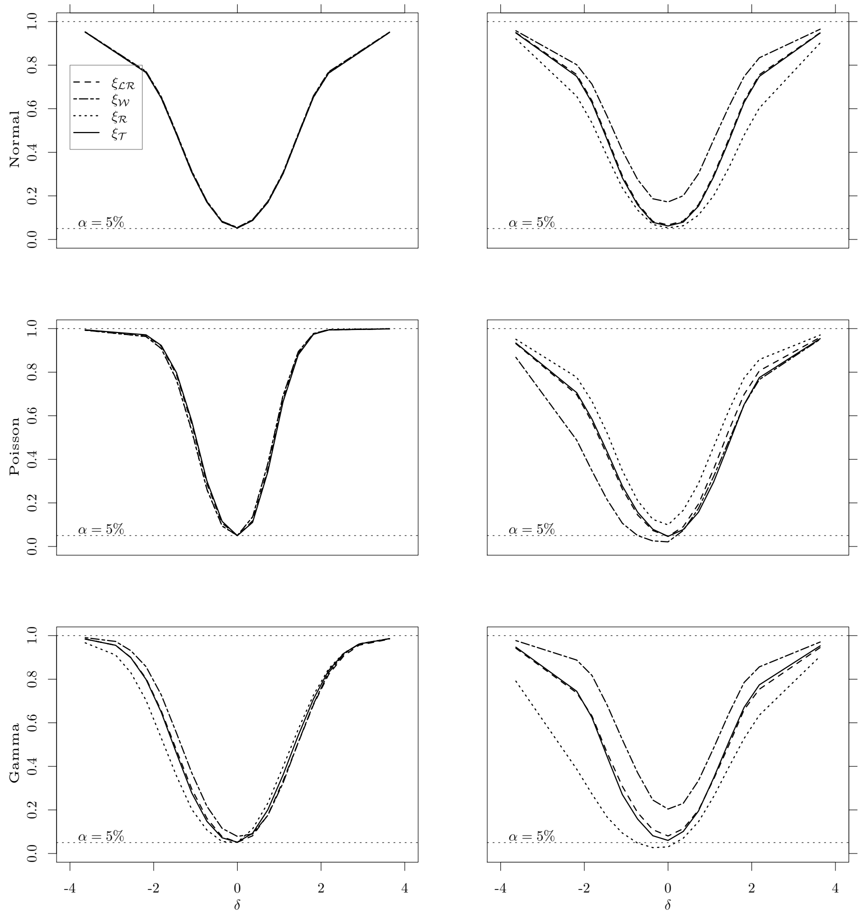


FIGURE 1. Power of the four tests: $n = 400$, $k = 3$, $\alpha = 5\%$. Left, for Gaussian quadrature fitting and right, for NPML fitting.

model and both gradient and LR tests have good rates for the Poisson response. Finally, our power simulations suggest that the gradient and LR tests have similar power properties. In sum, this indicates that the gradient tests should be preferred in the context of GLMwRE.

Acknowledgments: We gratefully acknowledge grant n° 9622/13-6 from CAPES foundation through Brazil's Science Without Borders Program.

References

Aitkin, M.A., Francis, B., Hinde, J. and Darnell, R. (2009). *Statistical Modelling in R.* Oxford: Oxford University Press.

Terrell, G.R. (2002). The gradient statistic. *Computing Science and Statistics*, **34** 206 – 215.

Distributional and quantile regression for quality control in fetal weight estimation

Andreas Mayr[1,2], Tobias Hepp[1], Elisabeth Waldmann[1], Matthias Schmid[2], Florian Faschingbauer[3]

[1] Friedrich-Alexander-University Erlangen-Nürnberg, Germany
[2] Rheinische Friedrich-Wilhelms-University Bonn, Germany
[3] Dept. of Obstetrics and Gynecology, University Hospital Erlangen, Germany

E-mail for correspondence: andreas.mayr@fau.de

Abstract: We propose two approaches to analyse measurement errors based on statistical modelling. The first incorporates distributional regression and aims to model systematic bias and random error simultaneously via generalized additive models for location, scale and shape (GAMLSS). The second approach focuses on quantile regression to evaluate the distribution of z-scores. All proposed models are illustrated with quality control in sonographic weight estimation, analysing the effect of the examiner and his experience on the accuracy.

Keywords: GAMLSS; Quantile regression; Measurement errors; Boosting.

1 Introduction

The estimated birth weight of the fetus is an important predictive parameter for neonatal morbidity and mortality. The estimates are based on the last sonography before birth and incorporate linear models for the different measured biometric parameters. Measurement errors are, however, inevitable and should therefore be subject to statistical analysis.
We analyse 4613 sonographic weight estimations of 18 examiners starting with the beginning of their ultrasound training (Figure 1). Typically, quality control in these kind of settings is done with the cumulative summation (CUSUM) technique (Balsyte et al., 2010) yielding individual learning curves. We propose alternative approaches based on statistical models that focus either on the distribution of the estimated birth-weight or analyse the z-scores of the underlying biometric parameters via quantile regression. Our

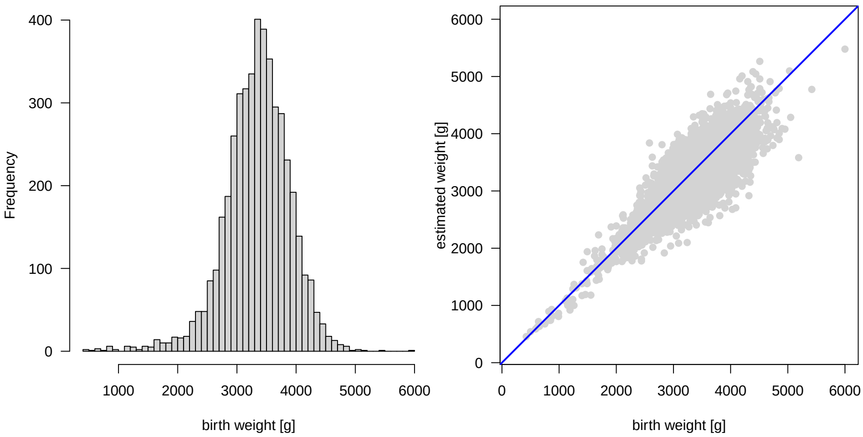


FIGURE 1. Distribution of birth weight (left) and the accuracy of sonographic weight estimation (right).

approaches allow to detect the sources of measurement errors while adjusting for confounders and provide the classical statistical inference.
We focus on the influence of the examiner performing the sonographic assessment and his experience on systematic bias and random errors via GAMLSS. Additionally, we investigate the performance of the examiner to detect clinical relevant cases via a comparison of the z-score distribution of the individual ultrasound parameters to the theoretically expected ones by quantile regression.

2 Distributional regression for measurement errors

Measurement errors can be separated into systematic bias and random error. We propose an approach to analyse both, simultaneously, via GAMLSS (Rigby and Stasinopoulos, 2005). In the easiest case, we assume the outcome to follow $N(\mu, \sigma^2)$. The basic idea is to model both parameters of this distribution for the measurements $\tilde{y}_1, ..., \tilde{y}_n$ while incorporating the true values $y_1, ..., y_n$ in the models:

$$\begin{aligned} \mu = \mathrm{E}(\tilde{Y}|Y,X) &= \beta_{0\mu} + \beta_{1\mu} y + \sum_{j=1}^{p} h_{\mu j}(x_j) \\ \log(\sigma) = \log\left(\sqrt{\mathrm{Var}(\tilde{Y}|Y,X)}\right) &= \beta_{0\sigma} + \beta_{1\sigma} y + \sum_{j=1}^{p} h_{\sigma j}(x_j) \end{aligned}$$

Variables that are assumed to have an effect on the accuracy of the measurement as well as possible confounders can be included in the additive

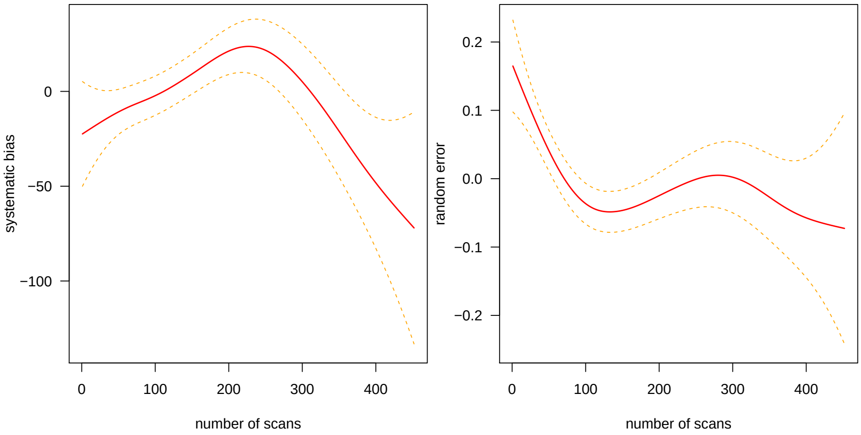


FIGURE 2. Effect of the individual experience of the examiner (number of scans) on the expected value (left) and standard deviation (right) of the estimated fetal weight, representing systematic bias and random error.

predictors via $\sum_{j=1}^{p} h_{\mu j}(x_j)$ or $\sum_{j=1}^{p} h_{\sigma j}(x_j)$, respectively. Examples can be the measurement device or the examiner. Variables that actually have an effect on the expected mean μ represent sources of systematic bias, while variables influencing the variance contribute to the random error of the measurements (Mayr et al., 2015).
In our analysis, we model the estimated fetal weight $\tilde{Y}$ while adjusting for the actual birth weight Y and other factors influencing the accuracy (e.g., maternal BMI). We are interested in the effect of the examiners and their evolution when they get more experienced: We therefore included both, the 18 different examiners as an categorical effect and the absolute number of scans they had performed before as a non-linear effect via P-splines (Figure 2).

3 Quantile regression for z-scores of parameters

The accuracy of the estimated fetal weight depends on the accuracy of the underlying four biometric parameters measured by ultrasound (head and abdomen circumference, femur length, biparietal diameter). For those measurements, however, we do not have the true observations but can only compare them to standardized reference values via z-scores

$$Z = \frac{Y_{\mathrm{GA}} - \mu(\mathrm{GA})}{\sigma(\mathrm{GA})},$$

where $\mu(\mathrm{GA})$ and $\sigma(\mathrm{GA})$ are models for the mean and standard deviation of parameter Y from a GAMLSS-based reference growth charts (Papageorghiou et al., 2014) which depend on gestational age (GA).

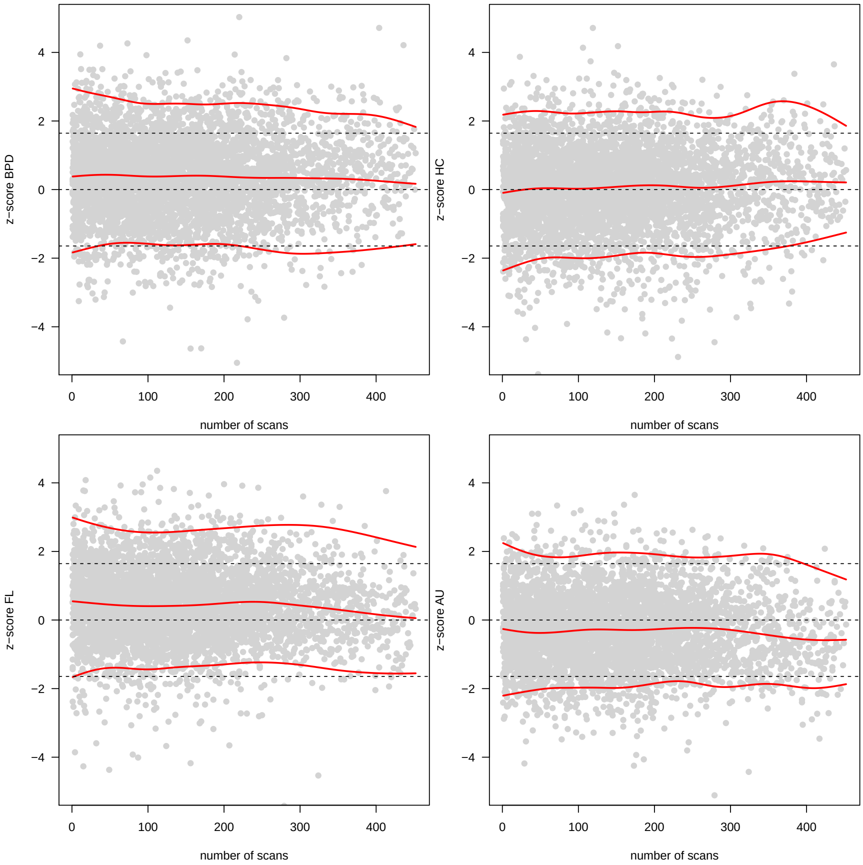


FIGURE 3. Comparing the theoretical 5%, 50% and 95% quantiles (horizontal dashed lines) with the ones resulting from quantile regression (solid red curves) for the z-scores of biomeric parameters depending on the number of scans of the examiner.

For accurate measurements, the distribution of the z-score should on average follow a standardized normal distribution. In clinical practice, pathological cases are often identified via comparing z-scores to the 5% and 95% quantiles of $N(0, 1)$ (Salomon et al, 2005). We propose to model these τ-quantiles

$$\mathrm{Q}_{\tau}(Z|X) = \beta_{0\tau} + \sum_{j=1}^{p} h_{\tau j}(x_j)$$

directly via quantile regression (Koenker, 2005) and compare them to the theoretical ones. Influential subject-specific factors as well as variables that are assumed to have an effect on the accuracy of the measurement are incorporated in $\sum_{j=1}^{p} h_{\tau j}(x_j)$.

In our analysis, we are again interested in the categorical effect of the examiner and the number of scans that were performed before. Additionally, we include the actual birth weight, maternal BMI and GA at the examination into the model (Figure 3).

4 Model inference

The gamlss package (Rigby and Stasinopoulos, 2005) must be considered the gold-standard for estimating GAMLSS via penalized maximum likelihood. During the last years, however, alternatives based on gradient boosting (Mayr et al., 2012) and Bayesian inference (Klein et al., 2015) emerged. Gradient boosting also works for high-dimensional data and can incorporate variable selection. A limitation of boosting is that standard errors or confidence intervals for effect estimates can only be computed based on resampling or permutations. Bayesian inference, on the other hand, provides the advantage of accurate credible intervals without relying on resampling or asymptotic approximations. In our setting to analyse measurement errors, all three inference schemes could be used to fit the proposed models. For quantile regression, the standard approach for model inference relies on linear programming (Koenker, 2005). However, again also Bayesian inference (Waldmann et al., 2013) and gradient boosting (Fenske et al., 2011) approaches are available. In our case, we followed the gradient boosting approach via the R add-on package mboost as it provides a very flexible implementation to incorporate different types of covariate effects.

5 Results

The examiner and its experience have a significant effect both on systematic bias and random error. As presented in Figure 2, the partial effect of the number of performed scans particularly contributes to the random error, yielding a higher variation for unexperienced examiners. This effect can also be observed in the distribution of z-scores for the biometric parameters (Figure 3). The estimated quantile curves clearly diverge from the theoretically expected values. However, the range between the quantiles for most parameters decreases with the number of scans, displaying the learning process.

From a methodological perspective, we think that distributional and quantile regression approaches provide suitable tools for quality control in cohorts of examiners. In contrast to commonly used descriptive techniques, the proposed statistical models additionally allow to detect sources of measurement errors while adjusting for confounders. For individual learning curves, one could incorporate interaction terms between the splines for the number of scans and the examiners.

Acknowledgments: The work on this article was supported by the Deutsche Forschungsgemeinschaft (http://www.dfg.de, grant number SCHM 2966/1-2) and the Interdisciplinary Center for Clinical Research of the Friedrich-Alexander University Erlangen-Nürnberg (project J49).

References

Balsyte, D., Schäfer, L., Burkhardt,J. et al. (2010). Continuous independent quality control for fetal ultrasound biometry provided by the cumulative summation technique. *Ultrasound Obstet Gynecol*, **35**: 449 – 455.

Fenske, N., Kneib, T., Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, **106**(494): 494 – 510.

Koenker, R. (2005): Quantile Regression. *Cambridge University Press.*

Klein, N., Kneib, T., Lang, S. (2015). Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data. *Journal of the American Statistical Association*, **110**(509): 405 – 419.

Mayr, A., Schmid, M., Uter, W., Pfahlberg, A. and Gefeller, O. (2015). A permutation test to analyse systematic bias and random measurement errors of medical devices via boosting location and scale models. *Stat Meth Med Res*, DOI: 10.1177/0962280215581855

Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012). Generalized additive models for location, scale and shape – a flexible approach based on boosting. *Applied Statistics*, **61**(3): 403 – 427.

Papageorghiou, A. T., Ohuma, E. O., Altman, D. G. et al. (2014). International standards for fetal growth based on serial ultrasound measurements. *The Lancet*, **384**(9946), 869 – 879.

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

Salomon, L.J., Bernard, J.P. and Ville, Y. (2005). Analysis of Z-score distribution for the quality control of fetal ultrasound measurements at 20-24 weeks. *Ultrasound Obstet Gynecol*, **26**, 750 – 754.

Waldmann, E., Kneib, T., Yue, Y.R., Lang, S. and Flexeder, C. (2013). Bayesian semiparametric additive quantile regression. *Statistical Modelling*, **13**(3), 223 – 252.

Probabilistic temperature forecasting based on an AR model fitted to forecast errors

Annette Möller[1], Jürgen Groß[2]

[1] Department of Animal Sciences, Biometrics & Bioinformatics Group, University of Göttingen, Germany
[2] Institute for Mathematical Stochastics, Faculty of Mathematics, Otto von Guericke University Magdeburg, Germany

E-mail for correspondence: annette.moeller@agr.uni-goettingen.de

Abstract: Statistical postprocessing models are widely applied to correct ensembles of deterministic numerical weather predictions for biases and dispersion errors and to obtain full predictive probability distributions. We found that the time series of forecast errors of the raw ensemble forecasts is not necessarily white noise, but can exhibit substantial autoregressive behavior. Thus, we propose to fit an AR process to the error series and develop an extension of a state-of-the-art postprocessing model based on numerical forecasts that are adjusted according to the respective AR-fit. Applied to temperature forecasts issued by the European Centre for Medium-Range Weather Forecasts (ECMWF), our proposed model shows significant improvement over a standard postprocessing model.

Keywords: Statistical postprocessing model; Predictive probability distribution; Autoregressive process; Spread-adjusted linear pool.

1 Introduction

Ensemble prediction systems aim to reflect and quantify sources of uncertainty in the deterministic numerical weather prediction (NWP) model forecasts (Gneiting et al., 2005; Leutbecher and Palmer, 2008).
However, they often fail to capture all sources of uncertainty and thus exhibit dispersion errors and biases. To deal with these issues, statistical postprocessing models have been developed and successfully applied over the last decades. Such a model employs the forecasts given by the individual ensemble members as covariates in a statistical model, where the response variable is given by the respective verifying observations. By fitting this type of model, the ensemble forecasts are corrected in accordance with

recent forecast errors and observations. A further advantage is that a full predictive probability distribution can be obtained (Gneiting and Katzfuss, 2014).

Investigations showed that the time series of the ensemble forecast errors is not necessarily white noise, but can exhibit substantial autoregressive behavior. We therefore propose to fit an AR model to the error series and develop an extended postprocessing model based on ensemble forecasts adjusted according to the AR-fit. The work presented here is part of a more extensive study, see Möller and Groß (2016).

2 Modeling autoregressive behaviour

Let $\{X_1(t), \ldots, X_m(t)\}$ denote an ensemble of forecasts for a univariate (normally distributed) weather variable $Y(t)$ at a fixed location. Let $\eta(t)$ denote a deterministic forecast of $Y(t)$ with corresponding forecast error

$$Z(t) := Y(t) - \eta(t) \; . \tag{1}$$

If a one-step-ahead forecast $\eta(t)$ made at origin $t-1$ had been obtained from an autoregressive-moving average model fit, then the residual series $Z(t)$ from (1) were a white noise process. Checking for autocorrelation in $Z(t)$ may then reveal some lack of fit. The unexplained autocorrelation information in $Z(t)$ can be utilized to improve the forecast. For this, assume that the series $\{Z(t)\}$ follows a weakly stationary AR(p) process, i.e.

$$Z(t) - \mu = \sum_{j=1}^{p} \alpha_j [Z(t-j) - \mu] + \varepsilon(t) \; , \tag{2}$$

where $\{\varepsilon_t\}$ is white noise. Combining (1) and (2) gives $Y(t) = \widetilde{\eta}(t) + \varepsilon(t)$, where

$$\widetilde{\eta}(t) = \eta(t) + \mu + \sum_{j=1}^{p} \alpha_j [Y(t-j) - \eta(t-j) - \mu] \tag{3}$$

can be seen as an AR adjusted forecast based on the actual forecast $\eta(t)$ and past values $Y(t-j)$ and $\eta(t-j)$, $j = 1, \ldots, p$. The coefficients μ, α_1, ..., α_p can be obtained by fitting an AR(p) process to the observed error series $\{Z(t)\}$ from a training period, where the order p of the process can automatically be chosen by applying a model selection criterion. This includes the incidence $p = 0$, in which case $\widetilde{\eta}(t)$ is a simple bias correction of $\eta(t)$.

3 Application to ECMWF temperature forecasts

For our case study we employ the $m = 50$ member ensemble of the European Centre for Medium-Range Weather Forecasts (ECMWF, see e.g.

Buizza et al., 2007). We consider 24-h ahead forecasts for 2-m surface temperature in Germany along with the verifying observations at different stations in the time period ranging from 2010-02-02 to 2011-04-30. Although there is a total of 518 stations in the full data set, only 383 stations with complete $T = 453$ observations were retained.

For each of the 383 stations we compute the series $Z(t) = Y(t) - \overline{X}(t)$ of forecast errors of the ensemble mean $\overline{X}$, where t ranges over the whole time period. To check for temporal independence, we apply the Ljung-Box test (Ljung and Box, 1978) based on lag 1. All 383 computed p-values are not greater than 0.046 (the largest occurring value), indicating substantial autocorrelation in the forecast error series for each station.

Figure 1 shows the series of temperature observations together with the ensemble mean, the corresponding forecast errors, and the autocorrelation function (ACF) of the series of forecast errors for the randomly chosen station Ruppertsecken.

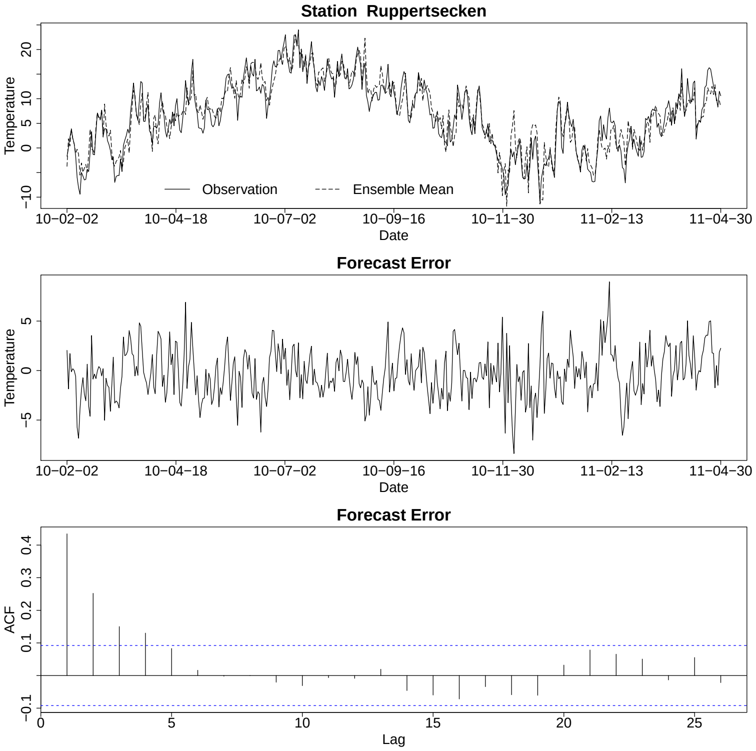


FIGURE 1. Series of temperature and ensemble mean (upper panel), series of forecast errors (middle panel), and ACF of series of forecast errors (lower panel) for station Ruppertsecken.

3.1 Predictive distribution based on autoregressive adjustment

Preliminary investigations showed that the error series belonging to the individual members exhibit quite different autoregressive behaviour. Therefore, the procedure performed on the ensemble mean is now applied to each ensemble member individually, meaning that the parameter estimates of the AR-fit are separately computed for each member.

To construct a predictive distribution based on the AR-adjusted forecast ensemble, we follow a similar approach as used for the state-of-the-art postprocessing method called Ensemble Model Output Statistics (EMOS, Gneiting et al., 2005). We assume a Gaussian predictive distribution

$$\mathcal{N}(\xi(t), \sigma^2(t)) \,, \tag{4}$$

for the weather quantity $Y(t)$ (in our case temperature), given the ensemble forecasts $\{X_1(t), \ldots, X_m(t)\}$.

In case of EMOS, $\xi(t)$ is a linear combination of the ensemble members and $\sigma^2(t)$ is a linear function of the ensemble variance. The coefficients are estimated by minimizing with respect to a proper scoring rule.

To obtain a predictive distribution based on our AR-adjustment of the forecast ensemble (which we call AR-EMOS) we suggest a different estimation procedure, that is, the following plug-in strategy.

The parameter $\xi(t)$ is now estimated by the mean of the AR adjusted forecast ensemble $\overline{\widetilde{X}}(t)$, with $\widetilde{X}_1(t), \ldots, \widetilde{X}_m(t)$ denoting the AR-adjusted ensemble. The parameter $\sigma^2(t)$ is estimated as mean over all estimated variances of the individual members $\eta(t) = X_i(t)$, which in turn are obtained via the moving average representation of the respective $Z(t)$ (see, e.g., Shumway and Stoffer, 2006).

For both, EMOS and AR-EMOS we proceed by estimating the coefficients station-wise (local approach).

TABLE 1. Verification statistics averaged over $T_2 = 338$ days and 383 stations.

	EMOS	AR-EMOS	SLP
MAE	2.042	2.036	1.969
CRPS	1.471	1.460	1.407
DSS	3.135	2.908	2.821

As visible in the first two columns of Table 1 our AR-EMOS approach based on the AR-adjusted ensemble shows slightly improved predictive performance with respect to several verification scores (Wilks, 2011). The respective PIT histograms (Wilks, 2011) presented in Figure 2 indicate that our proposed method improves calibration properties as well. However, the histograms also indicate slightly inverse dispersion properties: While the

EMOS PIT Histogram exhibits a U-shape that is typically for underdispersion, the AR-EMOS histogram has a slight hump-shape indicating the reverse effect, that is a small overdispersion.

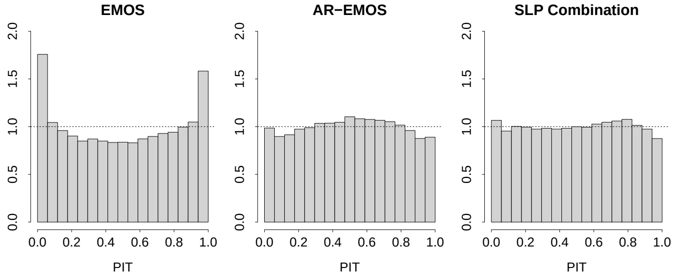


FIGURE 2. Univariate verification rank histogram and PIT histograms over 383 stations and all dates in the verification period.

3.2 Combination of predictive distributions

To neutralize the somewhat contradictory properties of EMOS and AR-EMOS, we proceed by combining the two distributions with a spread-adjusted linear pool (SLP, Gneiting and Ranjan, 2013). In our case the two component CDFs are Gaussian, so that $F_l^0(y) = \Phi(y/\sigma_l)$, $l = 1, 2$, and the SLP combined predictive CDF is

$$F(y) = w_1 G_1(y) + w_2 G_2(y), \quad G_l(y) = \Phi\left(\frac{y - \mu_l}{\sigma_l c}\right), \tag{5}$$

$l = 1, 2$, where w_1 is nonnegative, $w_2 = 1 - w_1$, and c is a strictly positive spread adjustment parameter (Gneiting and Ranjan, 2013).

For an appropriate choice of the SLP parameters when combining EMOS and AR-EMOS, we investigate a grid of 99 combinations of values for w_1 (w_2 is fully determined by w_1) and c.

For each of the investigated combinations, the average DSS and CRPS over all 338 days and 383 stations is computed, yielding (for both scores) a minimal average for the simple unfocused combination $w_1 = w_2 = 0.5$ and $c = 1$. As it seems, within the unfocused SLP combination the contradictory dispersion properties of EMOS and AR-EMOS mutually compensate, yielding a predictive distribution with further improved predictive performance and calibration, see the third column in Table 1 and the third panel in Figure 2.

4 Concluding Remarks

We propose a method that accounts for potential autoregressive structures in forecast errors of an NWP forecast ensemble. The AR-adjustment is straightforward to compute and can be utilized to simply obtain an AR-adjusted forecast ensemble or to construct different types of predictive distributions. In our case study we suggest to built an EMOS-like predictive distribution based on the AR-adjusted forecast ensemble and in a second step obtain an aggregated predictive distribution that comprises of the state-of-the-art EMOS predictive distribution and our AR-EMOS variant. This combined distribution improves dispersion and calibration properties to a high extend and thus leads to better predictive skill.

Acknowledgments: We are grateful to the European Centre for Medium-Range Weather Forecasts (ECMWF) and the German Weather Service (DWD) for providing forecast and observation data, respectively. We further wish to thank Tilmann Gneiting and Stephan Hemri for useful discussions and helpful comments.

References

Buizza, R., Bidlot, J.R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G. and Vitart, F. (2007). The new ECMWF VAREPS (variable resolution ensemble prediction system). *Quarterly Journal of the Royal Meteorological Society*, **133**, 681 – 695.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, **1**, 125 – 151.

Gneiting, T., Raftery, A.E., Westveld III, A.H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133**, 1098 – 1118.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, **7**, 1747 – 1782.

Leutbecher, M. and Palmer, T.N. (2008). Ensemble forecasting. *Journal of Computational Physics*, **227**, 3515 – 3539.

Ljung, G.M. and Box, G.E. (1978). On a measure of lack of fit in time series models. *Biometrika*, **65**, 297 – 303.

Möller, A. and Groß, J. (2016). Probabilistic temperature forecasting based on an ensemble AR modification. *Quarterly Journal of the Royal Meteorological Society*, DOI:10.1002/qj.2741.

Shumway R.H., Stoffer D.S. (2006). *Time Series Analysis and Its Applications. With R Examples.* Second Edition. Springer.

Wilks, D.S. (2011). *Statistical Methods in the Atmospheric Sciences.* Academic Press.

N-mixture models applied to zero-inflated insect abundance data

Rafael A. Moral[1], John Hinde[2], Clarice G. B. Demétrio[1],
Carolina Reigada[1], Wesley A. C. Godoy[1]

[1] ESALQ/USP, Brazil
[2] NUI Galway, Ireland

E-mail for correspondence: `rafael.moral@usp.br`

Abstract: In ecological field surveys it is often of interest to estimate the abundance of species. However detection is imperfect and hence it is important to model these data taking into account the ecological processes and sampling methodologies. In this context, N-mixture models and extensions are particularly useful, as it is possible to estimate population size and detection probabilities under different ecological assumptions. We apply extensions of this class of models to estimate the abundance of *Pachycrepoideus vindemiae*, a parasitoid of blowflies. We will also develop methods for assessing goodness-of-fit by proposing different types of residuals for this model class.

Keywords: Ecology of parasitoids; *Lucilia sericata*, *Pachycrepoideus vindemiae*; Population size estimation.

1 Introduction

It is very important in ecological contexts to measure animal abundance and understand how this abundance changes over time and space. There are different statistical models that may be used to estimate abundance as well as site-occupancy. N-mixture models were defined by Royle (2004) and have been generalised ever since, see Dail and Madsen (2010) and Hostetler and Chandler (2015). Here we develop and apply extensions of this class of models to estimate parasitoid abundance given different hosts in a field survey. So far specific forms of residuals and model diagnostics have not been proposed and we will develop goodness-of-fit assessment techniques for these models.

2 Case-study

Pachycrepoideus vindemiae is a generalist and solitary parasitoid which normally lays only one egg per host pupa. This species presents facultative hyperparasitism, i.e., when other parasitoid's larvae are present in a pupa, *P. vindemiae* larvae may kill their competitors.

A field survey was conducted from 2005 to 2007 in three different areas (rural, urban and forest) in the surroundings of the Brazilian municipality of Botucatu, in São Paulo state. Five different hosts (*Lucilia sericata*, *Chrysomya albiceps*, *Chrysomya megacephala*, *Chrysomya putoria*, and *Cochliomyia macellaria*) were placed in cages to attract the parasitoids. In each area, three cages were placed and observed on 49 occasions, totaling $3 \times 3 \times 49 = 441$ observations per host. After seven days in the field, the cages were removed and the number of parasitoids was counted.

3 Methodology

Let n_{it} represent insect counts for site $i, i = 1, \ldots, R$ over sampling occasion $t = 1, \ldots, T$. We are interested in estimating site abundance N_i, however there is a detection (or capture) probability p which is also unknown. Considering closed populations (i.e. no migration and constant birth and death rates), we may assume that n_{it} are independent and identically distributed as Binomial(N_i, p). The likelihood may be written as

$$L(N_1, \ldots, N_R, p | n_{11}, \ldots, n_{RT}) = \prod_{i=1}^{R} \left\{ \prod_{t=1}^{T} \binom{N_i}{n_{it}} p^{n_{it}} (1-p)^{N_i - n_{it}} \right\}. \quad (1)$$

The approach described by Royle (2004) takes N_i to be independent and identically distributed latent random variables with density $f(N_i; \theta)$, and marginalising (1) with respect to N_i. Hence the likelihood function of the N-mixture model may be written as

$$L(\theta, p | n_{11}, \ldots, n_{RT}) = \prod_{i=1}^{R} \left\{ \sum_{N_i = \max_t n_{it}}^{\infty} \prod_{t=1}^{T} \binom{N_i}{n_{it}} p^{n_{it}} (1-p)^{N_i - n_{it}} f(N_i; \theta) \right\} \quad (2)$$

Sensible choices for the distribution of N_i are the Poisson and negative binomial models, for example. This model may be extended for zero-inflated distributions and also to relax the closure assumption, including more latent variables to model mortality, recruitment and migration (Hostetler and Chandler, 2015). All analyses were carried out in software R (R Core Team, 2015).

4 Results and discussion

Exploratory analysis of the data show that abundance is probably higher in rural areas (see Fig. 1(a)), which may be due to the presence of carcasses in rural areas that attract the host. Also, detection probability may be lower in urban areas (see Fig. 1(b)), and this may be due to the fact that in the variety of chemical compounds in the air make it difficult for the parasitoids to be attracted by the traps.

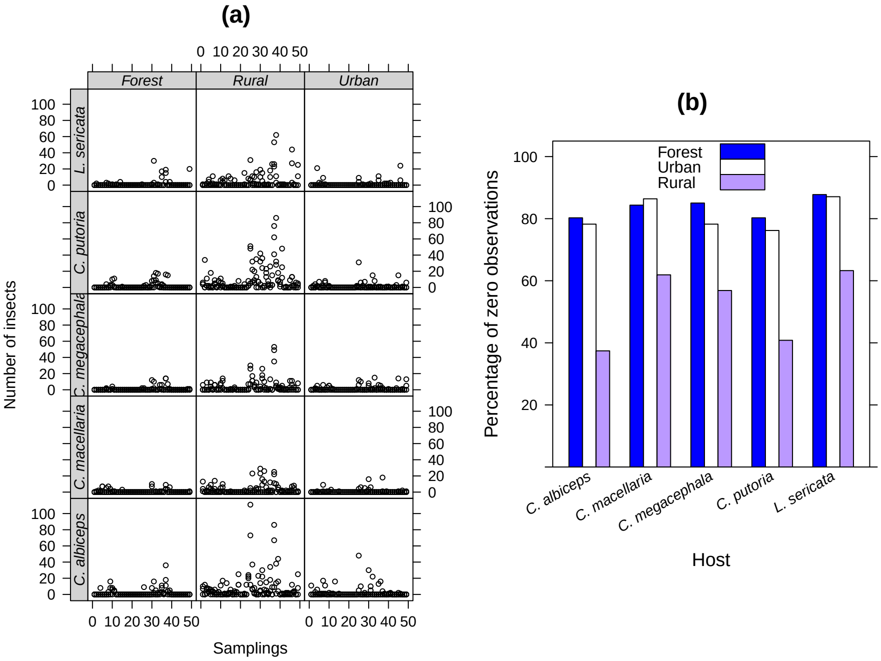


FIGURE 1. (a) Number of collected insects through time for each host × habitat combination and (b) percentage of zero observations for each host × habitat combination.

We fitted the N-mixture model for the abundance of *P. vindemiae* with *L. sericata* as a host taking the distribution of N_i as Poisson, negative binomial (NB), and zero-inflated Poisson (ZIP). The best model fit according to the Akaike Information Criterion (AIC) was the NB mixture including the effect of area in the detection probability and only an intercept for the abundance parameter (see Table 1).

Grouping forest and urban areas and fitting the NB model with area as a covariate for detection probability and only an intercept for abundance, the likelihood-ratio test indicates that the model including all three levels of area does not fit the data significantly better than the reduced model

TABLE 1. Akaike information criterion (AIC) for each N-mixture model fitted to parasitoid abundance data for host *L.* sericata (NB = negative binomial; ZIP = zero-inflated Poisson).

Abundance covariates	Detection covariates	N_i distribution	AIC
×	×	Poisson	3755.42
×	area	Poisson	3490.75
area	area	Poisson	3493.13
×	×	NB	3519.35
×	area	NB	**3443.36**
area	area	NB	3446.10
×	×	ZIP	3757.43
×	area	ZIP	3492.77
area	area	ZIP	3495.13

(LR=0.80, d.f.=1, p=0.37). We conclude that the detection probability in rural sites is significantly lower than for forest and urban sites, however the abundance is statistically the same for different area types (LR=1.25, d.f.=2, p=0.53).

It is useful to assess goodness-of-fit in this setting so that abundance may not be over- or underestimated. We propose using half-normal plots with simulation envelopes of residuals. However, no specific residual form for these models have yet been proposed other than the ordinary residuals, which are given by the difference between the observed data and the fitted values. Producing these plots for the three considered abundance mixtures and including only an intercept for the abundance model and area as a detection covariate indicates that the models fail to fit the data well (see Fig. 2) and hence the inclusion of other covariates and the use of other model extensions should be explored. This may also indicate that the distribution of the abundances may not be zero-inflated and that the many zero observations may be due to low detection probability.

Other subject of ongoing work is the joint modelling of the abundance of two species using the N-mixture framework, where the abundance of one species may affect the other's, such as in a predator-prey or host-parasitoid system.

Acknowledgments: Special Thanks to FAPESP for funding (proc no. 2014/12903-8).

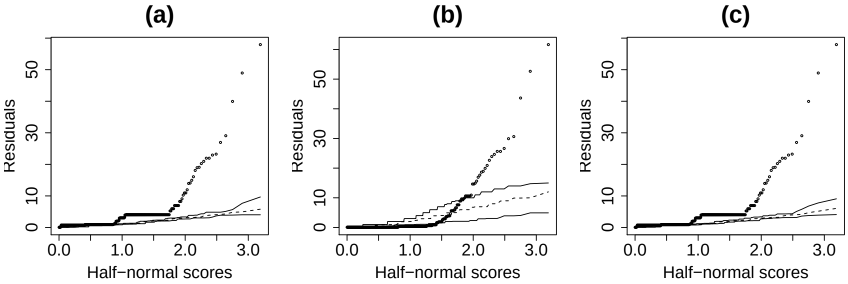


FIGURE 2. Half-normal plots with simulated envelopes of ordinary residuals for the models considering the abundance distribution as (a) Poisson, (b) negative binomial, and (c) zero-inflated Poisson including area as a detection covariate and only an intercept for the abundance model.

References

Dail, D. and Madsen, L. (2010). Models for estimating abundance from repeated counts of an open metapopulation *Biometrics*, **67**, 577 − 587.

Hostetler, J.A. and Chandler, R.B. (2015). Improved state-space models for inference about spatial and temporal variation in abundance from count data. *Ecology*, **96**, 1713 − 1723.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Royle, J.A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, **60**, 108 − 115.

Nonparametric imputation based on data depth

Pavlo Mozharovskyi[123], Julie Josse[34], François Husson[23]

[1] Centre Henri Lebesgue, Rennes, France
[2] Institut de Recherche Mathématique de Rennes, France
[3] Agrocampus Ouest, Rennes, France
[4] Institut National de Recherche en Informatique et en Automatique, Orsay, France

E-mail for correspondence: `pavlo.mozharovskyi@univ-rennes1.fr`

Abstract: A method for single imputation of missing values is presented. It consists in iterative maximization of data depth of each observation with missing values, and can be used with any properly defined depth. The method is robust, distribution-free, and applicable to general elliptically symmetric densities. Its particular case has direct connection to the well know treatments for multivariate normal model.

Keywords: Missing data; Data depth; Single imputation; Elliptical symmetry.

1 Introduction

The problem of missing values exists since the earliest attempts of exploiting data as a source of knowledge as it lies intrinsically in the process of obtaining, recording, and preparation of the data itself. The most naïve treatment consists in dropping rows or columns, depending on the view on the data, but by deleting the entire row (column) present data is removed as well. And if a data set contains one or a few missing values in a large portion of rows, substantial part of data can be missed by this list-wise deletion. To exploit all the information present in the data set, a statistical method may be adapted to missing values, but this requires developing such a one for each estimator and inference of interest. A more universal way is to impute missing data first, and then apply the statistical method of interest to the completed data set (Little and Rubin, 2002). Lastly, the multiple imputation has gained a lot of attention: for a data set containing missing values, a number of completed data sets is generated reflecting uncertainty of the imputation process, which enables not only estimating the value of interest but also drawing an inference on it (Van Buuren, 2012). Nevertheless, single imputation, *i.e.* just meaningfully replacing missing values, is still paid attention in the statistical literature. This can be appropriate when one needs just to complete a single data set, when no inference is required, when the applied statistical method is computationally too demanding for multiple data sets, or when a few values are missing only but one seeks an alternative to the list-wise deletion.

2 Proposal

One of the existing approaches to single imputation is to replace a missing value by its conditional mean, based on a specific joint model. Being of highest importance, multivariate normal distribution and perturbing this mechanisms have gained a lot of attention in the imputation literature. In the present work, we propose a single imputation method able to properly work for a broader class of elliptically symmetric distributions — a natural generalization of the multivariate normal model. The suggested technique is based on the notion of statistical centrality measure — data depth, and is generic in it. Before presenting the approach in Section 2.2, we refer to the notion of data depth in Section 2.1.

2.1 Data depth

Consider a point $\boldsymbol{x}_0 \in \mathbb{R}^d$ and a random sample $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ in $\mathbb{R}^d$. A statistical data depth is a function $D(\boldsymbol{x}_0|\boldsymbol{X}) : \mathbb{R}^d \to [0, 1]$ that describes how deep, or central the observation $\boldsymbol{x}_0$ is located w.r.t. $\boldsymbol{X}$. To be a well behaving depth, $D(\cdot|\cdot)$ should satisfy elementary postulates: be affine invariant, vanishing at infinity, non-increasing on any ray from the deepest point ($\arg\max_{\boldsymbol{x}_0 \in \mathbb{R}^d} D(\boldsymbol{x}_0|\boldsymbol{X})$) or even quasi-concave, and upper semi-continuous; see Mosler (2013) for a recent survey. $D(\boldsymbol{x}_0|\boldsymbol{X})$ provides a multivariate center-outward ordering, *i.e.* points closer to the center should have higher depth, and those more outlying smaller one. During the last decades, a number of notions of statistical depth function differing in properties and areas of application have been developed. For shortness and demonstrative reasons we proceed with the historically first Tukey depth below.

The Tukey (or halfspace, also location) depth (Tukey, 1975) of $\boldsymbol{x}_0$ w.r.t. $\boldsymbol{X}$ is defined as the smallest portion of $\boldsymbol{X}$ that can be contained in a closed halfspace with $\boldsymbol{x}_0$ on its boundary

$$D(\boldsymbol{x}_0|\boldsymbol{X}) = \frac{1}{n} \min_{\boldsymbol{r} \in S^{d-1}} \#\{i | \boldsymbol{x}_i'\boldsymbol{r} \geq \boldsymbol{x}_0'\boldsymbol{r}, i = 1, ..., n\}. \qquad (1)$$

2.2 Iterative approach

Given (a complete) $\boldsymbol{X}$, let $\boldsymbol{x} \in \mathbb{R}^d$ be an observation with missing coordinates, and index its existing entries by $\boldsymbol{x}_{obs}$ and missing with $\boldsymbol{x}_{miss}$. Denoting $D_\alpha(\boldsymbol{X})$ an α-upper-level set of $D(\cdot|\boldsymbol{X})$ (or depth-trimmed region), and denoting interior by *int*, let

$$\alpha^* = \inf_{\alpha\in(0;1)} \bigl\{\alpha \,|\, intD_\alpha(\boldsymbol{X}) \cap \{\boldsymbol{y} \,|\, \boldsymbol{y} \in \mathbb{R}^d\,,\, \boldsymbol{y}_{obs} = \boldsymbol{x}_{obs}\} = \emptyset\bigr\} \tag{2}$$

be the depth of the region with the smallest depth not touching the missing affine subspace of $\boldsymbol{x}$, or exactly touching it when $D(\cdot|\cdot)$ is continuous. We impute $\boldsymbol{x}$ by

$$\boldsymbol{x} = ave\bigl(\mathop{\arg\min}_{\boldsymbol{y}\in\mathbb{R}^d\,,\,\boldsymbol{y}_{obs}=\boldsymbol{x}_{obs}} \{\|\boldsymbol{y} - \boldsymbol{z}\| \,|\, \boldsymbol{z} \in \mathbb{R}^d\,,\, \boldsymbol{z} \in D_{\alpha^*}(\boldsymbol{X}), \|\}\bigr), \tag{3}$$

with *ave* being the averaging operator. In this way, discrete depth functions as well those vanishing immediately beyond the convex hull of data (as, *e.g.*, the Tukey depth) are accounted for, also computationally; see Figure 1 for a data set from `http://stat.ethz.ch/Teaching/Datasets/`. On the other hand, as noted above, if $D(\cdot|\cdot)$ is continuous, one can explicitly write

$$\boldsymbol{x} = \mathop{\arg\max}_{\boldsymbol{y}\in\mathbb{R}^d\,,\,\boldsymbol{y}_{obs}=\boldsymbol{x}_{obs}} D(\boldsymbol{y}|\boldsymbol{X}), \tag{4}$$

i.e. (instead of taking conditional mean) a point of the highest depth conditioned on $\boldsymbol{x}_{obs}$ and on $\boldsymbol{X}$ is taken.

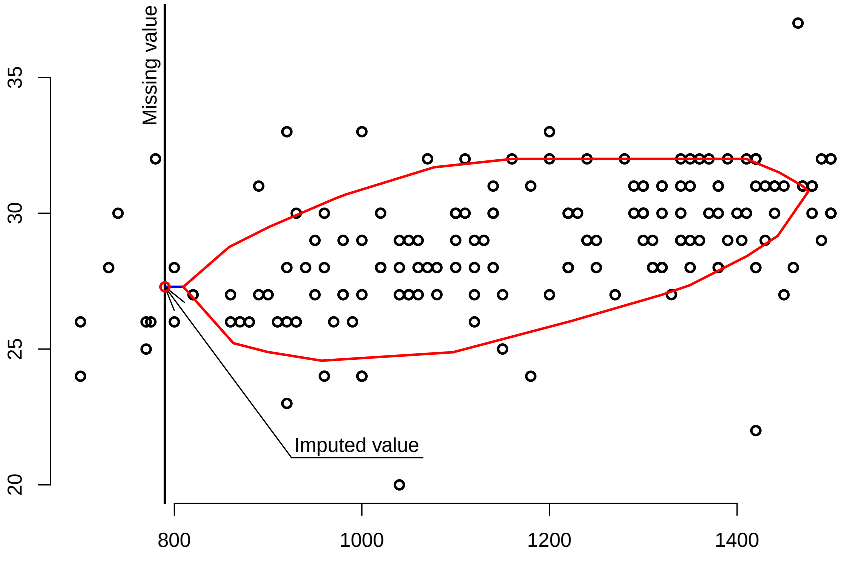


FIGURE 1. An imputation for the Babies data set using Tukey depth.

Given a data set with missing entries, we first fill not available data with starting values (say coordinate wise mean of the existing ones). Then to each observation with initially missing entries, (3, respectively 4) is applied, in this way updating all the missing entries. The process is iterated till convergence.

3 Discussion

The proposed method is general and generic, and can be coupled with any measure of centrality essentially defining its properties. Thus when employed with Mahalanobis (1936) depth, it imputes by iterated multiple-output regression, which coincides exactly with single imputation by iterated regression. Additionally, it can be shown that it yields exactly the same solution as imputation by the regularized PCA (Josse and Husson, 2012) when assuming rank equal to $d-1$ and any admissible variance of noise. Indeed, after convergence, each missing entry lies in the hyperplane of regressing on other coordinates; if for some $\boldsymbol{x}$ $\#miss > 1$, then on the intersection of several such regression hyperplanes, *i.e.* in general in a multiple output regression affine subspace of dimension $\#obs$.
With Tukey or projection depth, it yields a distribution-free imputation scheme, fitting missing value close to the data geometry. It does not exploit any estimates of location or scatter, avoiding problems with, *e.g.*, mean and covariance matrix, in a natural way. The approach is robust both in sense of outliers and heavy-tailed distributions, and for the class of continuous elliptically symmetric distributions imputed points converge to the points of the highest conditional density.

References

Josse, J. and Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, **153**, 1 – 21.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Hoboken: John Wiley & Sons.

Mosler, K. (2013) Depth statistics. In: *Robustness and Complex Data Structures, Festschrift in Honour of Ursula Gather*, Springer, Berlin, 17 – 34.

Tukey, J.W. (1975). Mathematics and the picturing of data. In: *Proceeding of the International Congress of Mathematicians (Volume 2)*, Canadian Mathematical Congress, Vancouver, 523 – 531.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Boca Raton: Chapman & Hall/CRC.

The latent scale covariogram: a tool for exploring the spatial dependence structure of non-normal responses

Samuel D. Oman[1], Jorge Mateu[2]

[1] Hebrew University, Jerusalem, Israel
[2] Universitat Jaume I de Castellon, Castellon, Spain

E-mail for correspondence: oman@mail.huji.ac.il

Abstract:
Let Y_i be spatially dependent non-normally distributed responses (e. g., disease prevalence in different regions) which we wish to model in terms of vectors $\mathbf{x}_i$ of explanatory variables, using a hierarchical generalized linear model (GLIM) in which the dependence structure is expressed via a latent Gaussian field $\mathbf{Z} = \{Z_i\}$. At the exploratory stage, it is common practice to first fit a GLIM assuming independence, and then examine the variogram of the residuals $Y_i - \hat{Y}_i$ to determine a possible parametric model for the autocorrelation function of $\mathbf{Z}$. This is not appropriate, however, since (unless an identity link function is used) Y_i and Z_i are on different scales. We propose here an alternative, the latent scale covariogram, whose graph reflects the autocorrelation structure of the underlying Gaussian field. We illustrate its use on a data set involving parasite counts, and obtain results quite different from those obtained using the variogram.

Keywords: Generalized Linear Model; Spatial Correlation; Variogram.

1 Introduction

When modelling spatially distributed normal responses Y_i in terms of vectors $\mathbf{x}_i$ of explanatory variables, one may fit a linear model assuming independence and then use the empirical variogram of the residuals to suggest an appropriate parametric form for the autocorrelation function. Suppose, however, that the responses are not normally distributed: for example, a Poisson or binomial distribution would be more appropriate if Y_i measures disease prevalence in region i. In such cases one may use a hierarchical generalized linear model (GLIM) in which, conditional on a latent Gaussian field $\mathbf{Z} = \{Z_i\}$, the Y_i have independent distributions from the exponential family, with an appropriate link function connecting their conditional means with the linear predictors $\mathbf{x}_i^T\beta + Z_i$. The question is then how to determine an appropriate model for the autocorrelation function of $\mathbf{Z}$. An empirical variogram of the residuals from fitting a GLIM without random effects is no longer appropriate, since (unless the link function is the identity) it is on the wrong scale. We propose here an alternative, the latent scale covariogram, whose graph reflects the autocorrelation structure of the underlying normal field. We illustrate its use on a data set involving parasite counts, and obtain results quite different from those obtained using the variogram.

2 The Latent Scale Covariogram

Consider $m_i \geq 1$ responses Y_{ij}, together with corresponding vectors $\mathbf{x}_{ij}$ of explanatory variables, defined at sites s_i, $i = 1, \ldots, n$; and let $d_{ij} = ||s_i - s_j||$ denote the distance between the sites. Let $Z_i \sim N(0, \sigma^2)$ be latent variables defined at s_i, with isotropic autocorrelation function $\rho(d_{ij}) \equiv \mathrm{cor}(Z_i, Z_j)$. For a link function $h(\cdot)$ and linear predictors $\eta_{ij} = \mathbf{x}_{ij}^T\beta$, the Y_{ij} are assumed independent, conditional upon $\mathbf{Z} = \{Z_i\}$, with (as in McCullagh and Nelder, 1989)

$$E(Y_{ij} \mid Z_i) = A_{ij}h(\eta_{ij} + Z_i) \equiv m_{ij}^*(Z_i)$$

for known A_{ij} (for example, Poisson offsets), and

$$\mathrm{var}(Y_{ij} \mid Z) = (\phi/w_{ij})b(m_{ij}^*(Z_i)) \equiv v_{ij}^*(Z_i) \tag{1}$$

for known weights w_{ij} and an overdispersion parameter $\phi > 0$ which either equals one or must be estimated. It follows that $E(Y_{ij}) = E(m_{ij}^*(Z_i))$ and $\mathrm{var}(Y_{ij}) = \mathrm{var}(m_{ij}^*(Z_i)) + E(v_{ij}^*(Z_i))$. As in Zeger et al (1988), we write

$$\begin{array}{ccccc} Y_{ij} & \approx & m_{ij}^*(Z_i) & + & \varepsilon_{ij} \\ & = & A_{ij}h(\eta_{ij} + Z_i) & + & \varepsilon_{ij} \end{array}$$

where $\{\varepsilon_{ij}\}$ and $\{Z_i\}$ are independent, with $\mathrm{var}(\varepsilon_{ij}) = E(v_{ij}^*(Z_i)) \equiv v_{ij}$. A Taylor approximation then gives

$$\begin{array}{ccccccc} Y_{ij} & \approx & A_{ij}h(\eta_{ij}) & + & A_{ij}h'(\eta_{ij})Z_i & + & \varepsilon_{ij} \\ & \equiv & \mu_{ij} & + & a_{ij}Z_i & + & \varepsilon_{ij}. \end{array}$$

Averaging on j, we obtain $\bar{Y}_{i\cdot} \approx \bar{\mu}_{i\cdot} + \bar{a}_{i\cdot}Z_i + \bar{\varepsilon}_{i\cdot}$, so that

$$(\bar{Y}_{i\cdot} - \bar{\mu}_{i\cdot})/\bar{a}_{i\cdot} \approx Z_i + \bar{\varepsilon}_{i\cdot}/\bar{a}_{i\cdot} \tag{2}$$

where Z_i and $\bar{\varepsilon}_{i\cdot}$ are independent. If $\hat{\mu}_{ij}$ and $\hat{a}_{ij}$ are computed from preliminary estimates $\hat{\eta}_{ij}$, we define the latent scale covariogram to be a graph of (a binned version of)

$$\frac{(\bar{Y}_{i\cdot} - \bar{\hat{\mu}}_{i\cdot})(\bar{Y}_{j\cdot} - \bar{\hat{\mu}}_{j\cdot})}{\bar{\hat{a}}_{i\cdot}\bar{\hat{a}}_{j\cdot}} \quad \text{vs} \quad \|s_i - s_j\|. \tag{3}$$

It follows from (2) that this should approximate a graph of $\text{cov}(Z_i, Z_j)$ vs $\|s_i - s_j\|$. ,
Observe (suppose for simplicity $m_i \equiv 1$) that fitting a GLIM assuming independence and then normalizing the Y_i by their estimated standard deviations gives the Pearson residuals $(Y_i - \hat{\mu}_i)/\sqrt{\hat{v}_i}$, where $\hat{v}_i = (\phi/w_i)b(\hat{\mu}_i)$ for the variance function $b(\cdot)$ given in (1). The estimated left-hand side of (2) is also a normalized residual, but with a different normalizing constant $\hat{a}_i$. Table 1 shows the values of $\sqrt{v_i}$ and a_i for several GLIM distributions and link functions. Note that in some cases, the LSC normalizes by an estimate of the variance, not of the standard deviation.

TABLE 1. Normalizing parameter a_i and standard deviation $\sqrt{v_i}$

Distribution	**Link**	μ_i	a_i	$\sqrt{v_i}$
Ber(p_i)	logit	$p_i = e^{\eta_i}/(1 + e^{\eta_i})$	$p_i(1 - p_i)$	$[p_i(1 - p_i)]^{1/2}$
Ber(p_i)	probit	$p_i = \Phi(\eta_i)$	$\phi(\eta_i)$	$[p_i(1 - p_i)]^{1/2}$
Poisson(λ_i)	log	$\lambda_i = e^{\eta_i}$	λ_i	$\lambda_i^{1/2}$
Neg-bin(μ, θ)	log	$\mu_i = e^{\eta_i}$	μ_i	$[\mu_i + \mu_i^2/\theta]^{1/2}$

3 An Example

These data (Bockarie et al, 1998) are from a drug trial concerning parasite counts in hamlets at $n = 147$ sites $\{s_i\}$ in a rural area of the East Sepic Province of Papua, New Guinea. For $j = 1, \ldots, m_i$ individuals in hamlet i the parasite count Y_{ij} was measured, together with the explanatory variables sex_{ij} ($= 0/1$ for female/male) and age_{ij} (in years); this resulted in a total of $N = m_1 + \cdots + m_n = 2219$ observations. Alexander et al (2000) modeled the data using a negative binomial distribution. Letting $\mathbf{x}_{ij}^t = (1, \quad \text{sex}_{ij}, \quad \text{age}_{ij})$ and denoting by $\mathbf{Z}$ a vector of latent hamlet effects, they assumed that $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = ((1/\phi)\exp(-\|s_i - s_j\|/\alpha))$, and that $Y_{ij} \mid \mathbf{Z} \sim_{ind} \text{NB}(\mu_{ij}, \theta)$ with $\mu_{ij} = \exp(\eta_{ij} + Z_i)$ for $\eta_{ij} = \mathbf{x_{ij}}^t\beta$. Here, the negative-binomial distribution is parameterized so that $\text{var}(Y_{ij} \mid \mathbf{Z}) = \mu_{ij}(1 + \mu_{ij}/\theta)$. Before fitting a model, they examined the form of the covariance structure using a variogram of the hamlet-level averages $\bar{Y}_{i\cdot}$ (their Figure 3a); this indicated spatial correlation up to approximately

6 km, and then leveled out. After fitting the spatial model, a variogram of the averaged Pearson residuals (their Figure 3b) was essentially constant, indicating no remaining spatial correlation or trend; moreover, the posterior median and mean of α (1.998 and 2.414) in fact correspond to ranges of 6.0 and 7.2 km.

However, using the LSC instead of the variogram gives a very different picture. The LSC of the hamlet-level averaged residuals from a model without spatial effects (Figure 1a) decreases to zero at approximately 4 km, but then becomes negative and continues to decrease, suggesting a spatial trend. A map of these residuals (Figure 1b) indicates that this apparent trend may be due to a difference between the eastern and western parts of the region; cf. the discussion at the bottom of p. 458 in Alexander et al, (2000). We therefore added a binary variable *east* to indicate whether or not the settlement is in the east (longitude $>$ 8 km in Figure 1). Table 2 shows the results of the two fits. We see that the estimated coefficient of *east* is positive, as expected from Figure 1a, and extremely significant; moreover, there is a 30% decrease in the coefficient of *age*. The map of the residuals from this model (Figure 1d) is improved, and the corresponding LSC (Figure 1c) now levels out and indicates an autocorrelation range of approximately 2 km. Note (Alexander (2000), p. 458) that the maximum distance flown by the main mosquito vector is 1.8 km.

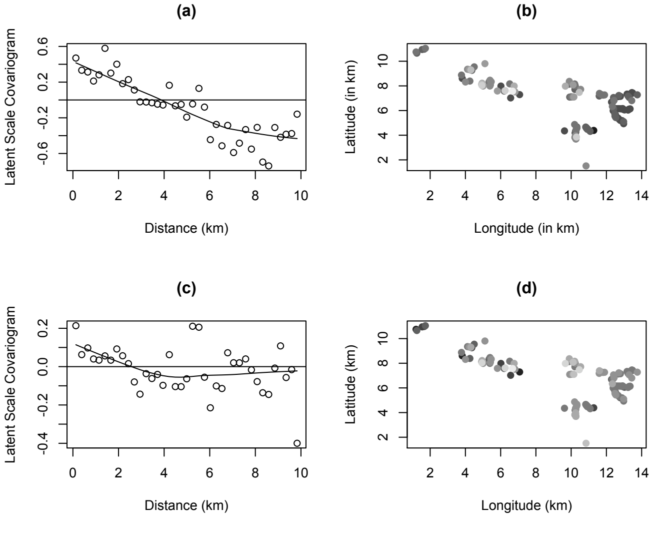


FIGURE 1. LSC (a) and map (b) of residuals from a negative binomial fit to the parasite data, without a regional effect; LSC (c) and map (d) of residuals from a fit including a regional effect.

In summary, an exploratory analysis using a variogram of the Pearson

TABLE 2. Coefficient estimates for non-spatial models.

Without regional effects				
Coefficient	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	5.781	0.158	36.66	< 2e-16
sex	0.153	0.147	1.04	0.299
age	0.028	0.004	6.49	8.82e-11
AIC: 20245				
With regional effects				
Coefficient	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	5.108	0.163	31.25	< 2e-16
east	1.232	0.171	7.20	6.24e-13
sex	0.155	0.145	1.07	0.286
age	0.037	0.004	8.49	< 2e-16
AIC: 20189				

residuals leads to a model with a 6 km range for the autocorrelation function and does not indicate a trend. However, using the LSC suggests a model with settlement clustering effects, together with a much shorter range for the autocorrelation function.

References

Alexander, N., Moyeed, R. and Stander, J. (2000). Spatial modelling of individual-level parasite counts using the negative binomial distribution. *Biostatistics*, **1**, 453– 463.

Bockarie, M.J., de Alexander, N., Hyuna, P., Dimbera, Z., Bockariea, F., Ibama, E., Alpers, M.P., and Kazura, J.W. (1998). Randomised community based trial of annual single-dose diethylcarbamazine with or without ivermectin against Wuchereria bancrofti infection in human beings and mosquitoes. *The Lancet*, **351**, 162– 168.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, Second Edition*. London: Chapman & Hall.

Zeger, S.L., Liang, K-L., and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049– 1060.

Bayesian Effect Fusion For Categorical Predictors With Sparse Finite Mixture Prior

Daniela Pauger[1], Gertraud Malsiner-Walli[1], Helga Wagner[1]

[1] Johannes Kepler University Linz, Austria

E-mail for correspondence: daniela.pauger@jku.at

Abstract: Sparse modelling is an important issue particularly in regression models with categorical predictors. One way to reduce the dimension of the model is to identify categories with the same effect on the response. We propose a modification of the usual spike and slab prior for the regression coefficients by combining the spike at zero with a finite location mixture of spiky Normal components. This prior encourages sparsity by clustering the regression effects. We use the suggested method to analyse the annual income in Austria based on EU-SILC data from 2010.

Keywords: Sparse Modelling; Regression Model; Location Mixture; EU-SILC Data.

1 Introduction

In regression type models variables collected as potential covariates are often categorical. The usual strategy of modelling the effect of a categorical covariate by defining dummy variables for level effects can lead to a high-dimensional vector of regression coefficients. A sparse representation of the model can be achieved by fusing category levels with essentially the same effect on the response and/or by removing variables without any effect. To achieve effect fusion we propose a prior distribution on the regression coefficients, which is specified as a finite location mixture of spiky components and encourages sparsity by eventually emptying some of these components. As an example we analyse the annual personal income in Austria as a function of social and demographic characteristics using data from the Survey on Income and Living Conditions (SILC) in 2010. As possible covariates in our analysis we use `age, gender, federal state of residence, citizenship` and `level of education`.

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 Specification

We consider a standard linear regression model with Normal response y and p categorical covariates with levels $0, ..., c_j$ where $j = 1, ..., p$. We define 0 as the reference category and denote by X_{jh} the dummy variable corresponding to the h-th level of covariate j. Hence, the regression model is given as

$$y = \beta_0 + \sum_{j=1}^{p} \sum_{h=1}^{c_j} X_{jh}\beta_{jh} + \epsilon \tag{1}$$

where β_{jh}, $h = 1, \ldots, c_j$ is the effect of the h-th level of covariate j with respect to the reference category and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the error term.

3 Prior Specification and Posterior Inference

Our goal is to specify a prior which allows to identify clusters of level specific effects of each covariate. As finite mixture distributions are a convenient tool to achieve model based clustering we will use a mixture prior on the level effects. Generally, we specify the prior on the model parameters as

$$p(\boldsymbol{\beta}, \sigma_\epsilon^2) = p(\boldsymbol{\beta})p(\sigma_\epsilon^2)$$

and assume that the vectors of regression effects are independent between covariates. Hence, the prior on the regression coefficients has the structure

$$p(\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p) = p(\beta_0) \prod_{j=1}^{p} p(\boldsymbol{\beta}_j).$$

Elements of $\boldsymbol{\beta}_j$ are assumed independent conditionally on hyper-parameters and the prior of the level effects β_{jh} is specified as a finite mixture of Normal distributions. In contrast to the popular spike and slab priors, which are employed for selection of regression effects we use a location mixture of more than two components. Each of these components has a small variance, i.e. all components are spiky. Further, to allow identification of practically zero effects, one of the mixture components has its mean at zero. The proposed prior is specified hierarchically as

$$\begin{aligned}
p(\beta_{jh}) &= \sum_{l=0}^{L_j} \eta_{jl} f_{\mathcal{N}}(\beta_{jh}|\mu_{jl}, \psi_j) \\
\boldsymbol{\eta}_j &\sim Dir_{L_j+1}(e_0) \\
\mu_{j0} &= 0 \\
\mu_{jl} &\sim \mathcal{N}(m_{j0}, M_{j0}) \quad \text{for} \quad l = 1, ..., L_j
\end{aligned}$$

where $L_j + 1$ is the number of mixture components for covariate j and $Dir_{L_j+1}(e_0)$ is a symmetric Dirichlet distribution of dimension $L_j + 1$ with parameter e_0. L_j has to be chosen reasonably large to capture all relevant differences of the level effects, but not larger than c_j. We use $L_j = c_j$ in our application in Section 4.

As shown in Rousseau and Mengersen (2011) and Malsiner-Walli et al. (2016) empty components of a mixture distribution can be encouraged by setting the parameters of the Dirichlet prior e_0 to small values, e.g. $e_0 = 0.01$.

The parameters of the Normal mixture components are specified by taking an empirical approach based on the OLS estimates $\hat{\boldsymbol{\beta}}$ of the regression effects. For the specification of the covariate specific spike variances ψ_j we use the variance decomposition of a mixture model where the total heterogeneity can be decomposed into the variation of the component means around the global mean and the heterogeneity within a mixture component (Frühwirth-Schnatter, 2006)

$$\mathrm{Var}(\beta_{jh}) = \sum_{l=0}^{L_j} \eta_{jl}(\mu_{jl} - \bar{\mu}_j)^2 + \sum_{l=0}^{L_j} \eta_{jl}\psi_j,$$

where $\bar{\mu}_j = \sum_{l=0}^{L_j} \eta_{jl}\mu_{jl}$. We suggest to set the covariate specific variances to $\psi_j = 0.005\, V_j$ where V_j is the variation of the estimated level effects $\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jc_j}$,

$$V_j = \frac{1}{c_j - 1} \sum_{h=1}^{c_j} (\hat{\beta}_{jh} - \bar{\beta}_j)^2$$

and $\bar{\beta}_j = \frac{1}{c_j} \sum_{h=1}^{c_j} \hat{\beta}_{jh}$ is their mean. With decreasing variance ψ_j the mixture components become more spiky, which suggests that ψ_j can control the size of the selected model.

As hyperprior on the component means μ_{jl} we use a Normal distribution and choose its parameters also based on $\hat{\boldsymbol{\beta}}_j$. We set the mean m_{j0} of the Normal hyperpriors to $\bar{\beta}_j$ and the variance M_{j0} to the squared range $\big(\max_h \hat{\beta}_{jh} - \min_h \hat{\beta}_{jh}\big)^2$.

Figure 1 shows the prior distributions for the level effects of two covariates in our application, `citizenship` and `education level`. For each covariate one mixture component is centred at zero and the others at their starting values $\hat{\beta}_{jh}$. The overlapping of components suggests that some level effects could be allocated to the same cluster.

Finally, for the intercept β_0 we specify a Normal prior with large variance $\psi_0 = 10,000$ and for the error variance σ_ϵ^2 an inverse Gamma prior $\sigma_\epsilon^2 \sim \mathcal{G}^{-1}(s_0, S_0)$ with $s_0 = S_0 = 0$.

Bayesian inference is accomplished by sampling from the posterior distribution using MCMC methods. Sampling from mixture distributions is more

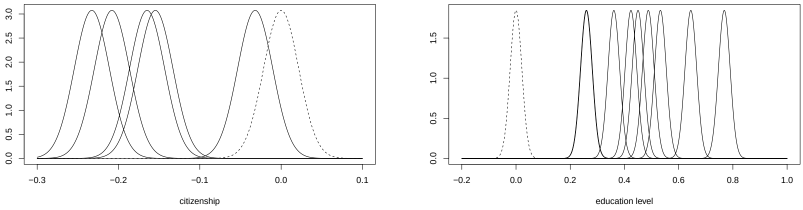


FIGURE 1. Finite mixture prior for level effects of covariate `citizenship` (left plot) and `education level` (right plot). One component is centred at zero (dashed), the others at $\hat{\beta}_{jh}$, $h = 1, \dots c_j$.

convenient using data augmentation by specifying latent allocation variables which indicate the component an object is assigned to. We therefore introduce a set of latent allocation variables $\boldsymbol{S}_j = (S_{j1}, ..., S_{jc_j})$ for each covariate j. S_{jh} takes values in $\{0, 1, ..., L_j\}$ and indicates the component to which the level specific effect β_{jh} is assigned. Posterior inference for the parameters $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \sigma_\epsilon^2, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{S})$ can be accomplished by sampling from the posterior distribution

$$p(\boldsymbol{\Theta}|\boldsymbol{y}) \propto p(\boldsymbol{\Theta})p(\boldsymbol{y}|\boldsymbol{\Theta})$$

where $p(\boldsymbol{y}|\boldsymbol{\Theta})$ is the likelihood of the regression model given in equation (1).

Sampling $\boldsymbol{\beta}$ from the corresponding multivariate posterior distribution conditional on $\boldsymbol{S}_j = (S_{j1}, ..., S_{jc_j})$ is straightforward as conditional on $S_{jh} = l$ the prior distribution for β_{jh} is the Normal distribution

$$p(\beta_{jh}|S_{jh} = l) \sim \mathcal{N}(\mu_{jl}, \psi_j).$$

4 Analysis of EU-SILC Data

We employ the sparse finite mixture prior described above in a linear regression analysis using Austrian EU-SILC (SILC = Survey on Income and Living Conditions) data from 2010. To model the log-transformed annual income of individuals we take into account only full-time employees (with a minimum annual income of EUR 2,000). As potential covariates we use `age` (linear and quadratic term), `gender`, `federal state`, `citizenship` and `highest education` achieved. After removing all data with missing values 3,909 observations remain in the data set.

We first fit a full model with regression effects for each covariate level using a flat prior. Posterior means of the regression coefficients are given in Table 1 (left).

To select a model with eventually fused categories we set the hyper - parameters as described in Section 3 and run MCMC for 50,000 iterations after a burn-in of 10,000.

As final model we select the model visited most often during MCMC and refit it with a flat prior on the regression effects. In this model the level effects of covariates `federal state`, `citizenship` and `education level` are fused to three, two and four effects, respectively. Posterior means of the regression effects are given in Table 1 (right) and Figure 2 compares posterior means and 95 % HPD intervals of the effects for covariates `citizenship` and `education level` in the full model and the refit of the selected model. Though none of the covariates is completely excluded from the model the number of regression coefficients (including intercept) is reduced from 26 to 10.

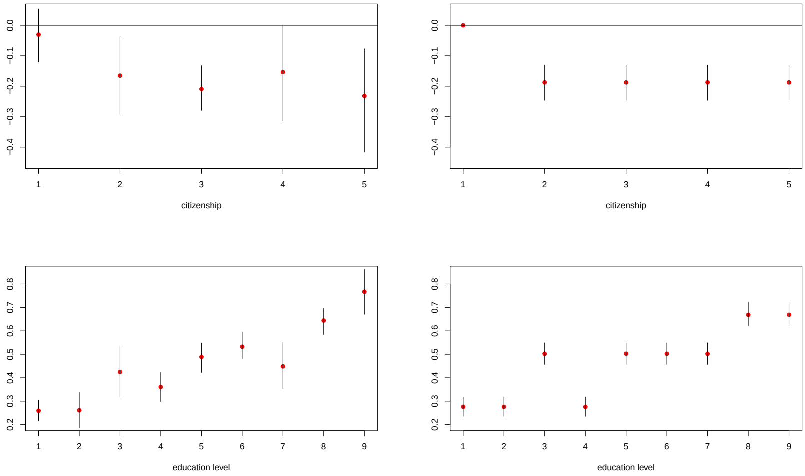


FIGURE 2. Posterior means and 95 % HPD intervals of covariate `citizenship` (upper panel) and `education level` (lower panel) for the full model (left) and the refitted selected model (right) under a flat prior.

The posterior mean of the error variance $\hat{\sigma}^2 = 0.181$ is negligibly higher compared to the full model ($\hat{\sigma}^2 = 0.180$), but BIC = 4493.06 is considerably smaller (full model: BIC = 4583.04).

Acknowledgments: We gratefully acknowledge financial support by the Austrian Science Fund FWF, project number P25850 'Sparse Bayesian modelling for categorical predictors'.

References

Frühwirth-Schnatter S. (2006). *Finite Mixture and Markov Switching Models.* Springer. New York.

Malsiner Walli G., Frühwirth-Schnatter S. and Grün B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, Vol.26, Issue 1, 303 – 324.

Rousseau J. and Mengersen K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society B*, Vol.73, Issue 5, 689 – 710.

TABLE 1. Posterior means for the full and the selected model

variable	full model	selected model
intercept	9.05	9.05
age (linear term)	0.62	0.62
age (squared term)	-0.43	-0.42
female	-0.22	-0.21
federal state (base: Upper Austria)		
Carinthia	-0.08	-0.06
Lower Austria	-0.03	-0.06
Burgenland	-0.08	-0.06
Salzburg	-0.04	-0.06
Styria	-0.10	-0.06
Tyrol	-0.05	-0.06
Vorarlberg	0.09	0.09
Vienna	-0.03	-0.06
citizenship (base: Austria)		
EU15/EFTA	-0.03	0.00
New EU10	-0.17	-0.18
Rest of Yugoslavia without Slovenia	-0.21	-0.18
Turkey	-0.15	-0.18
Others	-0.23	-0.18
Highest education achieved		
(base: max. secondary school degree)		
apprenticeship, trainee	0.26	0.28
master craftman's diploma	0.26	0.28
nurse's training school	0.42	0.51
other vocational school (medium level)	0.36	0.28
academic secondary school (upper level)	0.49	0.51
college for higher vocational education	0.53	0.51
vocational school for apprentices	0.45	0.51
university, academy, FH: first degree	0.65	0.67
university: doctoral studies	0.77	0.67

Modelling the covariance matrix for multivariate responses in hierarchical structures with correlated random effects

Adrian Quintero[1], Emmanuel Lesaffre[1]

[1] Leuven Biostatistics and Statistical Bioinformatics Centre, KU Leuven, Belgium

E-mail for correspondence: luisadrian.quinterosarmiento@kuleuven.be

Abstract: Multivariate regression methods generally assume a constant covariance matrix for the observations. In case a heteroscedastic model is needed, the parametric and nonparametric covariance regression approaches in the literature can be restrictive. We propose a multilevel regression model for the mean and covariance structure including random intercepts in both components and allowing for correlation between them. The implied conditional covariance function can be different across clusters as a result of the random effect in the variance structure. Furthermore, allowing for correlation between the random intercepts in the mean and covariance makes the model convenient for skewedly distributed responses. Parameter estimation is carried out via Gibbs sampling. The proposed model is applied to the RN4CAST data set to identify the variables that impact burnout of nurses in Belgium.

Keywords: Correlated responses; Covariance modelling; Gibbs sampling; Multivariate normal distribution.

1 Introduction

In multivariate regression analysis the mean vector is usually modelled assuming a constant covariance matrix for the observations. However, this assumption is often violated and understanding how the variance depends on the covariates may be of interest. In recent years, some approaches have been suggested to model simultaneously the mean and covariance structure. Specifically, Hoff and Niu (2012) proposed the rank-1 regression model $\mathbf{Y}_i = \mathbf{A}\mathbf{x}_i + \gamma_i \mathbf{B}\mathbf{x}_i + \boldsymbol{\varepsilon}_i$ where $\mathbf{A}$ is a matrix of regression coefficients, $\gamma_i \sim \mathcal{N}(0,1)$ is a latent variable that allows for extra-variation in the

response and $\mathbf{B}$ is a real matrix of factor loadings. The implied marginal covariance matrix for the response is a quadratic function of the covariates, i.e., $\boldsymbol{\Sigma}(\mathbf{x}_i) = \boldsymbol{\Psi} + \mathbf{B}\mathbf{x}_i\mathbf{x}_i^T\mathbf{B}^T$, where $\boldsymbol{\Psi}$ is the covariance matrix of $\boldsymbol{\varepsilon}_i$.
Li et al. (2013) proposed an extension of the covariance regression model suggested by Hoff and Niu (2012) to hierarchical data. The authors modelled three burnout outcomes in the European RN4CAST project (Sermeus et al., 2011) where nurses are clustered in nursing units and hospitals. Li et al. (2013) included a random intercept in the mean as well as a random intercept in the variance structure assuming independence between them. However, it could be the case that nursing units (or hospitals) with a high burnout level present a smaller variance for the outcomes, since burnout can be transmittable and all nurses might suffer homogeneously of psychological stress. Thus, a model that allows for correlation between the random effect in the mean and the variance structures appears necessary.

2 The RN4CAST dataset

The registered nurse forecasting (RN4CAST) project is a funded nurse workforce study conducted from 2009 to 2011 in 12 countries of Europe, see Sermeus et al. (2011) for details. It is of particular interest to identify the variables that have an effect on burnout of nurses in Belgium whilst explaining the covariance structure of the data based on the available covariates.
Burnout was measured using 22 items that were summarized in three variables: emotional exhaustion (EE), depersonalization (DP) and reduced personal accomplishment (PA). Several covariates were considered at hospital, nursing unit and nurse levels. There are in total 2492 female nurses in Belgium with full information for the analysis, grouped in 269 nursing units and 66 hospitals.

3 The correlated random effects model

The model is introduced here for a two-level structure assuming that the mean model and covariance structure depend on the same covariates. Let $\mathbf{Y}_{ij} \in \mathbb{R}^p$ be the multivariate response variable for subject j in cluster i, $\mathbf{x}_{ij} \in \mathbb{R}^q$ be the vector of explanatory variables and $\mathbf{B}$ the factor loadings matrix of size $p \times q$. The proposed model is

$$\mathbf{Y}_{ij} = \boldsymbol{\mu}_{\mathbf{x}_{ij}} + \mathbf{U}_i + \gamma_{ij}[\mathbf{B}\mathbf{x}_{ij} + \mathbf{U}_i^*] + \boldsymbol{\varepsilon}_{ij},\ i = 1, \ldots, I,\ j = 1, \ldots, n_i,$$

where $\boldsymbol{\mu}_{\mathbf{x}_{ij}} = \mathbf{A}\mathbf{x}_{ij}$ is the expectation $\mathrm{E}(\mathbf{Y}_{ij}|\mathbf{x}_{ij})$. The latent factor γ_{ij} has a standard normal distribution while the random error $\boldsymbol{\varepsilon}_{ij}$ follows a multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi})$. Similarly, it is assumed that $\mathbf{U}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi}_u)$ and $\mathbf{U}_i^* \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Psi}_{u^*})$. The random effect in the mean is allowed to be correlated with the random effect in the variance structure as $\text{Cov}(\mathbf{U}_i, \mathbf{U}_i^*) = \text{E}(\mathbf{U}_i\mathbf{U}_i^{*T}) = \boldsymbol{\Psi}_{uu^*}$.

It is also assumed that the random error, the latent factor and the random effects joint vector $\mathbf{U}_i^b = (\mathbf{U}_i, \mathbf{U}_i^*)$ are mutually independent. The distribution of the joint vector is $\mathbf{U}_i^b \sim \mathcal{N}_{2p}(\mathbf{0}, \boldsymbol{\Psi}^b)$. This model is an extension of the model proposed by Li et al. (2013) which assumes $\boldsymbol{\Psi}_{uu^*} = \mathbf{0}$.

The implied covariance matrix for $\mathbf{Y}_{ij}$ given the model is a quadratic function of the explanatory variables, i.e.,

$$\boldsymbol{\Sigma}(\mathbf{x}_{ij}) = \boldsymbol{\Psi} + \boldsymbol{\Psi}_u + \boldsymbol{\Psi}_{u^*} + \mathbf{B}\mathbf{x}_{ij}\mathbf{x}_{ij}^T\mathbf{B}^T. \tag{1}$$

Hence, the model has two possible solutions since $\boldsymbol{\Sigma}(\mathbf{x}_{ij})$ is the same given γ, $\mathbf{B}$, $\mathbf{U}^*$ and given $-\gamma$, $-\mathbf{B}$ and $-\mathbf{U}^*$. Both alternatives lead to the same interpretation of the parameters in the variance function. The identifiability of $\mathbf{B}$ up to the sign can be shown exactly as for the rank-1 model in Hoff and Niu (2012), given sufficient variability in the covariates.

The distribution of the response variable under this model is the product of two normal densities added to a multivariate normal distribution. The skewness of the l-th component of $\mathbf{Y}$ is equal to

$$6\mathbf{b}^{(l)}\mathbf{x}_{ij}\sigma_{uu_l^*}/\sigma^3_{y_{ijl}},\ l = 1, \ldots, p,$$

where $\sigma_{uu_l^*}$ is the l-th diagonal element of $\boldsymbol{\Psi}_{uu^*}$, $\mathbf{b}^{(l)}$ corresponds to the l-th row of $\mathbf{B}$ and $\sigma^2_{y_{ijl}}$ is the l-th diagonal element of the marginal variance in (1). Hence, allowing for correlation between the random effects can be advantageous in cases with skewedly distributed response variables.

3.1 Distribution of the response given the random effects

The conditional distribution of the response variable given the random effects is

$$\mathbf{Y}_{ij}|\mathbf{U}_i, \mathbf{U}_i^* \sim \mathcal{N}_p(\boldsymbol{\mu}_{\mathbf{x}_{ij}} + \mathbf{U}_i, \boldsymbol{\Psi} + (\mathbf{B}\mathbf{x}_{ij} + \mathbf{U}_i^*)(\mathbf{B}\mathbf{x}_{ij} + \mathbf{U}_i^*)^T).$$

Thus, the conditional covariance matrix of the response is different in each cluster depending on the random effect $\mathbf{U}_i^*$. To illustrate the implications of this conditional variance, let us take the simple case of $p = 2$ with only one explanatory variable x and the matrix $\mathbf{B}$ of columns $\mathbf{b}_1 = (1, 1)^T$ and $\mathbf{b}_2 = (-2, -2)^T$. Figure 1 presents the conditional variance σ_1^2 and covariance σ_{12} in function of the covariate for the two cases $\mathbf{U}_i^* = (-5, 0)^T$ and $\mathbf{U}_i^* = (5, 0)^T$ assuming that $\boldsymbol{\Psi}$ is the identity matrix. In the first case, the expected variance of Y_1 is an increasing function of x for the observations in the cluster, while the expected covariance between Y_1 and Y_2 is decreasing. The effect of the covariate on the expected variance and covariance terms is the opposite in the second cluster compared to the first case.

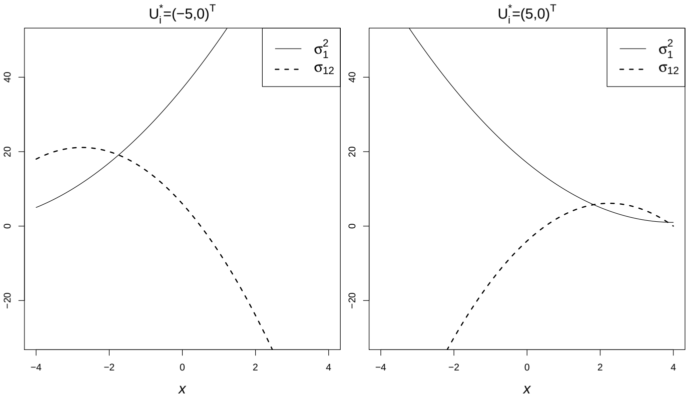


FIGURE 1. Effect of including the random effect $\mathbf{U}^*$. The conditional variance of the response variable σ_1^2 and covariance σ_{12} is presented for a cluster with $\mathbf{U}_i^* = (-5,0)^T$ (left panel) and another cluster with $\mathbf{U}_i^* = (5,0)^T$ (right panel).

3.2 Variability given the random effect in the mean

Under the uncorrelated model ($\boldsymbol{\Psi}_{uu^*} = \mathbf{0}$), the covariance matrix of the response depends only on the random effect $\mathbf{U}_i^*$. On the other hand, when assuming the correlated model, the variance depends indirectly on $\mathbf{U}_i$ due to its correlation with the random effect in the covariance structure.

Let us define $\mathbf{U}_i^{bl} = (\mathbf{U}_i, U_{il}^*)$ as the vector of size $p+1$ that contains the random effect for the mean and the l-th component of $\mathbf{U}_i^*$, for $l = 1,\dots,p$. The covariance matrix of this vector is $\mathbf{K}_l = \text{Var}(\mathbf{U}_i^{bl})$ which can be obtained from the corresponding elements of $\boldsymbol{\Psi}^b$ and let us define $\mathbf{Q}^l = \mathbf{K}_l^{-1} = [q_{ij}^l]$ as its inverse. Thus, the conditional expectation of the random effect in the factor loadings given the random effect in the mean part is $\text{E}(\mathbf{U}_i^*|\mathbf{U}_i) = \mathbf{Q}\mathbf{U}_i$, where

$$\mathbf{Q} = -\left[diag\left(q^1_{(p+1)(p+1)},\dots,q^p_{(p+1)(p+1)}\right)\right]^{-1}\begin{pmatrix} q^1_{(p+1)1} & \cdots & q^1_{(p+1)p} \\ \vdots & \vdots & \ddots & \vdots \\ q^p_{(p+1)1} & \cdots & q^p_{(p+1)p}\end{pmatrix},$$

and the conditional variance of the response variable is

$$\text{Var}(\mathbf{Y}_{ij}|\mathbf{U}_i) = \boldsymbol{\Psi} + \boldsymbol{\Psi}_{\mathbf{U}^*|\mathbf{U}} + \mathbf{B}\mathbf{x}_{ij}\mathbf{x}_{ij}^T\mathbf{B}^T + \mathbf{B}\mathbf{x}_{ij}\mathbf{U}_i^T\mathbf{Q}^T + \mathbf{Q}\mathbf{U}_i\mathbf{x}_{ij}^T\mathbf{B}^T, \quad (2)$$

where $\boldsymbol{\Psi}_{\mathbf{U}^*|\mathbf{U}} = \text{Var}(\mathbf{U}_i^*|\mathbf{U}_i)$ is a constant matrix. Thus, the implied conditional variance of the response depends linearly on the random effect in the mean. It is in contrast with the uncorrelated model, where this random effect has no effect at all.

TABLE 1. Estimates of the correlated model for the burnout measures.

Parameter	Mean	2.5%	97.5%	Parameter	Mean	2.5%	97.5%
$\beta_{expe.n}[EE]$	-0.057	-0.099	-0.017	$\beta_{envi.h}[EE]$	-0.134	-0.236	-0.032
$\beta_{expe.n}[DP]$	-0.24	-0.298	-0.181	$\beta_{envi.h}[DP]$	-0.161	-0.263	-0.06
$\beta_{expe.n}[PA]$	-0.015	-0.053	0.024	$\beta_{envi.h}[PA]$	-0.115	-0.182	-0.048
$\beta_{size.u}[EE]$	-0.042	-0.104	0.021	$\lambda_0[EE]$	0.341	0.219	0.465
$\beta_{size.u}[DP]$	-0.081	-0.151	-0.012	$\lambda_0[DP]$	0.359	0.174	0.54
$\beta_{size.u}[PA]$	0.017	-0.03	0.065	$\lambda_0[PA]$	0.563	0.422	0.726
$\beta_{envi.u}[EE]$	-0.184	-0.236	-0.13	$\lambda_{expe.n}[EE]$	0.034	-0.029	0.099
$\beta_{envi.u}[DP]$	-0.239	-0.303	-0.176	$\lambda_{expe.n}[DP]$	0.076	-0.011	0.163
$\beta_{envi.u}[PA]$	-0.093	-0.136	-0.049	$\lambda_{expe.n}[PA]$	0.192	0.114	0.266

4 Modelling the burnout outcomes

The uncorrelated model proposed by Li et al. (2013) (assuming $\mathbf{\Psi}_{uu^*} = \mathbf{0}$) and the correlated model introduced here were fit to the RN4CAST data set. The burnout outcomes were modelled considering all the covariates in the three-level hierarchical structure with Belgian nurses clustered in nursing units and hospitals. The models were estimated using Gibbs sampling with normal vague priors for the regression coefficients and vague inverse Wishart density for each covariance matrix. We compared the two models based on the deviance information criterion (DIC). For the uncorrelated model, the DIC is 44236 ($p_D = 1921$), whereas for the correlated model DIC $= 44160$ ($p_D = 1833$), indicating a strong preference for the latter.
The estimated correlation matrices between the random effects in the mean and variance structure at hospital and nursing unit levels with the correlated model are respectively

$$\mathrm{Cor}(\mathbf{U}_h.\mathbf{U}_h^*) = \begin{pmatrix} -0.82 & -0.58 & -0.77 \\ -0.77 & -0.65 & -0.70 \\ -0.75 & -0.57 & -0.78 \end{pmatrix}, \quad \mathrm{Cor}(\mathbf{U}_u.\mathbf{U}_u^*) = \begin{pmatrix} -0.58 & -0.21 & -0.52 \\ -0.31 & -0.15 & -0.51 \\ -0.38 & -0.01 & -0.78 \end{pmatrix}.$$

Thus, most correlations between the random effects are strongly negative, corroborating the necessity of a model that accounts for it. The Bayesian estimates of the model are presented in Table 1, indicating that a better work environment tends to reduce the three burnout measures. Likewise, a higher number of nurses in the nursing unit tends to diminish the level of depersonalization and nurses with higher experience report lower emotional exhaustion and depersonalization levels.
The estimated variance function (1) for personal accomplishment is presented in Figure 2, as well as the conditional variance (2) given the random effect in the mean at hospital level. It can be seen that nurses with longer experience present higher variability for the personal accomplishment measure compared to nurses with short experience. Additionally, hospitals (it is also the case for nursing units) with high levels of personal accomplishment present smaller variance for this burnout outcome.

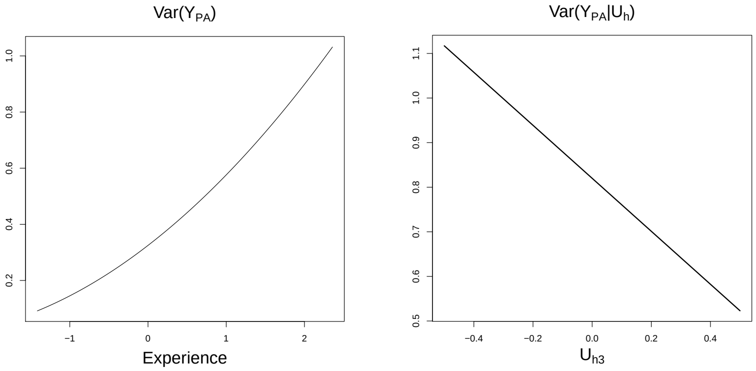


FIGURE 2. Marginal variance of PA and conditional variance given $\mathbf{U}_h$.

5 Conclusions

The proposed model that allows for correlation between random effects is useful to model skewed distributed responses. This alternative appears to be more appropriate for the RN4CAST dataset based on the DIC. The random effects in the mean showed to be strongly negatively correlated with the random effects in the variance structure, pointing out the necessity of the proposed model. Regarding the variance part, we found that nurses with longer experience present generally higher variability for personal accomplishment and hospitals and nursing units with high levels of this burnout measure suffer more homogeneously of that inferiority complex.

Acknowledgments: Special thanks to Luk Bruyneel and Walter Sermeus for their input in this paper.

References

Hoff, P. and Niu, X. (2012). Analysis of Longitudinal Data. *Statistica Sinica*, **22**, 729–753.

Li, B., Lesaffre E. and Bruyneel, L. (2013). A multivariate multilevel gaussian model with a mixed effects structure in the mean and covariance part. *Statistics in Medicine*, **33**, 1877,–1899.

Sermeus W., Aiken L., Van den Heede K., Rafferty A., Griffiths P., Moreno-Casbas M., Busse R., Lindqvist R., Scott A., Bruyneel L., Brzostek T., Kinnunen J., Schubert M., Schoonhoven L., Zikos D. and RN4CAST consortium. (2011). Nurse forecasting in Europe (RN4CAST): rationale, design and methodology. *BMC Nursing*, **10**.

Nearest Neighbor Imputation for Categorical Data by Weighting of Attributes

Shahla Ramzan[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: shahla.ramzan@stat.uni-muenchen.de

Abstract: Missing values are a common phenomenon in applied research. While various imputation methods are available for metrically scaled variables, methods for categorical data are scarce. An imputation method that has been shown to work well for high dimensional metrically scaled variables is imputation by nearest neighbor methods. In this paper, we extend the weighted nearest neighbors approach to impute missing values to the case of categorical variables. The proposed method explicitly uses the information on association among attributes. A version of L_q-distance based on dummy variables is proposed. The performance of different imputation methods is compared in terms of the proportion of falsely imputed values. Simulation results show that the weighting of attributes yields smaller imputation errors than existing approaches.

Keywords: Categorical data; Weighted nearest neighbors; Kernel function.

1 Introduction

Categorical data often come with missing values but approaches to the imputation of categorical variables are scarce. The k-nearest neighbors method originally developed for continuous data (Troyanskaya et al., 2001), cannot by employed to non-metric data such as unordered categorical or ordinal data (Schwendler, 2012). Some existing methods to impute attributes are based on the mode or weighted mode of k-nearest neighbors.

For categorical data one has to use specific distances or similarity measures. Distance measures for categorical data are typically based on an $R \times C$ contingency table, where R and C are the number of values that the two attributes can assume. Some commonly used distance measures include the Simple Matching Coefficient (SMC), Cohen's κ , or the Manhattan or L_1 distance.

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 Euclidean or variants of the Minkowski distance give an equal importance to all the variables in the data matrix when computing the distance. But for a larger number of variables, the equal weighting ignores the complex structure of correlation/association among these variables. Then it is helpful to utilize this information to obtain better distance measures. We propose a weighted distance that explicitly takes the association among covariates into account. More specifically, highly associated covariates are given higher weights forcing them to contribute more strongly to the computation of the distance than weakly associated covariates.

2 Weighted distance measure for categorical data

Let data be collected in $\boldsymbol{Z}_{(n \times p)} = (Z_{is})$ and $\boldsymbol{O}_{(n \times p)} = (o_{is})$, where Z_{is} is the i^{th} observation on the s^{th} attribute, and $o_{is} = 1$ if Z_{is} was observed, $o_{is} = 0$ if it was missing. The categorical observations Z_{ij} in the data matrix $\boldsymbol{Z}$, can assume values $c_j \in \{1, \ldots, k_j\}, j = 1, \ldots, p$, where k_j is the number of categories that the j^{th} attribute can take. For the computation of distances, the categorical variables are transformed into binary variables. Thus the observation Z_{ij} becomes a vector, $\boldsymbol{z}_{ij}^T = (z_{ij1}, \ldots, z_{ijk_j})$ with $z_{ij1} = 1$ if $Z_{ij} = r$. The dummy vectors $\boldsymbol{z}_{ij}^T$ for a nominal variable with four categories can be written as

category	z_{ij1}	z_{ij2}	z_{ij3}	z_{ij4}
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

The observation vectors in the transformed data matrix can be written as $\boldsymbol{z}_i^T = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{ip}^T)$. Let now z_{is} be a missing entry in the i-th observation, that is $O_{is} = 0$. Then the distance between the i-th and the j-th observation is defined by

$$d_{q,C}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \left(\frac{1}{n_{ij}} \sum_{l=1}^{p} \sum_{c=1}^{k_l} |z_{ilc} - z_{jlc}|^q I_{(o_{il}=1)} I_{(o_{jl}=1)} C(\delta_{sl}) \right)^{1/q}, \quad (1)$$

where $n_{ij} = \sum_{l=1}^{p} I_{(o_{is}=1)} . I_{(o_{js}=1)}$ denotes the number of valid components in the computation of distances. The crucial part in the definition of the distance is the weight $C(\delta_{sl})$. $C(.)$ is a convex function defined on the interval $[-1, 1]$ that transforms the measures of association into weights and δ_{sl} is a measure of association between attributes s and l. It is worth noting that the distance is specific to the s^{th} attribute, which is to be imputed.

For $C(.)$ we use the power function $C(\delta_{sl}) = |\delta_{sl}|^m$. So the attributes that have a higher association with the s^{th} attribute are contributing more to the distance and vice versa. The higher the value of association, the more it contributes to the computation of the distance. Note also that only the *available* pairs i.e., $I_{(o_{il}=1)}I_{(o_{rl}=1)}$, are used in the computation of the distance. We use Cramer's V, which is based on the χ^2-statistic, to measure the association among attributes.

3 Weighted NN Imputation

Let z_{is} be a missing value in $\boldsymbol{Z}_{(n\times p)}$ matrix. Then one finds the k nearest neighbor observation vectors $\boldsymbol{z}_{(k)}$ based on the distances defined in equation (1)

$$\boldsymbol{z}_{(1)}, \dots, \boldsymbol{z}_{(k)} \quad \text{with} \quad d(\boldsymbol{z}_i, \boldsymbol{z}_{(1)}) \leq \cdots \leq d(\boldsymbol{z}_i, \boldsymbol{z}_{(k)}).$$

The weighted imputation estimate is obtained by the relative probabilities $(\hat{\pi}_c)$ for each class,

$$\hat{\pi}_c = \hat{z}_{.jc} = \sum_{l=1}^{k} w(\boldsymbol{z}_i, \boldsymbol{z}_{(j)}) z_{(l)jc},$$

where $c = 1, \dots, k_j$. One may draw at random from the distribution. The weights are defined by

$$w(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{k(d(\boldsymbol{z}_i, \boldsymbol{z}_j)/\lambda)}{\sum_{l=1}^{k} k(d(\boldsymbol{z}_i, \boldsymbol{z}_j)/\lambda)}, \tag{2}$$

where$k(.)$ is a kernel functions (tricube, Gaussian etc.) and λ is tuning parameter. If one uses all the available neighbors that is $k = \tilde{n}$, then λ is the only and crucial tuning parameter.

The imputed estimate is the value of $c \in \{1, \dots, k_j\}$ with highest value of $\hat{\pi}$. In other words, the weighted imputation estimate of a categorical missing value z_{is} is

$$\hat{z}_{is} = \arg\max{}_{c=1}^{k_j} \hat{\pi}_c,$$

One may obtain more than one values with the same probability $(\hat{\pi})$. In this case, only one value at random is selected.

4 Simulation Studies

This section includes the application of the proposed method using simulated data to check if the suggested distance measure contributes to better imputation or not.

We generated 200 samples of size $n = 100$ and $p = 10, 50$ predictors drawn from a multivariate normal distribution with $N(\mathbf{0}, \mathbf{\Sigma})$. The correlation matrix $\mathbf{\Sigma}$ has an autoregressive type of order 1 with $\rho = 0.9$. We construct categories from the continuous data by setting cut points. For example, for four categories $n_{cat} = 4$, with equal probability (π_c) for all the predictors, the quartiles Q_1, Q_2, Q_3 are used as cut points, where Q_1, Q_2, Q_3 are the usual lower quartile, median and upper quartile respectively, which divide the data into four equal parts. So in this case, $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$. In general, to create c categories of a variable one needs $c - 1$ cut points and each category has $\pi_c = 1/c$. In each sample, 10%, 20%, 30% of the total values were replaced by missing values completely at random.

The missing values are imputed using Mode imputation, random forests (RF) and proposed weighted nearest imputation methods. In proposed method ($wNNSel_{cat}$), the distance (1) is computed using $q = 1, 2$. The tuning parameters λ and m are estimated by cross validation procedure. In the data matrix some values, for example 5% of the total available values, are removed randomly and then imputed using specific values of λ and m. The pair of values that provides smaller imputation error is chosen as λ_{opt} and m_{opt} for that particular data matrix.

To compare the performance of different imputation methods, the proportion of falsely imputed categories (PFC) is computed for each imputation method.

$$\text{PFC} = \frac{1}{n^*} \sum_{z_{is}:o_{is}=0} I_{(z_{is} \neq \hat{z}_{is})},$$

where n^* is the number of missing values in the data matrix, z_{is} is the true value and $\hat{z}_{is}$ is the imputed value.

When n_{cat} is same for all the attributes

In our first simulation setting, the number of categories (n_{cat}) of all the attributes is same but the categories within each attribute $(c = 1, \cdots, n_{cat})$ may have an unequal chances of their occurrence (π_c). The purpose is to investigate whether π_c do have any effect on the imputation results.

We use $q = 1, 2$ in the distance calculation of $wNNSel_{cat}$ method to get L_1 and L_2 metrics. The tuning parameters are estimated by cross validation and these optimal values, λ_{opt} and m_{opt}, are used to estimate the final imputed values.

The method by Schwendler (2012) is used as benchmark only in this simulation setting where all the attribute have an equal number of categories. In this method, k neighbors are chosen based on some distance measure and their weighted average is used as an imputation estimate. But this method requires the selection of suitable value of k (the number of nearest neighbours) and a distance metric. We use Cohen, Pearson corrected coefficient

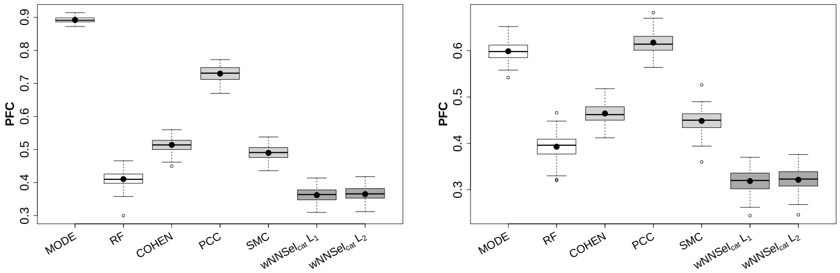


FIGURE 1. Boxplots of proportion of falsely imputed categories for MCAR missing pattern with $n_{cat} = 4$ and 10% missing data. Left panel shows when probability of occurrence of each category (π_c) is same and right panel for probability of occurrence of each category (π_c) is not same.

(PCC) and simple matching coefficient (SMC) distances to compare the performance. Cross-validation is used to find the suitable value of k and the value with smaller PFC is selected.
The resulting average PFC for $n_{cat} = 4$ with 10% missing values are shown in Figure 1. Clearly, the weighted method ($wNNSel_{cat}$) for imputation provides smaller imputation error whether the categories have the same probability of occurrence (Fig.1: left panel) or not (Fig.1: right panel). It is also worth noting that the difference of L_1 and L_2 metric is not noticeable.

When n_{cat} is different for the attributes

The second simulation setting investigates a more general case when neither the number of categories n_{cat} nor the the probability of occurrence of categories (π_c)is same. we use $n_{cat} = 3, 4$ and $n_{cat} = 3, 4, 5$ for the variables.

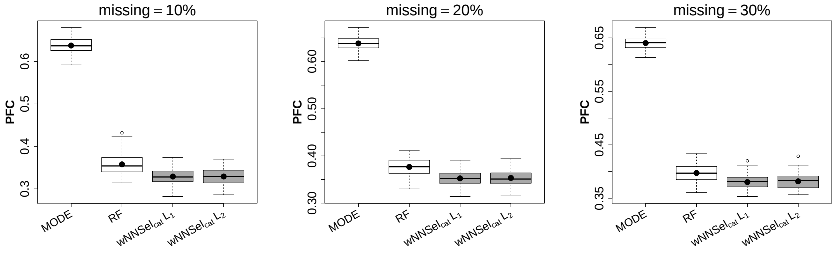


FIGURE 2. Boxplots of proportion of falsely imputed categories for MCAR missing pattern. $S = 200$ samples $n = 100$, $p = 50$ were drawn from multivariate normal distribution using autoregressive correlation structures to form $n_{cat} = 3, 4$ categories.

The results only for $n = 100$, $p = 50$, $n_{cat} = 3, 4$ with 10%, 20%, 30% missing values are shown in Figure 2. In this data matrix, some of the variables chosen randomly are divided into $n_{cat} = 3$ and the rest are divided into $n_{cat} = 4$ categories. So the number of values that a variable can assume are different for all the variables in a 100×50 data matrix. The cut points are also chosen such that the probability of occurrence of each category (π_c) is not the same. Figure 2 shows that mode imputation perform poor as compared to random forest and $(wNNSel_{cat})$ method. The random forest is a good competitor but nevertheless $(wNNSel_{cat})$ provides smallest imputation errors. The choice of L_1 or L_2 metric do not have any significant effect on the results.

References

Schwender, H. (2012). Imputing missing genotypes with weighted k nearest neighbors. *Journal of Toxicology and Environmental Health*, Part A, 75(8-10), 438 – 446.

Stekhoven, D. J., and Bühlmann, P. (2012). MissForest: non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28 (1)**, 112 – 118.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17 (6)**, 520 – 525.

Spatio-temporal adaptive penalized splines with application to Neuroscience

María Xosé Rodríguez-Álvarez[1], María Durbán[2], Dae-Jin Lee[3], Paul H.C. Eilers[4], Francisco Gonzalez[5]

[1] Department of Statistics and Operations Research and Biomedical Research Centre (CINBIO), University of Vigo, Vigo, Spain
[2] Department of Statistics, University Carlos III of Madrid, Leganés, Spain
[3] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[4] Erasmus University Medical Center, Rotterdam, The Netherlands
[5] Department of Surgery and CIMUS, University of Santiago de Compostela, Santiago de Compostela, Spain.

E-mail for correspondence: mxrodriguez@uvigo.es

Abstract: Data analysed here derive from experiments conducted to study neurons' activity in the visual cortex of behaving monkeys. We consider a spatio-temporal adaptive penalized spline (P-spline) approach for modelling the firing rate of visual neurons. To the best of our knowledge, this is the first attempt in the statistical literature for locally adaptive smoothing in three dimensions. Estimation is based on the *Separation of Overlapping Penalties* (SOP) algorithm, which provides the stability and speed we look for.

Keywords: Visual neuron; Visual receptive field; Adaptive Smoothing; P-splines; SOP algorithm

1 Visual receptive fields

Electrophysiology studies record the electrical activity produced by neurons. They allow the study of the association between sensory stimuli and neural response in any part of the brain. Neurons produce sudden changes in their membrane potential known as 'spikes', that can be recorded using microelectrodes. The analysis of the frequency of spike discharges provides insights on how the neurons and the nervous system work.

Visual receptive fields (RFs) are small areas of the visual field that a particular visual neuron 'sees'. Reverse cross-correlation is a receptive field mapping technique used for studying how visual neurons process signals from different positions in their receptive field. From the neuron responses (spikes) we can infer the spatio-temporal properties of the RFs (i.e., when and where a sensory stimulus produces a response). A detailed explanation on how the reverse cross-correlation technique was used in the experiments analyzed here can be found elsewhere (Rodríguez-Álvarez et al., 2012). Schematically, the subject (a monkey) was viewing two monitors (one for each eye) with a fixation target. Within a square area a bright or dark spot was flashed at different positions in a pseudorandom manner. Neuron spikes were recorded while the stimulus was delivered. When a spike was produced, the stimulus position at several pre-spike times was read. As a result, a set of numerical matrices (one for each pre-spike time) containing the number (counts) of times the stimulus was at that given position when a spike occurred is obtained. The graphical representation of each of these matrices is called receptive field map (RFmap), and can be regarded as a representation of the firing rate of the neuron.

2 Three-dimensional adaptive P-spline

For each neuron, the reverse cross-correlation technique provides a dataset consisting of a series of 16 matrices of dimension 16 × 16, each matrix corresponding to the different pre-spike times considered (between −20 to −320 milliseconds). We adopted a Poisson model which expresses the neuron response (i.e., number of spikes) as a smooth function of both space and time

$$log\left(E\left[y \mid r, c, t\right]\right) = log\left(n_{rc}\lambda_{rct}\right) = log\left(n_{rc}\right) + f\left(r, c, t\right), \tag{1}$$

where r indicates the row of the matrix, c the column $(r, c = 1, \ldots, 16)$, and t the pre-spike time $(t = -20, \ldots, -320)$. n_{rc} denotes the number of stimulus presentations on each particular grid position (the offset) and λ_{rct} is the intensity parameter (or firing rate). The smooth function $f(\cdot,\cdot,\cdot)$ was represented by the tensor product of three univariate B-spline basis (Eilers and Marx, 2003), i.e., $f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{t}) = \left(\boldsymbol{B}_3^{(16\times c_3)} \otimes \boldsymbol{B}_2^{(16\times c_2)} \otimes \boldsymbol{B}_1^{(16\times c_1)}\right)\boldsymbol{\theta}$, where $\otimes$ denotes the Kronecker product.

In order to avoid over-fitting, the previous model can be estimated by penalized-likelihood methods (Eilers and Marx, 2003). In the absence of locally adaptive smoothness, the anisotropic penalty matrix is defined as

$$\lambda_1\left(\mathbf{I}_{c_3} \otimes \mathbf{I}_{c_2} \otimes \boldsymbol{D}_1^{\mathrm{T}}\boldsymbol{D}_1\right) + \lambda_2\left(\mathbf{I}_{c_3} \otimes \boldsymbol{D}_2^{\mathrm{T}}\boldsymbol{D}_2 \otimes \mathbf{I}_{c_1}\right) + \lambda_3\left(\boldsymbol{D}_3^{\mathrm{T}}\boldsymbol{D}_3 \otimes \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_1}\right), \tag{2}$$

where λ_1, λ_2 and λ_3 are the smoothing parameters, and $\boldsymbol{D}_d$ $(d = 1,\ 2,\ 3)$ are difference matrices of possibly different order q_d.

In adaptive P-spline smoothing (see, e.g., Rodríguez-Álvarez et al., 2015a) each λ_d in (2) is replaced by a vector of smoothing parameters $\boldsymbol{\lambda}_d$, where each component is associated with one coefficient difference (along the d-direction). However, this approach would imply as many smoothing parameters as coefficient differences, which could lead to under-smoothing and unstable computations. To reduce the dimension, $\boldsymbol{\lambda}_d$ is modelled by means of B-splines, i.e.,

$$\begin{aligned}
\boldsymbol{\lambda}_1 &= \left(\boldsymbol{C}_{11}^{((c_1-q_1)\times p_{11})} \otimes \boldsymbol{C}_{12}^{(c_2\times p_{12})} \otimes \boldsymbol{C}_{13}^{(c_3\times p_{13})}\right) \boldsymbol{\phi}_1 = \boldsymbol{C}_1\boldsymbol{\phi}_1,\\
\boldsymbol{\lambda}_2 &= \left(\boldsymbol{C}_{21}^{(c_1\times p_{21})} \otimes \boldsymbol{C}_{22}^{((c_2-q_2)\times p_{22})} \otimes \boldsymbol{C}_{23}^{(c_3-p_{23})}\right) \boldsymbol{\phi}_2 = \boldsymbol{C}_2\boldsymbol{\phi}_2,\\
\boldsymbol{\lambda}_3 &= \left(\boldsymbol{C}_{31}^{(c_1\times p_{31})} \otimes \boldsymbol{C}_{32}^{(c_2\times p_{32})} \otimes \boldsymbol{C}_{33}^{((c_3-q_3)\times p_{33})}\right) \boldsymbol{\phi}_3 = \boldsymbol{C}_3\boldsymbol{\phi}_3,
\end{aligned}$$

where $\boldsymbol{C}_{ij}$ $(i, j = 1, 2, 3)$ are B-spline regression matrices, with less columns than rows to ensure that the dimension is in fact reduced. The *adaptive* penalty matrix in three dimensions can be then expressed as

$$\begin{aligned}
&\sum_{s=1}^{p_{11}p_{12}p_{13}} \phi_{1s} \left(\mathbf{I}_{c_3} \otimes \mathbf{I}_{c_2} \otimes \boldsymbol{D}_1\right)^{\mathrm{T}} diag\left(\boldsymbol{c}_{1,s}\right) \left(\mathbf{I}_{c_3} \otimes \mathbf{I}_{c_2} \otimes \boldsymbol{D}_1\right) \qquad +\\
&\sum_{u=1}^{p_{21}p_{22}p_{23}} \phi_{2u} \left(\mathbf{I}_{c_3} \otimes \boldsymbol{D}_2 \otimes \mathbf{I}_{c_1}\right)^{\mathrm{T}} diag\left(\boldsymbol{c}_{2,u}\right) \left(\mathbf{I}_{c_3} \otimes \boldsymbol{D}_2 \otimes \mathbf{I}_{c_1}\right) \qquad +\\
&\sum_{v=1}^{p_{31}p_{32}p_{33}} \phi_{3v} \left(\boldsymbol{D}_3 \otimes \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_1}\right)^{\mathrm{T}} diag\left(\boldsymbol{c}_{3,v}\right) \left(\boldsymbol{D}_3 \otimes \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_1}\right),
\end{aligned} \tag{3}$$

where $\boldsymbol{c}_{d,l}$ denotes the column l of $\boldsymbol{C}_d$.
Estimation of the three-dimensional P-spline model for Poisson data (1) subject to the adaptive penalty defined in (3) can be based on its mixed-model representation. Restricted maximum likelihood (REML) estimates of the variance components (or smoothing parameters) are obtained by means of the *Separation of Overlapping Penalties* (SOP) algorithm, recently proposed by Rodríguez-Álvarez et al. (2015a,b). It should be noted that the reformulation of model (1) as a mixed model does not gives rise to a diagonal precision matrix, and thus, some of the computational advantages of SOP are lost. Nevertheless, even in this case, the algorithm provides reasonable computing times. Besides, Generalized Linear Array Models (GLAM, Currie et al., 2006) can be used to compute the inner products involved in the mixed model equations as well as the penalty matrices given in (3), thus improving the speed of the estimation algorithm.

3 Results

For illustration purposes, we show the results for a single visual neuron from area V1 (primary visual cortical area). Model (1) was estimated with and without assuming locally adaptive smoothness by means of the SOP algorithm and GLAM . In both cases, we used second-order differences ($q_d = 2$) and marginal B-splines bases of dimension $c_d = 7$. For the adaptive approach, we chose $p_{ij} = 4$ (i, j = 1,2,3), yielding a total of 192 (3×4^3) smoothing parameters (or variance components). Figure 1 shows the observed and estimated series of smooth RFmaps for several pre-spike times using both approaches. As it can be seen, both analyses show a central area of high values that represents the visual RF of the neuron, which is in concordance with the raw data. However, there are two major differences: whereas the non adaptive approach seems to indicate that the time between sensory stimulus and response spans from 20 to 100 ms, the adaptive method reduces this time span from 40 to 100 ms. Also the adaptive approach shows a sharper increase and a larger estimate of the firing rate than the non-adaptive approach (see also Figure 2). In terms of computational effort, in the absence of adaptive smoothness the algorithm needed about 14 seconds, whereas the complexity afforded by the adaptive approach increased the computing time to 133 seconds.

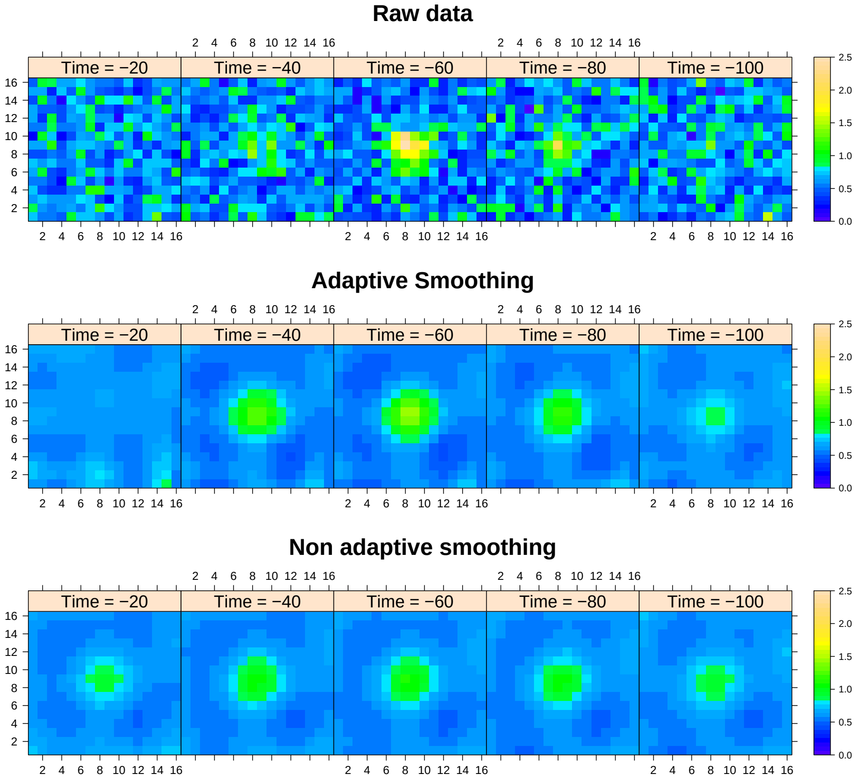


FIGURE 1. Level plot of the observed and smoothed firing rates of the RFmap with and without locally adaptive smoothness.

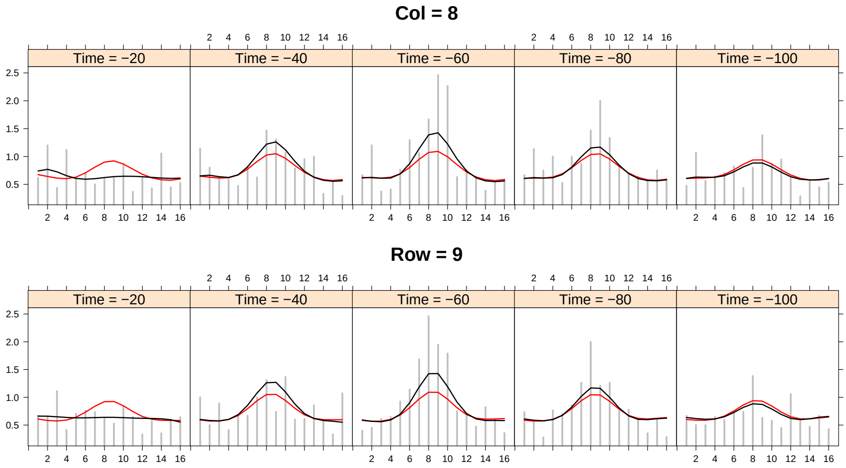


FIGURE 2. Observed and smoothed firing rates of the RFmap by row for column 8 (top figure); and by column for row 9 (bottom figure). Gray vertical lines: observed. Black line: adaptive approach. Red line: non adaptive approach.

Acknowledgments: This study was supported by grants IT620-13, MTM2014-55966-P, MTM2014-52184-P and Severo Ochoa Program SEV-2013-0323 and Basque Government BERC Program 2014-2017. The first author thanks the Agrupamento INBIOMED from DXPCTSUG-FEDER unha maneira de facer Europa (2012/273). This study was also partially supported by Plan Nacional 2013-2016, ISCIII (RETICS, Oftared) RD12/0034/0017, and FEDER.

References

Currie, I., Durbán, M. and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, **68**, 259 – 280.

Eilers, P.H.C. and Marx, B.D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*, **66**, 159 – 174.

Rodríguez-Álvarez, M.X., Cadarso-Suárez, C., González, F. (2012). Analysing Visual Receptive Fields through Generalised Additive Models with Interactions. *SORT*, **36**, 3 – 32.

Rodríguez-Álvarez, M.X., Durbán, M., Lee, D-J. and Eilers, P.H.C. (2015a). Fast estimation of multidimensional adaptive P-spline models. In:

Proceedings of the 30th Workshop on Statistical Modelling, Friedl, H. and Wagner, H. (Eds.), pp. 330 – 335.

Rodríguez-Álvarez, M.X., Lee, D-J., Kneib, T., Durbán, M. and Eilers, P.H.C. (2015b). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing*, **25**, 941 – 957.

Comparison of methods to analyse longitudinal data truncated by death

Anaïs Rouanet[1,2], Hélène Jacqmin-Gadda[12]

[1] University of Bordeaux, ISPED, Centre Inserm U1219, Bordeaux FRANCE
[2] Inserm, ISPED, Centre Inserm U1219, Bordeaux FRANCE

E-mail for correspondence: anais.rouanet@isped.u-bordeaux2.fr

Abstract: In epidemiologic cohorts of elderly, follow-up is often interrupted by drop-out or death. The two main methods used for handling this type of data are mixed models, estimated by likelihood maximisation and marginal models, estimated by GEE. When focusing on covariate effects, there is a consensus on the fact that mixed models estimate a subject-specific effect while marginal models estimate a population-based effect. When the models are linear, regression parameters have both interpretations. Nevertheless, when the follow-up may be truncated by death, the interpretation of the estimands is under debate. On one hand, marginal models estimate the effect among the population that is alive, prone to attrition. As the surviving population is likely to be healthier, several weighting methods of GEE were proposed to correct for this selection. On the other hand, mixed models estimate the effect of the covariates adjusted on the random effects ('subject-specific effect') but some authors interpret these parameters as the effect in a hypothetic population with no risk of death, said immortal cohort. In this work, we compare mixed models estimated by likelihood maximisation and the marginal models estimated by unweighted/weighted GEE in terms of parameter interpretation, efficiency and robustness through a simulation study. We investigate different associations between the marker and the risks of death and drop-out and we consider two weighting methods: by the inverse probability to be observed and by the inverse probability to be observed given that the subject is alive. These models are also applied to the French prospective cohort Paquid which includes 3777 subjects who took cognitive tests every 2/3 years during 25 years.

Keywords: Death; Longitudinal data; Marginal models; Missing data; Mixed models.

1 Introduction

In elderly cohorts, longitudinal data may be interrupted as subjects drop out or die, leading to possibly informative missing data. The issue is even more critical if death has common risk factors with the process under consideration.

The most frequently used methods are mixed models and marginal models. Mixed models are estimated by likelihood maximisation on available data and this estimation procedure is equivalent to imputing data after drop-out and after death. Thus, this method is criticized since the estimators are interpretable among an ”immortal” population, as stated in Kurland (2009), where no one is at risk of death nor drop-out.

On the other hand, marginal models, estimated by Generalized Estimating Equations, provide the ”partly conditional” expectation of the process, defined in Kurland (2005) as the expectation among subjects that are currently alive and observed. Since the considered population tend to be selected, weighted methods were proposed to correct for the selection biases due to death and drop-out.

As there is no consensus on the best method, the aim of this work is to compare linear mixed models, estimated by likelihood maximisation, and linear marginal models, estimated by unweighted/weighted GEE, when the longitudinal follow-up is truncated by death and any other type of drop-out.

2 Interpretations

When considering a linear mixed model, formulated by:

$$E(Y_i(t)|X_i(t)) = X_i(t)\beta + Z_i(t)u_i + \epsilon_i(t) \text{ with } u_i \sim \mathcal{N}(0,B), \epsilon_i(t) \sim \mathcal{N}(0,\sigma^2) \tag{1}$$

$$\text{then } \beta = E(Y_i(t)|X_i(t)=1,u_i) - E(Y_i(t)|X_i(t)=0,u_i)$$

and the effect β of the covariate X on the mean of the marker Y is adjusted on the random effects u_i, which represent the unobserved individual predictors. In other words, β represents the effect of X on the individual change, called ’subject-specific’ effect.

From model (1), the linear marginal model is formulated by:

$$E(Y_i(t)|X_i(t)) = X_i(t)\beta$$

$$\text{then } \beta = E(Y_i(t)|X_i(t)=1) - E(Y_i(t)|X_i(t)=0),$$

and β represents the effect of X on the population mean of Y, averaged on the random effects, called ’population-averaged’ effect. Thus, with complete data, the parameters from the linear mixed model have both the subject-specific and population-averaged interpretations.

When the follow-up is truncated by death and dropout, several expectations are of interest:

- $E(Y_{ij}|X_i(t), u_i)$

  Subject-specific mean in the immortal population

- $E(Y_{ij}|X_i(t), u_i, T_i > t)$

  Subject-specific mean in the population currently alive

- $E(Y_{ij}|X_i(t))$

  Population-averaged mean in the immortal population

- $E(Y_{ij}|X_i(t), T_i > t)$

  Population-averaged mean in the population currently alive

where T_i is the age at death. Under classical mechanisms of informative missing data, we show that the subject-specific interpretation still holds in the population currently alive as:

$$E(Y_i(t)|X_i(t), u_i) = \beta X_i(t) + Z_i(t)u_i = E(Y_i(t)|X_i(t), u_i, T_i > t)$$

Thus, the subject-specific expectations in the population currently alive and in the immortal population are equal and the parameter from mixed models can be interpreted as the subject-specific effect on the individual change in the immortal cohort but also in the mortal cohort.

Nevertheless, with incomplete data, their population-averaged interpretation holds only among the immortal population as:

$$E(Y_i(t)|X_i(t)) = \beta X_i(t) \neq E(Y_i(t)|X_i(t), T_i > t) = \beta X_i(t) + E(u_i|X_i(t), T_i > t)$$

The marginal expectation among subjects currently alive and in the immortal population are different. We propose an approximation of $E(u_i|X_i, T_i > t)$ to quantify the difference between the regression parameters in the immortal cohort and the ones in the population currently alive.

To correct for the selection biases due to drop-out and death, two weighted versions of GEE were proposed. First, Dufouil (2004) proposed to define the weights as follow:

$$w_{ij}^{(1)} = \frac{P(R_{ij} = 1|X_{ij}, S_{ij} = 1)}{P(R_{ij} = 1|X_{ij}, S_{ij} = 1, \mathcal{H}_i(t_{ij}))},$$

with $R_{ij} = 1$ if subject i is observed at visit j and 0 if not, $S_{ij} = 1$ if subject i is alive at visit j and 0 if not, $\mathcal{H}_i(t_{ij})$ the history of the marker prior to time t_{ij}, time of the j^{th} visit for subject i. Thus, the contribution to the score equation of subjects that are alive and who are likely to drop out is overweighted to correct for the selection bias due to drop-out only. The estimated marginal parameters are interpretable among a mortal population with no drop-out.

In a second version, Weuve (2012) defined the weights by:

$$w_{ij}^{(2)} = \frac{P(R_{ij} = 1|X_{ij})}{P(R_{ij} = 1|X_{ij}, \mathcal{H}_i(t_{ij}))},$$

The contribution of subjects who are likely to drop out or to die is overweighted to correct for the selection biases due to drop-out and death. This method estimates the marginal effect among an immortal population with no drop-out nor death.

3 Simulations

To check and illustrate our results, we performed a simulation study to compare linear mixed models estimated by likelihood maximisation and linear marginal models estimated by GEE under different assumptions regarding the association between the process of interest and the risks of death and drop-out, described in Little (2002).

For instance, we show that a linear individual trend may lead to a quadratic trend of the marginal mean of Y in the population still alive. Besides, if there is no subject-specific effect of the covariate X on Y and if the risk of death depends on an interaction between X and Y, an association may be observed between X and the marginal mean of Y among subjects still alive.

4 Application

The mixed models estimated by likelihood maximisation and marginal models estimated by GEE were applied and compared on the French prospective PAQUID cohort, designed to study the normal and pathological brain aging. This cohort includes 3777 subjects from two French departments, who were visited every two or three years during 25 years and completed a battery of psychometric tests. We considered gender and educational level as risk factors of the cognitive decline of the ISAACS Set Test, which assesses verbal fluency.

References

Dufouil, D., Brayne, C. and Clayton, D. (2004). Analysis of longitudinal studies with death and drop-out: a case study. *Statistical Medicine* **23**, 2215–2226.

Kurland, B.F. and Heagerty, P. J. (2005). Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics* **6**, 241–58.

Kurland, B.F., Johnson, L.L., Egleston, B.L. and Diehr, P.H. (2009). Longitudinal Data with Follow-up Truncated by Death: Match the Analysis Method to Research Aims. *Statistical Science* **24**, 211.

Little, R. and Rubin, D. (2002). Statistical Analysis with Missing Data. *Statistical Medicine* Second edition.

Weuve, J., Tchetgen, E. J. T., Glymour, M. M., Beck, T. L., Aggarwal, N. T., Wilson, R. S., Evans, D. A. and de Leon, C. F. M. (2012). Accounting for bias due to selective attrition: the example of smoking and cognitive decline. *Epidemiology*, 23(1), 119.

A Bayesian Analysis of a Three-Level Moderated Mediation Model with Ordinal Outcome

Šárka Rusá[1], Arnošt Komárek[1], Emmanuel Lesaffre[2], Luk Bruyneel[3]

[1] Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
[2] Leuven Biostatistics and Statistical Bioinformatics Centre, KU Leuven, Belgium
[3] Leuven Institute for Healthcare Policy, KU Leuven, Belgium

E-mail for correspondence: `rusa@karlin.mff.cuni.cz`

Abstract: This study analyzes the impact of nurse staffing (patient-to-nurse ratio) and nurse work environment on patient experiences with hospital care and its possible moderation by nurse education. A Bayesian three-level moderated mediation model is fitted to the data collected during a large European nurse workforce survey, the RN4CAST study, whose multi-level structure is taken into account. The measure of nursing care undone behaves as a mediator in this relationship. We treat the ordinal outcome as an observation coming from a latent continuous normal distribution with unknown threshold specification.

Keywords: Bayesian multi-level modelling; Latent variables; Ordinal outcome; Moderated mediation model.

1 The RN4CAST Study and the Research Question

The Registered Nurse Forecasting (RN4CAST) study (Sermeus et al., 2011) is a cross-sectional survey of patients and nurses in 12 European countries, in which the patients and nurses are further clustered in hospitals and nursing units. The data collected in 2009 − 2010 during this FP7-funded project contain information on various hospital characteristics such as nurse staffing, nurse education, number of beds, overtime work, etc. The patients provided information on their satisfaction with hospital care and hospital rating.

The research question we will address in this paper is how the nurse staffing influences patient experiences with hospital care and if this effect changes with nurse education. As an outcome in our model, we used the measure of patients' willingness to recommend the hospital which was an ordinal outcome with 4 levels (*definitely yes, probably yes, probably no, definitely no*).

2 Three-Level Mediation Model with Ordinal Outcome

To answer the research question with the use of the subset of RN4CAST data which corresponds to eight countries where both patient and nurse survey were carried out (Belgium, Finland, Germany, Greece, Ireland, Poland, Spain, and Switzerland). There were several issues we needed to tackle by our modelling approach. The complex multi-level data structure (patients–hospitals–countries) led us to take the clustering of patients into account. Moreover, it was necessary to consider the ordinal nature of the patient outcome, to decide which variables should be included in the mediating part of the model and which variables could act as confounders. The proposed model also allows for a moderating effect of nurse education.

2.1 Definition of Response

Our model is closely related to the three-level moderated mediation model considered by Bruyneel et al. (2015) who analyzed the mediating effect of *nursing care undone* on the relationship of nursing and patient experiences with hospital care. In their paper, the original ordinal outcome variables were transformed to binary responses in order to conduct the analysis with a multi-level probit model, which is a rather artificial approach. Here we base the analysis on full utilization of the information provided by the ordinal outcome.

We denote the ordinal patient outcome (patients' willingness to recommend the hospital in our case) by z_{ijk} which represents the value of the outcome measured on the i-th patient from hospital j and country k.

A natural approach to be used in case the outcome of a model is ordinal, is to assume that the ordinal variable is derived from a continuous latent variable y_{ijk} which is of our primary interest. Note that this approach has become increasingly popular especially in context of the Bayesian approach. We start by exploiting the approach of Song and Lee (2012, p. 87), and assume that the latent variable is normally distributed. In addition, we assume there exist unknown thresholds $\alpha_1, \dots, \alpha_M$, such that if the value of the latent variable occurs between two specified thresholds then it corresponds to a given value of the ordinal outcome. To be more specific,

$$z_{ijk} = m, \qquad \text{if} \quad \alpha_m < y_{ijk} \leq \alpha_{m+1}, \qquad m = 0, \dots, M,$$

$-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_M < \alpha_{M+1} = \infty$ with $0, \ldots, M$ the unique values of z.

2.2 Definition of Covariates

The hospital-level nurse staffing variable was calculated as the mean number of patients assigned to nurses on their last shift. The nurse working environment was measured by the composite nursing work environment score (Aiken et al., 2012) which was calculated from the 32-item Practice Environment Scale of the Nursing Work Index. The education is a hospital-level variable computed as the proportion of nurses with at least a bachelor's degree in the hospital. Other nurse characteristics (averaged by hospital) included in the model were overtime work, performing non-nursing tasks, years of experience, type of employment (full-time, part-time), etc. Those main explanatory variables will be further denoted by $\boldsymbol{x}_{jk}$.

The nurse staffing and the nurse education variables were both group-mean centered in order to facilitate easier interpretation (the country-level mean was subtracted from the original hospital-level mean). Next to the previously mentioned explanatory variables, we also controlled for the effect of several hospital structural characteristics such as the number of beds, the teaching status (teaching hospital or nonteaching hospital) and technology level (with open heart surgery, organ transplantation, or both defining high-technology hospitals). We denote those explanatory variables by $\boldsymbol{c}_{jk}$.

When taking into account the extent of nursing care undone, it is logical that poor staffing in some hospitals can lead to more nursing care left undone which in turn will affect patient experiences with care. In other words, nursing care undone ($\boldsymbol{m}_{jk}$) mediates the relationship of patient satisfaction and the main explanatory variables. The dependence of the outcome y_{ijk} on the independent variables $\boldsymbol{x}_{jk}$ is then partially explained by two measures of nursing care undone $\boldsymbol{m}_{jk}$(*clinical care activities left undone* and *planning and communication activities left undone*).

2.3 Model

The purpose of our modelling was to ascertain whether there was a significant effect of nurse staffing and the quality of nurse work environment on the patient outcome. Moreover, we added the moderating effect of nurse education such that we included the interaction term between the nurse staffing and the nurse education variables into the model.

In a mediation analysis, we focus on the estimation of the indirect effect of X on Y through a mediator variable M which can be illustrated using two linear models: $M = \gamma_0 + \gamma_X X + \epsilon_M$ and $Y = \beta_0 + \beta_X X + \beta_M M + \epsilon_Y$. The mediation is said to be moderated by another variable W if the interaction between X and W is added to both equations. In other words, the equations in a moderated mediation model have the forms $M = \gamma_0 +$

$\gamma_X X + \gamma_W W + \gamma_{XW} XW + \epsilon_M$, $Y = \beta_0 + \beta_X X + \beta_W X + \beta_{XW} WX + \beta_M M + \epsilon_Y$. In our application, the effect of nurse staffing was moderated by nurse education (in the following model, the interaction is included in the vector of covariates $\boldsymbol{x}_{jk}$).

In order to deal with the multi-level structure of the RN4CAST data, we also included both hospital-level (u_{jk}) and country-level (u_k) random effects to capture dependencies between observations measured at common hierarchical levels. Consequently, the considered model equation for the latent outcome variables has the following form:

$$y_{ijk} = \beta_0 + \boldsymbol{\beta}_c^\top \boldsymbol{c}_{jk} + \boldsymbol{\beta}_x^\top \boldsymbol{x}_{jk} + \boldsymbol{\beta}_m^\top \boldsymbol{m}_{jk} + u_{jk} + u_k + \varepsilon_{ijk},$$

where

$$u_{jk} \sim \mathcal{N}(0, \sigma^2_{hospital}), \quad u_k \sim \mathcal{N}(0, \sigma^2_{country}), \quad \varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2),$$

$$\boldsymbol{m}_{jk} = (m_{1,jk}, m_{2,jk})^\top, \qquad m_{t,jk} = \gamma_{t0} + \boldsymbol{\gamma}_{tx}^\top \boldsymbol{x}_{jk} + \xi_{t,jk},$$
$$\xi_{t,jk} \sim \mathcal{N}(0, \sigma^2_{mt}),\ t = 1, 2.$$

We assume that $u_{jk}, u_k, \varepsilon_{ijk}, \xi_{1,jk}, \xi_{2,jk}$ are mutually independent. To avoid identification issues, we fix the thresholds α_1 and α_K as was suggested in Song and Lee (2012, p. 89).

Bayesian inference based on MCMC simulation was used. To this end, semi-conjugate vague priors on the model parameters were specified. Namely, normal priors for all β and γ parameters ($\beta_0, \boldsymbol{\beta}_c^\top, \boldsymbol{\beta}_x^\top, \boldsymbol{\beta}_m^\top, \gamma_{10}, \gamma_{20}, \boldsymbol{\gamma}_{1x}^\top, \boldsymbol{\gamma}_{2x}^\top$), an inverse gamma prior for ($\sigma^2_{hospital}, \sigma^2_{country}, \sigma^2, \sigma^2_{m1}, \sigma^2_{m2}$). For the thresholds, the following noninformative prior was used (Song and Lee, 2012, p.117):

$$p(\alpha_2, \ldots, \alpha_{M-1}) \propto C, \qquad \text{for} \quad \alpha_2 < \cdots < \alpha_{M-1}.$$

The samples from the posterior distributions were obtained using the Stan software. The inference is based on MCMC methods, mainly the Hamiltonian Monte Carlo and the Metropolis algorithm.

3 Results

The Bayesian estimates and the numerical standard errors were computed from 10000 observations generated in two chains which were collected after discarding the burn-in of the same length. Using *Intel(R) Xeon(R) CPU E5520 @ 2.27GHz, 24 GB RAM*, the computing time for getting the results was approximately 6.5 hours. Table 1 shows the estimated posterior means and HPD intervals of the parameters in the moderated mediation model. Because of the limited space we do not show the estimates of the parameters in the *planning/communication left undone* mediating equation.

	Patients recommending the hospital	
	Estimate	95% HPD CI
Organization of nursing care		
Nurse staffing	−0.025*	(−0.051; −0.001)
Nurse education	0.336	(−0.048; 0.696)
Staffing * Education	−0.072	(−0.225; 0.099)
Nurse working env.	0.184	(−0.064; 0.435)
Years of experience	−0.006	(−0.024; 0.010)
Type of employment	−0.206	(−0.573; 0.153)
Non-nursing tasks	0.070	(−0.265; 0.412)
Overtime work	0.379*	(0.039; 0.735)
Care left undone (mediators)		
Clinical care	−0.205*	(−0.302; −0.106)
Planning/communication	−0.004	(−0.126; 0.136)
Confounding variables		
Number of beds	0.000	(0.000; 0.000)
Technology level	−0.021	(−0.155; 0.115)
Teaching status	−0.004	(−0.172; 0.171)

	Clinical care left undone	
	Estimate	95% HPD CI
Organization of nursing care		
Nurse staffing	−0.010	(−0.056; 0.036)
Nurse education	0.722*	(0.002; 1.454)
Staffing * Education	−0.415*	(−0.711; −0.100)
Nurse working env.	−1.305*	(−1.651; −0.948)
Years of experience	−0.033*	(−0.063; −0.004)
Type of employment	0.225	(−0.338; 0.837)
Non-nursing tasks	−0.075	(−0.655; 0.505)
Overtime work	0.831*	(0.203; 1.453)

Note. HPD CI = Highest posterior density credible interval, * indicates statistical significance, Staffing * Education = Interaction of Nurse staffing and Nurse education.

TABLE 1. Findings for the Eight-Country Moderated Mediation Analysis Estimating the Moderating Effect of Nurse Education on the Effect of Nurse Staffing on Patient Recommending the Hospital Through Care Left Undone.

Worse nurse staffing (= higher patient-to-nurse ratio) leads to lower values of patients' willingness to recommend the hospital.
With respect to the *clinical care left undone* mediator, the effect of nurse staffing in its mediating equations is not significant but the inclusion of nurse staffing and nurse education interaction is essential. In hospitals with poor nurse staffing, its effect is smaller if the proportion of nurse with a bachelor degree is higher. Neither the effect of nurse staffing, nor the interaction between nurse staffing and nurse education is significant in the other mediator equation.
The effect of nurse working environment and work experience is fully mediated through *clinical care undone*. More favourable nurse working environment leads to less care left undone which in turn results in better patient satisfaction with care. Similarly, longer experience of working as a nurse leads to fewer tasks left undone.
More *planning / communication left undone* is related to higher proportion of nurses performing non-nursing tasks. The negative effect of overtime work on patients' willingness to recommend the hospital is partially mediated through *clinical care left undone*. The type of employment was not associated with either one of the mediators or the patient outcome.
Clinical care left undone is significantly associated with patients' recommending the hospital, while the relationship between *planning / communication left undone* is not.

Acknowledgments: The first author gratefully acknowledges support from the grant GACR 15-04774Y. Support from the Interuniversity Attraction Poles Programme (IAP-network P7/06), Belgian Science Policy Office, is also gratefully acknowledged.

References

Aiken, L. H., Sermeus, W., Van den Heede, K. et al. (2012). Patient safety, satisfaction, and quality of hospital care: cross sectional surveys of nurses and patients in 12 countries in Europe and the United States. *British Medical Journal*, **344**, e1717.

Bruyneel L., Li B., Ausserhofer D., Lesaffre E. et al. (2015). Organization of Hospital Nursing, Provision of Nursing Care, and Patient Experiences With Care in Europe. *Medical Care Research and Review*, **72(6)**, 643 – 664.

Song, X. Y. and Lee, S.Y. (2012). *Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences.* Chichester, West Sussex: John Wiley.

Sermeus, W., Aiken, L. H., Van den Heede, K. et al. (2011). Nurse forecasting in Europe (RN4CAST): Rationale, design and methodology. *BMC Nursing*, **10(6)**, 643 – 664.

Subject-Object-Specific Covariates in Paired Comparison Models – An Application to Data from the German Bundesliga

Gunther Schauberger[1], Andreas Groll[1], Gerhard Tutz[1]

[1] LMU Munich, Germany

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

Abstract: A model for results of football matches is proposed that is able to take into account match-specific covariates as, for example, the total distance a team runs in the specific match. The model extends the Bradley-Terry model in many different ways. In addition to the inclusion of covariates, it considers ordered response values and (possibly team-specific) home effects. Penalty terms are used to reduce the complexity of the model and to find clusters of teams with equal covariate effects.

Keywords: Bradley-Terry; BTLLasso; Paired Comparison; Football data.

1 Introduction

Paired Comparisons occur if two objects are compared with respect to an underlying latent trait. In this work, we consider football matches and treat them as paired comparisons between two teams where the underlying latent traits are the playing abilities of the teams. The data we consider are data from the season 2014/15 of the German Bundesliga. In particular, match-specific covariates are used to model the results from single matches. In general, if covariates are to be considered in paired comparison, one has to distinguish between subjects and objects of the paired comparisons and, accordingly, between subject-specific, object-specific and subject-object-specific covariates. In football matches, the teams are the objects while a single match can be considered to be the subject that makes the comparison between the two objects/teams. In our application, subject-object-specific covariates are considered.

The Bradley-Terry model (Bradley and Terry, 1952) is the standard model for paired comparison data. Assuming a set of objects $\{a_1, \ldots, a_m\}$, in its most simple form the Bradley-Terry model is given by

$$P(a_r \succ a_s) = P(Y_{(r,s)} = 1) = \frac{\exp(\gamma_r - \gamma_s)}{1 + \exp(\gamma_r - \gamma_s)}.$$

One models the probability that a certain object a_r dominates or is preferred over another object a_s, $a_r \succ a_s$. The random variable $Y_{(r,s)}$ is defined to be $Y_{(r,s)} = 1$ if a_r dominates a_s and $Y_{(r,s)} = 0$ otherwise. The parameters γ_r represent the attractiveness or strength of the respective objects.

2 Bundesliga Data

The main goal of this work is to analyze if (and which) match-specific covariates influence the result of football matches. Match-specific covariates are information on specific measurements of the teams in each match, as for example the number of kilometers a team runs (*Distance*). In total, all the following covariates are known per team and per match:

Distance Total amount of km run

BallPossession Percentage of ball possession

TacklingRate Rate of won tacklings

ShotsonGoal Total number of shots on goal

Passes Total number of passes

CompletionRate Percentage of passes reaching teammates

FoulsSuffered Number of fouls suffered

Offside Number of offsides (in attack)

In particular, it is interesting which covariates have an influence at all and for which covariates there are different effects for single teams. As the covariates we consider are collected per team and per match, they generally can be termed as subject-object-specific covariates.

3 A Paired Comparison Model for Football Matches Including Subject-Object-Specific Covariates

When using a paired comparison model for football matches several extensions compared to the standard Bradley-Terry model are needed. The model has to be able to handle an ordinal response (in particular draws), home effects and subject-object-specific covariates.

For that purpose, we propose to use the general model for ordinal response data $Y_{i(r,s)} \in \{1, \ldots, K\}$ denoted by

$$\begin{aligned} P(Y_{i(r,s)} \leq k) &= \frac{\exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})}{1 + \exp(\delta_r + \theta_k + \gamma_{ir} - \gamma_{is})} \\ &= \frac{\exp(\delta_r + \theta_k + \beta_{r0} - \beta_{s0} + \boldsymbol{z}_{ir}^{\mathrm{T}}\boldsymbol{\alpha}_r - \boldsymbol{z}_{is}^{\mathrm{T}}\boldsymbol{\alpha}_s)}{1 + \exp(\delta_r + \theta_k + \beta_{r0} - \beta_{s0} + \boldsymbol{z}_{ir}^{\mathrm{T}}\boldsymbol{\alpha}_r - \boldsymbol{z}_{is}^{\mathrm{T}}\boldsymbol{\alpha}_s)}. \end{aligned}$$

Basically, the model is a special case of a cumulative logit model and allows for the inclusion of so-called subject-object-specific covariates $\boldsymbol{z}_{ir}$. See also Tutz and Schauberger (2015) for a model including object-specific covariates $\boldsymbol{z}_r$ and Schauberger and Tutz (2015) for a model including subject-specific covariates $\boldsymbol{z}_i$. $Y_{i(r,s)}$ encodes an ordered response with K categories (including a category for draws) for a match between team a_r and team a_s on matchday i where a_r played at its home ground. The linear predictor of the model contains the following terms:

δ_r team-specific home effects of team a_r

θ_k category-specific threshold parameters

β_{r0} team-specific intercepts

$\boldsymbol{z}_{ir}$ p-dimensional covariate vector that varies over teams and matches

$\boldsymbol{\alpha}_r$ p-dimensional parameter vector that varies over teams.

In general, for ordinal paired comparisons it can be assumed that the response categories have a symmetric interpretation so that $P(Y_{(r,s)} = k) = P(Y_{(s,r)} = K - k + 1)$ holds. Therefore, the threshold parameters should be restricted with $\theta_k = -\theta_{K-k}$ and, if K is even, $\theta_{K/2} = 0$ to guarantee for symmetric probabilities. Instead, the home effects now cover the possible order effects (the advantage of the home team a_r over the away team a_s). Instead of fixed abilities of the teams γ_r, the teams have matchday-specific abilities $\gamma_{ir} = \beta_{r0} + \boldsymbol{z}_{ir}^{\mathrm{T}}\boldsymbol{\alpha}_r$, depending on the covariates of team $a_r$on matchday i.

Both the home effect and the covariate effects could also be included as global parameters instead of team-specific parameters. To decide, whether the home effect or single covariate effects should be considered with team-specific or global parameters, penalty terms will be used. In particular, the absolute values of all pairwise differences between the team-specific home advantages are penalized using the penalty term

$$P(\delta_1, \ldots, \delta_m) = \sum_{r<s} |\delta_r - \delta_s|.$$

The penalty term enforces the clustering of teams with equal home effects as it is able to set differences between parameters to exactly zero. As an extreme case, the penalty leads to one global home effect parameter if all differences are set zero. Additionally, also for the team-specific covariate effects a penalty term is introduced that penalizes the absolute values of all pairwise differences of the covariate parameters and of the parameters themselves by

$$J(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m) = \sum_{j=1}^{p} \sum_{r<s} |\alpha_{rj} - \alpha_{sj}| + \sum_{j=1}^{p} \sum_{r=1}^{m} |\alpha_{rj}|.$$

The penalty enforces clustering of teams with respect to certain covariates, possibly leading to global effects instead of team-specific effects. Moreover, due to the penalization of the absolute values covariates can be eliminated completely from the model. For comparability of the penalties and the resulting effects, all covariates have to transformed to a joint scale.
Both penalty terms are combined and the respective penalized likelihood

$$l_p(\cdot) = l_p(\cdot) - \lambda(P(\delta_1, \ldots, \delta_m) + J(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m))$$

is maximized. The tuning parameter λ is chosen by 10-fold cross-validation.

4 Application to Bundesliga Data

For easier interpretation of the intercepts, the covariates were centered (per team around the team-specific means). Centering the covariates only changes the paths (and interpretation) of the team-specific intercepts which now represent the ability of a team with every covariate at the team-specific mean of this covariate. The paths and the interpretation of the covariate effects stay the same, representing the effect of a covariate for the team ability when the respective covariate changes (deviates from the team-specific means).
Figure 1 illustrates the parameters paths for the proposed model, separately for each covariate along the tuning parameter λ. The dashed vertical line indicates the model that was selected by 10-fold cross-validation. The paths illustrate the clustering effect of the penalty terms. It can be seen, that the home effect seems to be equal for almost all teams, only Eintracht Frankfurt has its own team-specific home effect. The home effect is positive, therefore, playing at the home ground is an advantage for all teams. The greatest effect of all covariates can be seen for *Distance*. It has a positive effect for all teams (except Borussia Mönchengladbach). Therefore, the teams gain better results in matches where they had a good running performance. Interestingly, the covariate *BallPossession* has rather negative effects for all teams. The covariates *ShotsonGoal*, *CompletionRate* and *Offside* were eliminated completely from the model.

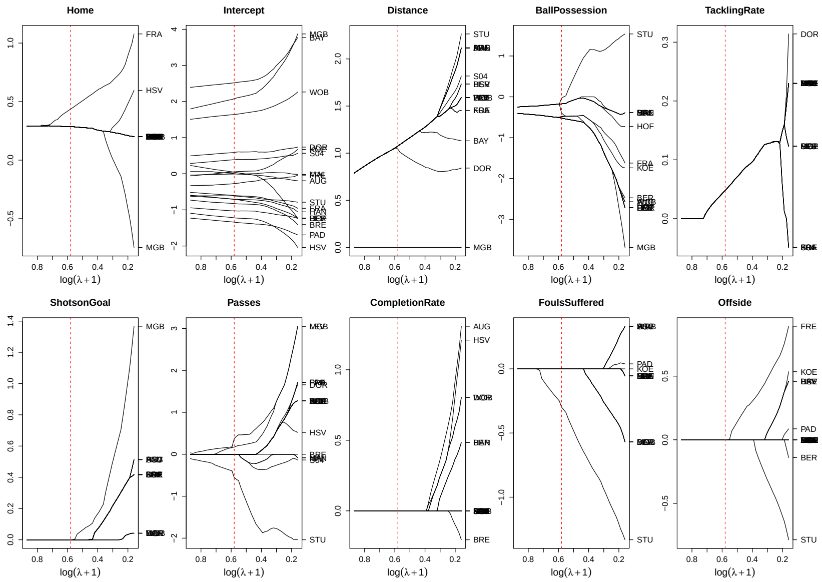


FIGURE 1. Parameter paths, separately for home effect, intercepts and all (centered) covariates. Dashed vertical line represents the optimal model according to 10-fold cross-validation.

5 Alternative Modeling Approach

The covariate effects of the model proposed in Section 3 have a very specific interpretation. Every team has an (unpenalized) intercept that reflects the average ability of the team over the season. Therefore, the intercepts also already cover the mean covariate effects of all teams. Accordingly, the covariate effects captured in the respective parameter vectors $\boldsymbol{\alpha}_r$ represent effects where covariates can explain deviations of the team performance from its average performance.

However, if one is interested in the total effect of a covariate for the performance of single teams, a different parameterization becomes necessary. In an alternative approach, the team-specific intercepts are simply eliminated from the model. In this parameterization, the specific ability of team a_r on matchday i is specified by $\gamma_{ir} = \boldsymbol{z}_{ir}^{\mathrm{T}}\boldsymbol{\alpha}_r$ instead of $\gamma_{ir} = \beta_{r0} + \boldsymbol{z}_{ir}^{\mathrm{T}}\boldsymbol{\alpha}_r$ as in Section 3.

In this alternative approach, the mean abilities of the teams can not be modelled by the team-specific intercepts and have to be replaced by covariate effects. That also implies, that in this alternative model the average values of the covariates for each team matter and the covariates are not centered anymore. In Figure 2, for the optimal model (according to 10-fold cross-validation) the estimated parameters multiplied by the means of the respective covariates are plotted. Per covariate, each of these 'total covariate effects' is represented by a circle, larger circles represent (the size of) clusters of teams. For illustration, the teams Bayern München, Borussia Dortmund and SC Paderborn are highlighted. Now the covariates *BallPossession* and *Passes* seem to be the most influential covariates. For example, Bayern München has a very large effect for *BallPossession*, and, therefore, the dominance of Bayern München is strongly related to the *BallPossession* of Bayern München. For SC Paderborn (relegated to the 2. Bundesliga), its rather bad performance is mainly explainable by the covariate *Passes*.

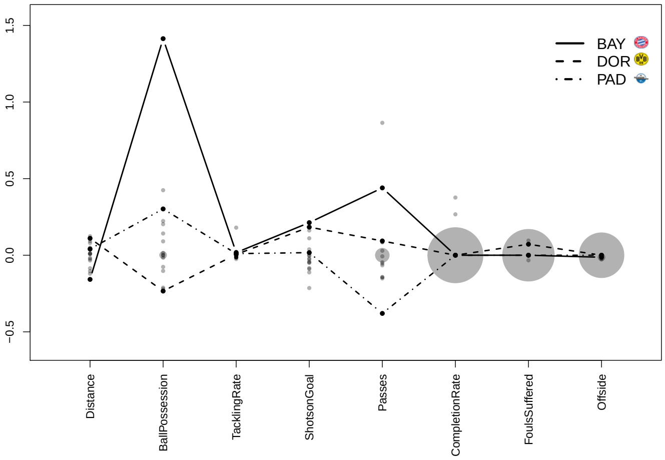


FIGURE 2. Total covariate effects (estimated parameters multiplied by team-specific covariate means) for alternative modeling approach for optimal model according to 10-fold cross-validation.

References

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs, I: The method of pair comparisons. *Biometrika*, **39**, 324 – 345.

Schauberger, G. and Tutz, G. (2015). Modelling Heterogeneity in Paired Comparison Data - an L1 Penalty Approach with an Application to Party Preference Data. *Department of Statistics, LMU München*, Technical Report 183.

Tutz, G. and Schauberger, G. (2015). Extended Ordered Paired Comparison Models with Application to Football Data from German Bundesliga. *Advances in Statistical Analysis*, **99(2)**, 209 – 227.

NEAT: an efficient Network Enrichment Analysis Test

Mirko Signorelli[1,2], Veronica Vinciotti[3], Ernst C. Wit[1]

[1] Johann Bernoulli Institute, University of Groningen, Netherlands
[2] Department of Statistical Sciences, University of Padova, Italy
[3] Department of Mathematics, Brunel University London, United Kingdom

E-mail for correspondence: m.signorelli@rug.nl

Abstract: Network enrichment analysis (NEA) integrates gene enrichment analysis with information on dependences between genes. Existing tests for NEA rely on normality assumptions, they can deal only with undirected networks and are computationally slow. We propose NEAT, an alternative test based on the hypergeometric distribution. NEAT can be applied also to directed and mixed networks, and it is faster and more powerful than existing NEA tests.

Keywords: networks; enrichment analysis; gene expression.

1 Introduction

When the first data on gene expression became available, they were analysed considering each gene separately. However, researchers soon realized that genes act in a concerted manner, and that cellular processes are often the result of complex interactions between different genes and molecules. Nowadays, sets of genes that are responsible for many cellular functions have been identified, and are collected in publicly available databases (such as GO and KEGG). These sets of genes, whose function is already known, can be used to characterize and interpret ("enrich") the results of new experiments. This characterization is typically done by means of gene enrichment analysis (GEA) tests, which allow to compare gene expression levels between two conditions (experimental and control) and to detect functional sets of genes that are activated or repressed in the experimental condition. The power of GEA tests is often low, mostly because they consider the level of overlap between sets of genes only, and they ignore associations and dependences that exist between genes.

This paper was published as a part of the proceedings of the 31st International Workshop on Statistical Modelling, INSA Rennes, 4–8 July 2016.

Recently, Alexeyenko *et al.* (2012) and McCormack *et al.* (2013) have proposed to integrate GEA with information on dependences between genes by making use of gene networks. The idea is that "enrichment" between two sets of genes A and B can be assessed by comparing the number of links connecting nodes in A and B, n_{AB}, with a reference distribution that assumes that no relation exists between the two sets. Their tests rely on a normal approximation for the reference distribution (which is discrete), they require the computation of many network permutations (an activity that can be highly time consuming) and are restricted to the analysis of undirected networks.

In the sequel we propose NEAT, an alternative Network Enrichment Analysis Test based on the hypergeometric distribution. The assumption that in absence of enrichment N_{AB} is distributed as an hypergeometric arises quite naturally, and enables us to avoid normal approximations and network permutations. We develop NEAT not only for undirected, but also for directed and mixed networks, thus providing a common framework for the analysis of different types of networks.

2 Methods

A graph is a pair $\mathcal{G} = (V, E)$, which consists of a set of nodes V connected by a set of directed or undirected edges $E \subseteq V \times V$. In gene regulatory networks each gene is represented as a node of the graph, and an edge between two nodes is drawn to signify dependence between the corresponding genes. In the inferred network, we expect that individual links may be slightly unstable and noisy. However, we do expect that inferred links contain a sign of the relationships between functional gene sets. So, if there is a functional relationship (i.e., enrichment) between functions described by sets $A \subset V$ and $B \subset V$, then we expect the number of links between the two groups to be larger (or smaller) than expected by chance.

2.1 Directed and mixed networks

In directed networks, we assess the presence of enrichment from A to B by considering the number of arrows n_{AB} going from genes in A to genes belonging to B. The observed n_{AB} can be thought as a realization from the random variable N_{AB}, with expected value μ_{AB}. We compare μ_{AB} with the number of arrows μ_0 that we would expect to observe from A to B by chance, and test $H_0 : \mu_{AB} = \mu_0$ versus $H_1 : \mu_{AB} \neq \mu_0$. We say that there is enrichment from A to B if μ_{AB} is significantly different from μ_0.

We use the hypergeometric distribution to model the null distribution of N_{AB}. The hypergeometric models the number of successes in a random sample without replacement: in our case, let's mark arrows that reach genes in B as "successful", and the remaining ones as "unsuccessful". If there is no relation between A and B, we can view the arrows that go out from genes in A as a random sample without replacement from the population of arrows present in the graph, and n_{AB} as the number of successes in that sample. Thus, the distribution of N_{AB} when H_0 is true is

$$N_{AB} \sim hypergeom(n = o_A, K = i_B, N = i_V), \qquad (1)$$

where the sample size o_A is the outdegree of A, the number of successes in the population i_B is the indegree of B and the population size i_V is the total indegree of the network. So, we expect $\mu_0 = o_A \frac{i_B}{i_V}$ to increase as the indegree of A, or the outdegree of B, increases. A toy example that explains the rationale behind NEAT is presented in Figure 1.

Bearing in mind the fact that for a discrete test statistic T the usual formula for p-values $p_1 = 2 \min P_0[(T \leq t), P(T \geq t)]$ can exceed 1, we compute the p-value using

$$p = 2 \min \left[P_0(N_{AB} > n_{AB}), P_0(N_{AB} < n_{AB})\right] + P_0(N_{AB} = n_{AB}), \qquad (2)$$

which differs from p_1 by a factor equal to $P_0(T = t)$. A p-value close to 0 can be regarded as evidence of enrichment, because it entails that n_{AB} is significantly higher/smaller than we would expect it to be under H_0. For a given type I error α, one can then conclude that there is enrichment from A to B if $p < \alpha$.

A mixed network is a network where both directed and undirected edges are present. It is possible to regard a mixed network as a directed network, where every undirected edge $v \sim w$ stands for two directed arrows, $v \to w$ and $w \to v$. NEAT adopts such convention for the analysis of mixed networks.

2.2 Undirected networks.

When dealing with undirected networks, the presence of enrichment between A and B depends on the number of links n_{AB} that connect genes in A to genes in B. Here, there is no distinction between indegree and outdegree of a node, and it only makes sense to consider the degree of a node: thus, assumption (1) needs to be properly modified. Define the total degree of a set as the sum of the degrees of nodes that belong to it: then, the null distribution is $N_{AB} \sim hypergeom(n = d_A, K = d_B, N = d_V)$, where d_A, d_B and d_V are the total degrees of sets A, B and V.

2.3 Software.

NEAT is implemented in the `R` package `neat`, which is available on `CRAN` (Signorelli *et al.*, 2016). `neat` allows the user to specify the network in different formats, and it includes a set of data and examples.

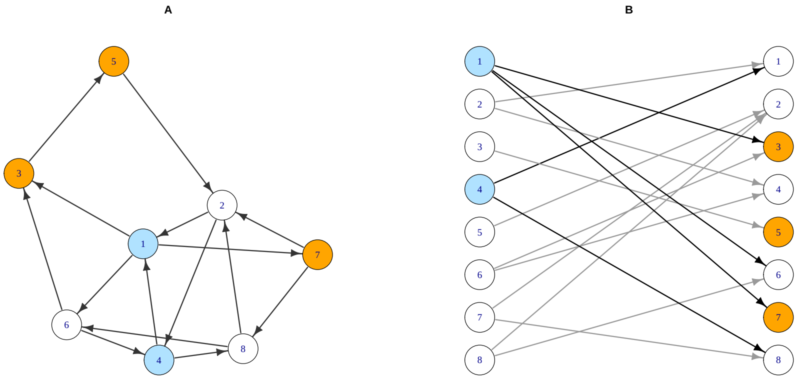


FIGURE 1. *A directed network with 8 nodes (A) and its bipartite representation (B).* Suppose that one wants to know whether there is enrichment from set $A = \{1, 4\}$ to set $B = \{3, 5, 7\}$. There are 5 arrows going out from A, and 2 of them reach B. The whole network consists of 15 arrows, of which 4 reach B. Thus, $n_{AB} = 2$, $o_A = 5$, $i_B = 4$ and $i_V = 15$. The idea behind NEAT is that, if the 5 arrows that are going out from A are a random sample (without replacement) from the population of 15 arrows that are present in the network, then the proportion of arrows reaching B from A should be close to the proportion of arrows reaching B in the whole network. In this case, it seems that arrows going out from A tend to reach B more frequently (40%) than other arrows do (27% of the 15 arrows in the network reach B). However, the computation of the test leads to $p = 0.48$: the observed $n_{AB} = 2$ does not provide enough evidence to reject the null hypothesis that there is no enrichment from A to B.

3 Simulations

We compare the performance of NEAT with the NEA test of Alexeyenko *et al.* (2012) and with the LP, LA, LA+S and NP tests of McCormack *et al.* (2013) by means of two simulations. We simulate two undirected random networks with 1000 nodes, whose degree distributions are a power law in simulation S1, and a mixture of Poisson distributions in simulation S2. We test enrichment between 50 sets of nodes, with cardinality ranging from 50 to 100 nodes. We modify the original networks to introduce enrichments between 100 pairs of these sets, by either increasing or reducing n_{AB} by a proportion uniformly ranging from 10 to 50%. The results (see Table 1) show that the distribution of p-values is uniform in both cases for NEAT and LA, and in one case for LA+S (S1) and NP (S2). NEA and LP, instead, do not produce uniform distributions in any case. In both S1 and S2, NEAT turns out to have the highest discriminatory capacity (AUC) and to be by far the fastest method, from 22 to 3000 times faster than alternative tests.

TABLE 1. Results of simulation S1 and S2. The best results in each column are bolded. Abbreviations: p^{KS} denotes the p-value of the Kolmogorov-Smirnov test for $H_0 : X \sim U(0, 1)$; AUC stands for "area under the ROC curve". Time is expressed in seconds.

	Simulation S1			Simulation S2		
Test	p^{KS}	AUC	Time	p^{KS}	AUC	Time
NEAT	**0.399**	**0.920**	**0.6**	**0.343**	**0.925**	**0.7**
NEA	0.001	**0.918**	2125.4	0.024	0.912	2151.5
LP	0	0.908	28.6	0	0.904	44.7
LA	**0.255**	0.897	14.4	**0.111**	0.908	18.0
LA+S	**0.409**	0.913	21.8	0.024	0.910	27.6
NP	0.037	0.884	12.9	**0.323**	0.908	15.8

4 Data analysis

After analysing gene expression patterns of yeast *Saccaromyces cerevisiae* in response to different stressful stimuli, Gasch *et al.* (2000) inferred the existence of two set of genes, collectively called *Environmental Stress Response* (ESR), that constitute a coordinated, initial reaction to the emergence of any hostile condition in the cell. The original study made use of a GEA test to characterize the two sets. Here, we incorporate into the analysis known associations between genes, as represented in the *YeastNet* network (Kim *et al.*, 2013). For lack of space, we do not show here the lists of enrichments detected by NEAT for the two ESR sets; however, such lists can be retrieved running the example in the help page `?yeast` of the `R` package `neat` (Signorelli *et al.*, 2016). In short, NEAT detects most of the enrichments that were found in the original study for the two ESR sets; besides, it unveils some further enrichments related to molecular transportation and amino-acid biosynthesis for the set of induced ESR genes, which would be overlooked if functional couplings between genes were ignored.

5 Conclusion

Traditional gene enrichment analysis assesses enrichment between gene sets solely on the basis of the extent of their overlap. Network enrichment analysis is a powerful extension of traditional GEA tests, which makes use of genetic networks to integrate enrichment analyses with information on associations and dependences that exist between genes.

We have developed NEAT, a test for network enrichment analysis that aims to overcome some limitations of the resampling-based tests of Alexeyenko *et al.* (2012) and McCormack *et al.* (2013). First of all, we believe that a normal approximation does not make justice to the discrete nature of N_{AB}. We have showed that this approximation can be avoided, if one models N_{AB} with the hypergeometric distribution. In addition, existing NEA tests require the computation of many network permutations: this operation can be highly time consuming, slowing down computations considerably. NEAT, instead, fully specifies the null distribution of N_{AB} without resorting to permutations, thus speeding up the computation of the test. A further drawback of existing resampling-based tests is that they have been implemented only for undirected networks: we address this problem proposing two different parametrizations for NEAT, that take into account the different nature of directed and undirected edges.

The test is implemented in the `R` package `neat`, which is freely available on `CRAN` (Signorelli *et al.*, 2016). Our simulations show that NEAT behaves well under the null hypothesis, is more powerful and faster than existing NEA tests. Application to the Environmental Stress Response data shows that NEAT can detect most of the enrichments that were found with GEA methods, and unveils further enrichments that would be overlooked, if depencences between genes were ignored. We believe that NEAT could constitute a flexible and computationally efficient test for network enrichment analysis. Potential applications of NEAT extend beyond gene regulatory networks, and include social networks, brain networks and other situations where one attempts to understand the relation between groups of vertices in a network.

References

Alexeyenko, A., Lee, W., Pernemalm, M., Guegan, J., Dessen, P. *et al.* (2012). Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, **13**:226.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B. *et al.* (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, **11**(12), 4241 – 4257.

Kim, H., Shin, J., Kim, E., Kim, H., Hwang, S. *et al.* (2013). YeastNet v3: a public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. *Nucleic Acids Research*, 1 – 13.

McCormack, T., Frings, O., Alexeyenko, A., Sonnhammer, E. L. (2013). Statistical assessment of crosstalk enrichment between gene groups in biological networks. *PLoS One*, **8**(1):e54945.

Signorelli, M., Vinciotti, V., Wit E. C. (2016). NEAT: efficient Network Enrichment Analysis Test. https://cran.r-project.org/package=neat.

An Adaptive Subspace Method for High-Dimensional Variable Selection

Christian Staerk[1], Maria Kateri[1], Ioannis Ntzoufras[2]

[1] Institute of Statistics, RWTH Aachen University, Germany
[2] Department of Statistics, Athens University of Economics and Business, Greece

E-mail for correspondence: christian.staerk@rwth-aachen.de

Abstract: We propose a variable selection algorithm called the "adaptive subspace method", which aims at finding the best model with respect to a given selection criterion in a sparse high-dimensional situation. Possible selection criteria include ℓ_0-type information criteria like the recently proposed extended BIC. The method is based on the idea of solving many low-dimensional problems in order to solve the given high-dimensional problem using a certain form of adaptive learning. We provide a Bayesian interpretation of the algorithm and investigate its performance in a simulation study.

Keywords: Variable Selection, Information Criteria, Bayesian Learning, Subsampling, Generalized Linear Model.

1 Introduction

Variable selection in high-dimensional settings where the number of explanatory variables p is possibly larger than the sample size n has been a challenging problem in recent applications such as Genomics or Text Categorization. Many different methods have been proposed to solve the variable selection problem in a generalized linear model (GLM) set-up.
The classical approach to the variable selection problem is to come up with a selection criterion and to solve the resulting optimization problem. Selection criteria include, among many others, the Akaike Information Criterion AIC, the Bayesian Information Criterion BIC and the recently proposed Extended Bayesian Information Criterion EBIC (Chen and Chen (2008)), which particularly aims at high-dimensional settings. Chen and Chen (2012) show that under moderate conditions EBIC is a consistent variable selection procedure for a GLM if $p = \mathcal{O}(n^k), k > 0$. The challenging problem with these ℓ_0-type selection criteria is that the resulting

combinatorial optimization problems are in general very difficult to solve if there are many possible explanatory variables p, since there are 2^p possible models for which the criterion has to be evaluated in a full enumeration.

In the 90's the focus shifted from solving discrete optimization problems to solving continuous, convex optimization problems, which actually could be solved. It was Tibshirani (1996) who promoted the use of the famous Lasso, which solves a convex optimization problem with an ℓ_1-penalty on the regression coefficients and then selects those variables whose corresponding regression coefficients are non-zero in the optimal solution. A big drawback of ℓ_1-regularization methods like the Lasso is that they typically require strong conditions on the design matrix X to be variable selection consistent (see e.g. Zhao and Yu (2006)).

Another general problem of criterion based procedures is that per se they do not provide any information about the uncertainty concerning the best model. In fact one observes that the optimal Lasso solution is not very stable with respect to small changes in the sample. Therefore Meinshausen and Bühlmann (2010) propose a procedure called stability selection. It is based on the idea of applying a given variable selection method (e.g. the Lasso) multiple times on subsamples of the data. Finally one selects those explanatory variables whose relative selection frequencies exceed some threshold. The subsampling scheme is to draw a set I of size $\lfloor \frac{n}{2} \rfloor$ without replacement from $\{1, \ldots, n\}$ and then consider the model adjusted to the rows in I.

Even though the stability selection procedure has nice theoretical properties and seems to become more and more popular in practice, one might ask whether it is the best thing that can be done in a high-dimensional situation where the sample size n is typically in the tenths or small hundredths and $p > n$. Note that stability selection successively applies a possibly inconsistent selection procedure like the Lasso on even more severe high-dimensional problems with $p \gg \lfloor \frac{n}{2} \rfloor$. In contrast, the main idea of the proposed adaptive subspace method is to successively apply a consistent model selection procedure (like EBIC) on data with original sample size n and a subset of the p covariates of size less than n.

So the ideology behind the adaptive subspace method can be summarized as: "Solve a high-dimensional problem by solving several low-dimensional ones." Two issues naturally arise in this regime: Which low-dimensional problems should be solved? And how can the information from the solved low-dimensional problems be combined in order to solve the original problem? The proposed algorithm addresses these two issues using a certain form of adaptive learning.

2 The Adaptive Subspace Method

We introduce some general notation in a setting with a criterion-based variable selection procedure. We denote the set of explanatory variables by

$\{X_j;\ j \in \mathcal{P}\}$ with index set $\mathcal{P} = \{1, \ldots, p\}$ and denote the corresponding model space by $\mathcal{M} = \mathfrak{P}(\mathcal{P}) = \{S \subseteq \{1, \ldots, p\}\}$. Let $C : \mathcal{M} \to \mathbb{R}$ be any model selection criterion. We assume w.l.o.g. that we want to find the model $S^* \in \mathcal{M}$ that maximizes the given criterion C, i.e. $S^* := \arg\max_{S \in \mathcal{M}} C(S)$. Examples include posterior model probabilities (within the Bayesian set-up) or the negative EBIC (within the ℓ_0-penalized criteria framework). Let

$$f_C : \mathcal{M} \to \mathcal{M},\ f_C(S) := \arg\max_{\tilde{S} \subseteq S} C(\tilde{S}).$$

So for a given $S \subseteq \{1, \ldots, p\}$, $f_C(S)$ is the best model according to criterion C among all models included in S. We assume w.l.o.g. that $C(S) \neq C(S')$ for all $S, S' \in \mathcal{M}$ with $S \neq S'$, so that f_C is a well-defined map.

Now suppose we have observed some data $\mathcal{D} = (X, Y)$ and we want to identify the best model $S^* = \arg\max_{S \in \mathcal{M}} C(S)$ according to criterion C for a GLM. As explained above, the basic idea of the adaptive subspace method is to solve many low-dimensional problems (i.e. compute $f_C(V)$ for many $V \in \mathcal{M}$ with $|V|$ relatively small) in order to solve the given high-dimensional problem (i.e. compute $S^* = f_C(\mathcal{P})$). More precisely, the steps of the adaptive subspace method are given by:

(1) Initialize the expected size q of the low-dimensional problems to be considered at the beginning, i.e. $q \in [1, p]$, as well as the adaptation rate $K > 0$ and the number of iterations $T \in \mathbb{N}$.

(2) For $j \in \mathcal{P}$ initialize $r_j^{(0)} = \frac{q}{p}$.

(3) For $t = 1, \ldots, T$:

   (a) Draw $b_j^{(t)} \sim \text{Bernoulli}(r_j^{(t-1)})$ independently for $j \in \mathcal{P}$.

   (b) Set $V^{(t)} = \{j \in \mathcal{P};\ b_j^{(t)} = 1\}$.

   (c) Compute $S^{(t)} = f_C(V^{(t)})$.

   (d) For $j \in \mathcal{P}$ update $r_j^{(t)} = \frac{q + K \sum_{i=1}^{t} 1_{S^{(i)}}(j)}{p + K \sum_{i=1}^{t} 1_{V^{(i)}}(j)}$, where 1_A denotes the indicator function of a set A.

So the adaptive subspace method is a stochastic algorithm which samples a subset $V^{(t)} \subseteq \mathcal{P}$ in each iteration t. The probability that $j \in \mathcal{P}$ is included in $V^{(t)}$ is given by $r_j^{(t-1)}$. The selection probabilities $r_j^{(t)}$ are adapted after each iteration t. Note that we implicitly assume that it is computationally feasible to compute $S^{(t)} = f_C(V^{(t)})$ in each iteration t. In fact, if the underlying "truth" is sparse, $|V^{(t)}|$ is expected to be relatively small. Otherwise, if $|V^{(t)}|$ is bigger than some computational upper bound U_C, one might replace $V^{(t)}$ by a subsample of $V^{(t)}$ of size U_C. Alternatively one might use heuristic algorithms in place of a full enumeration. In principle, there are two different choices for the final subset selected by the algorithm after iteration T. One can choose the "best" sampled model $\hat{S}_{\mathrm{b}}$ for which $C(\hat{S}_{\mathrm{b}}) = \max\{C(S^{(1)}), \dots, C(S^{(T)})\}$, or one can consider $\hat{S}_\rho = \{j \in \mathcal{P};\ r_j^{(T)} > \rho\}$ with some threshold $\rho \in (0, 1)$.

The evolution of the algorithm can formally be described by a Markov chain. So the adaptive subspace method is in fact nothing else than a Markov Chain Monte Carlo (MCMC) method. The fundamental difference in comparison to standard MCMC algorithms is that instead of sampling from the (unknown) posterior distribution, the adaptive subspace method constructs a Markov chain which converges (in a sense and under conditions that need to be specified) to the (unknown) true solution $S^* = \arg\max_{S \in \mathcal{M}} C(S)$ of the optimization problem.

Furthermore, the adaptive subspace method can be viewed as a form of Bayesian learning. Let π_j denote the subjective belief that variable X_j is in the best model S^* with respect to some criterion C. Note that under knowledge of $S^* = \arg\max_{S \in \mathcal{M}} C(S)$, in fact π_j has a Dirac distribution concentrated at 1 if $j \in S^*$ or concentrated at 0 if $j \notin S^*$. But since we cannot solve the high-dimensional optimization problem exactly, we sequentially solve low-dimensional problems of the form $S^{(t)} = \arg\max_{S \subseteq V^{(t)}} C(S)$ and sequentially update our belief π_j about variable X_j. This process can be viewed as an adaptive experimental design where in each step we set up the design with explanatory variables given by $V^{(t)}$ and observe "new" data $\mathcal{D}^{(t)} = (X_{|V^{(t)}}, Y)$. Using an appropriate beta-prior for π_j, one can interpret $r_j^{(t)}$ as the (pseudo) posterior mean of π_j after observing $\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(t)}$.

3 Simulation Study

Relatively low-dimensional ($p \leq 40$) simulation studies using BIC as a selection criterion in linear models show that in many cases the adaptive subspace method indeed identifies the best model according to BIC. We illustrate the performance of the algorithm in the special design case of equal underlying correlation $c \in (0, 1)$ between the explanatory variables.

Let $p = 40$ and $n = 100$. We simulate $X = (X_{ij}) \in \mathbb{R}^{n \times p}$ with i-th row $X_{i\cdot} \sim \mathcal{N}_p(0, \Sigma)$, where $\Sigma_{kl} = c$ for $k \neq l$ and $\Sigma_{kk} = 1$ for $k = 1, \dots, p$. For each $c \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ we simulate 100 datasets in this way. Furthermore let $\beta_0 = (1, -1, 1, 2, -2, 2, 0, \dots, 0)^T \in \mathbb{R}^p$ be the true underlying vector of coefficients, so that the true active set is $S_0 = \{1, \dots, 6\}$. The response $Y = (Y_1, \dots, Y_n)^T$ is simulated via $Y_i \sim N(X_{i\cdot}\beta_0, \sigma^2)$ with variance $\sigma^2 = 4$. In the adaptive subspace method we initialize $q = 10$, $K = p$ and $T = 1000$. We use the "leaps and bounds" algorithm implemented in the R-package "leaps" (Lumley and Miller (2009)) to compute at iteration t the best model $S^{(t)}$ according to BIC contained in $V^{(t)}$.

The results of the adaptive subspace method are shown in Table 1, where we compare $\hat{S}_{0.9}$ (with threshold $\rho = 0.9$) and $\hat{S}_{\mathrm{b}}$ with the best possible BIC model S^* found by "leaps and bounds" when applied to the full dataset.

The adaptive subspace method identifies the best BIC model ($\hat{S}_{\mathrm{b}} = S^*$) in most of the cases. Moreover, we often have $\hat{S}_{0.9} = S^*$, while interestingly $\hat{S}_{0.9}$ selects on average less false positives than the best BIC model S^* itself (at the price of a slightly increased mean of false negatives, if c is large). The computational time of the adaptive subspace method, without aiming for optimal code, is a few seconds (max. 5 sec.) for each dataset, while full enumeration with "leaps and bounds" needs up to 1 min. for $c = 0.9$.

TABLE 1. Low-dimensional simulation study ($p = 40$ with 100 simulated datasets for each design with correlation c). Comparison of $\hat{S}_{0.9}$ and $\hat{S}_{\mathrm{b}}$ found by adaptive subspace method with best BIC model S^* in terms of mean numbers of false positives and false negatives, as well as the percentage of cases where they agree.

	mean false positives			mean false negatives			% of agreement	
c	$\hat{S}_{0.9}$	$\hat{S}_{\mathrm{b}}$	S^*	$\hat{S}_{0.9}$	$\hat{S}_{\mathrm{b}}$	S^*	$\hat{S}_{0.9} = S^*$	$\hat{S}_{\mathrm{b}} = S^*$
0.1	1.47	1.76	1.92	0.05	0.03	0.03	0.82	0.93
0.3	1.25	1.53	1.62	0.09	0.08	0.09	0.85	0.95
0.5	1.27	1.73	1.81	0.41	0.37	0.36	0.74	0.94
0.7	1.40	1.66	1.71	1.20	1.07	1.07	0.76	0.96
0.9	1.21	1.72	2.03	2.85	2.71	2.60	0.58	0.82

Furthermore, we have investigated sparse high-dimensional linear models via simulation and observe that in many situations the adaptive subspace method with EBIC performs very well in terms of finding the "true" underlying variables when we compare it to algorithms like stability selection with Lasso. The presentation of the results of an extensive simulation study comparing the proposed adaptive subspace method with different well-established algorithms is beyond the scope of this article.

As an illustration for the effectiveness of the adaptive subspace method, the evolution of the algorithm is displayed in Figure 1 for a high-dimensional dataset that is generated in the same way as above with moderate correlation $c = 0.3$ and with $p = 500$ instead of $p = 40$. We use the adaptive subspace method with EBIC (with constant $\gamma = 0.5$) as a selection criterion and initialize $q = 10$, $K = p$ and $T = 2000$. In this example, the adaptive subspace method yields $\hat{S}_{0.9} = \hat{S}_{\mathrm{b}} = \{1, 3, 4, 5, 6\}$, i.e. we have one false negative variable and no false positives. Figure 1 shows exemplarily the convergence of $r_1^{(t)}$ and $r_3^{(t)}$ against 1 and the convergence of $r_2^{(t)}$ and $r_7^{(t)}$ against 0. The convergence is very fast, with the algorithm taking 5.5 sec. for $T = 2000$ iterations.

4 Discussion

It is desirable to theoretically understand the limiting properties of the proposed adaptive subspace method and to find sufficient conditions for the

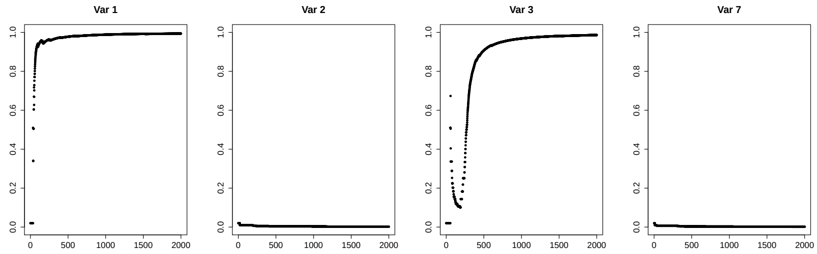


FIGURE 1. Adaptive subspace method for high-dimensional example ($p = 500$, correlation $c = 0.3$, true active set $S_0 = \{1, \ldots, 6\}$). Plot of the evolution of $r_1^{(t)}, r_2^{(t)}, r_3^{(t)}, r_7^{(t)}$ along the iterations t.

"correct convergence" of the algorithm, where by "correct convergence" it is meant that $r_j^{(t)} \to 1$ (a.s.) if $j \in S^*$ and $r_j^{(t)} \to 0$ (a.s.) if $j \notin S^*$, as $t \to \infty$. Recently, we have found a simple sufficient condition which ensures the correct convergence of the algorithm in the above sense. We are currently investigating, in which particular situations this condition is satisfied and how it might be relaxed in order to find weaker sufficient conditions. The theoretical details will be addressed in a subsequent paper. However, simulation studies indicate that even when such sufficient conditions do not hold, the algorithm provides useful information about the underlying variable selection problem and can be valuable as a heuristic algorithm for tracing well-fitted models.

References

Chen, J. and Chen, Z. (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, **95(3)**, 759–771.

Chen, J. and Chen, Z. (2012). Extended BIC for small–n–large–p sparse GLM. *Statistica Sinica*, **22(2)**, 555–574.

Lumley, T. and Miller, A. (2009). leaps: Regression Subset Selection. R package version 2.9, *http://CRAN.R-project.org/package=leaps.*

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Statist. Soc. B*, **72(4)**, 417–473.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc. B*, **58(1)**, 267–288.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**, 2541–2563.

A simple model for analyzing multi-tissue gene expression data

Anestis Touloumis

[1] University of Brighton, Brighton, U.K.

E-mail for correspondence: `A.Touloumis@brighton.ac.uk`

Abstract: In multiple-tissue experiments, gene expression is measured across multiple tissues for each subject. For each subject, the data measurements can be written as a matrix with the different multiple tissues indexing the columns and the genes indexing the rows. In this context, neither the genes nor the tissues are expected to be independent and straightforward application of traditional statistical methods that ignore this two-way dependence structure might lead to erroneous statistical conclusions. We present a non-parametric set of statistical tools for conducting inference in multi-tissue gene-expression studies.

Keywords: High-dimensional data; Hypothesis testing; Estimation; Gene expression data.

1 Introduction

Consider biological studies that use microarrays to study gene expression patterns in multiple tissue samples taken from the same subject (e.g., Melé *et al.*, 2015; Piccirillo *et al.*, 2015 and Sottoriva *et al.*, 2013). For each subject, the row variables correspond to genes, the column variables to tissue samples and the measurements are gene expression levels. A complex and high-dimensional dependence structure is expected to occur as neither the genes nor the tissue samples are likely to be independent. From a statistical perspective, it is extremely important to acknowledge both sources of correlation while keeping a parsimonious representation for the dependence structure so as to ease the key inferential purposes regarding the tissue- or gene- wise correlation pattern.
To this direction, the matrix-variate distribution can be utilized. We say that the random matrix $\mathbf{X}$ follows a matrix-variate normal distribution with matrix parameters $\mathbf{M}$, $\mathbf{\Sigma}_{\text{R}}$ and $\mathbf{\Sigma}_{\text{C}}$ if its vectorized form, $\text{vec}(\mathbf{X})$, follows a

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).

multivariate normal distribution with mean vector $\text{vec}(\mathbf{M})$ and covariance matrix the Kronecker product $\mathbf{\Sigma}_{\text{C}} \otimes \mathbf{\Sigma}_{\text{R}}$, where $\mathbf{\Sigma}_{\text{R}}$ and $\mathbf{\Sigma}_{\text{C}}$ are recognized as the row and column covariance matrix. Thus the covariance structure of the elements of a typical random matrix $\mathbf{X}$ that follows the matrix-variate normal distribution is given by the Kronecker product model:

$$\text{Cov}(X_{ij}, X_{lm}) = \mathbf{\Sigma}_{\text{R}il}\, \mathbf{\Sigma}_{\text{C}jm}.$$

Application of the matrix-variate normal model to multiple tissue gene experiments that measure gene expression, implies that the inference about the mean relationship of the genes across the tissues and about the dependence structure relies on estimating and/or testing hypotheses the mean matrix $\mathbf{M}$, the gene covariance matrix $\mathbf{\Sigma}_{\text{R}}$, and the tissue covariance matrix $\mathbf{\Sigma}_{\text{C}}$. In this context, the (a, b)-th element of $\mathbf{M}$ determines the mean expression level for gene a in tissue b, the (c, d)-th element of $\mathbf{\Sigma}_{\text{R}}$ the covariance of genes c and d, and the (e, f)-th element of $\mathbf{\Sigma}_{\text{C}}$ the covariance of tissues e and f.

Despite the nice interpretation, the normality assumption and the lack of statistical methods for estimating and hypothesis testing in high-dimensional settings could discourage researchers from employing the matrix-variate normal model. Herein, we present a non-parametric extension of this model that addresses these two challenges while maintaining the Kronecker product assumption about the dependence structure.

2 Inference based on a simple non-parametric model

Suppose that the gene expression levels for subject i are recorded in an $r \times c$ matrix $\mathbf{X}_i$ with rows the same set of r genes and columns the same set of c tissues. Let $\mathbf{X}_1, \dots, \mathbf{X}_N$ be a sample of N i.i.d. $r \times c$ random matrices generated by the non-parametric model

$$\mathbf{X}_i = \mathbf{\Sigma}_{\text{R}}^{1/2} \mathbf{Z}_i \mathbf{\Sigma}_{\text{C}}^{1/2} + \mathbf{M}, \tag{1}$$

where $\mathbf{M} = \text{E}[\mathbf{X}_i]$ is the $r \times c$ mean matrix, $\mathbf{\Sigma}_m = \mathbf{\Sigma}_m^{1/2} \mathbf{\Sigma}_m^{1/2}$ is a positive definite matrix ($m \in \{\text{R}, \text{C}\}$) and $\{\mathbf{Z}_i : i = 1, \dots, N\}$ is a sequence of i.i.d. $r \times c$ random matrices. The random variables $\{Z_{iab} : a = 1, \dots, r \text{ and } b = 1, \dots, c\}$ within $\mathbf{Z}_i$ are assumed to be independent with mean zero, unit variance and finite eighth moment. Model (1) includes the matrix-variate normal distribution as a special case obtained if Z_{iab} are i.i.d. $\text{N}(0, 1)$ random variables. Since $\text{cov}[\text{vec}(\mathbf{X}_i)] = \mathbf{\Sigma}_{\text{C}} \otimes \mathbf{\Sigma}_{\text{R}}$, we will refer to $\mathbf{\Sigma}_{\text{R}}$ and $\mathbf{\Sigma}_{\text{C}}$ as gene- and tissue- covariance matrix respectively.

To manage the high-dimensional setting, we assume that as $N \to \infty$ and $rc \to \infty$

$$\frac{\text{tr}(\mathbf{\Sigma}_{\text{R}}^4)}{\text{tr}^2(\mathbf{\Sigma}_{\text{R}}^2)} \to 0 \text{ and } \frac{\text{tr}(\mathbf{\Sigma}_{\text{C}}^4)}{\text{tr}^2(\mathbf{\Sigma}_{\text{C}}^2)} \to 0. \tag{2}$$

Assumption (2) does not specify the pairwise limiting ratios of the triplet (N, r, c), which allows us to consider data from multiple tissue gene expression studies. Options for $\mathbf{\Sigma}_{\mathrm{R}}$ and $\mathbf{\Sigma}_{\mathrm{C}}$ include covariance matrices with eigenvalues bounded away from 0 and ∞ or that have a few divergent eigenvalues as long as they diverge slowly, and thus the class of dependence structures under consideration is not seriously limited. Therefore, model (1) and assumption (2) constitute a flexible working framework to model multiple tissues gene expression studies in a non-parametric fashion while maintaing the same interpretation of the three matrix parameters as if the data were generated from a matrix-variate normal distribution.

2.1 Inference about the mean relationship

In multi-tissue studies, it is often of interest to identify differentially expressed genes. For example, one needs to assess whether the overall mean pattern of gene expression levels remains constant across all or pre-specified tissue groups. This amounts to testing the general hypothesis

$$\mathrm{H}_0 : \mathbf{M} = \mathbf{M}_0 \equiv [\mu_1 \mathbf{1}_{c_1}^{\mathrm{T}}, \mu_2 \mathbf{1}_{c_2}^{\mathrm{T}}, \dots, \mu_g \mathbf{1}_{c_g}^{\mathrm{T}}] \text{ vs. } \mathrm{H}_1 : \mathbf{M} \neq \mathbf{M}_0 \tag{3}$$

for known positive integers $c_1, \dots, c_g$ such that $\sum_{q=1}^{g} c_q = c$ with at least one $c_q \geq 2$ and for g (fixed but arbitrary) unknown parameter vectors $\mu_1, \dots, \mu_g$. For example, testing conservation of the mean gene expression across all tissues implies that $c_1 = c$. Touloumis *et al.* (2015) developed a statistical procedure for testing (3) which appeared to be more powerful that traditional ANOVA-type tests after applying multiple testing corrections. From a practical point of view, the null hypothesis in (3) would be dictated by the experiment's design. For example, in Piccirillo *et al.* (2015), one objectives was to investigate whether the mean gene-expression in glioblastoma patients was conserved in 5 different tumor fragments but varied from that in a normal tissue and from that taken from a subventricular zone tissue (that is $c_1 = 5$, $c_2 = c_3 = 1$).

The testing procedure proposed by Touloumis *et al.* (2015) can be also applied to a subset of genes or to the row mean vectors in order to obtain a more parsimonious form for $\mathbf{M}$ than the fully unstructured. Once the plausible form of $\mathbf{M}$ is induced, the mean relationship can be estimated by taking the corresponding sample analogue.

2.2 Inference about the covariance matrices

Estimation of $\mathbf{\Sigma}_{\mathrm{R}}$ and $\mathbf{\Sigma}_{\mathrm{C}}$ relies on shrinking approaches that extend the results of Touloumis (2015) for *vector*-valued random variables. Touloumis *et al.* (2016a) proposed the column covariance matrix estimator

$$\widehat{\mathbf{\Sigma}}_{\mathrm{C}} = (1 - \widehat{\lambda}_{\mathrm{C}}) \frac{1}{(N-1)r} \sum_{i=1}^{N} \mathbf{Y}_i^T \mathbf{Y}_i + \widehat{\lambda}_{\mathrm{C}} \widehat{\mu}_{\mathrm{C}} \mathbf{I}_c$$

and the row covariance matrix estimator

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{R}} = (1 - \widehat{\lambda}_{\mathrm{R}}) \frac{1}{(N-1)\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathrm{C}})} \sum_{i=1}^{N} \mathbf{Y}_i \mathbf{Y}_i^T + \widehat{\lambda}_{\mathrm{R}} \mathbf{I}_r$$

where $\mathbf{Y}_i = \mathbf{X}_i - \widehat{\mathbf{M}}$, $\mathbf{I}_k$ is the identity matrix of size k, and where the exact formulas for $0 \leq \widehat{\lambda}_{\mathrm{C}} \leq 1$, $0 \leq \widehat{\lambda}_{\mathrm{R}} \leq 1$ and $\widehat{\mu}_{\mathrm{C}}$ can be found in the Supplementary Material in Touloumis *et al.* (2016a). These shrinkage estimators are easy to calculate regardless the number of genes and tissues and are expected to be useful in the construction of relevance networks for genes and/or tissues (see Schäfer and Strimmer, 2005).
To study the correlation patterns of genes or tissues, Touloumis *et al.* (2016b) developed testing methodologies to assess whether known covariance structures are plausible ($\mathrm{H}_0 : \boldsymbol{\Sigma}_{\mathrm{R}} = \boldsymbol{\Sigma}$ or $\mathrm{H}_0 : \boldsymbol{\Sigma}_{\mathrm{C}} = \boldsymbol{\Sigma}$ for known $\boldsymbol{\Sigma}$) and to assess whether the genes or tissues are uncorrelated with the same variance but differing mean vectors ($\mathrm{H}_0 : \boldsymbol{\Sigma}_{\mathrm{R}} = \sigma^2 \mathbf{I}_r$ or $\mathrm{H}_0 : \boldsymbol{\Sigma}_{\mathrm{C}} = \sigma^2 \mathbf{I}_c$ for $\sigma^2 > 0$).

2.3 Software availability

The R package HDTD (Touloumis *et al.*, 2016a) implements the proposed methodologies. The user can estimate the mean matrix (`meanmat.hat`), the gene- and tissue- covariance matrix (`covmat.hat`) and conduct hypothesis testing for the mean matrix (`meanmat.ts`) and either of the two covariance matrices (`covmat.hat`).

3 Multiple tissue example

Melé *et al.* (2015) investigated variability in the human transcriptome across multiple tissues by analyzing RNA sequencing. This study identified, among other things, genes whose expression signature characterized particular tissues by using all available tissue-samples from each of the 175 individuals and comparing the gene expression levels of the tissue tested and that of the remaining tissues. This approach does not acknowledge the tissue-wise correlation and consequently, the discovery of tissue-specific gene lists might be hindered. To check this, we considered a subset of this dataset including only the subjects ($N = 11$) with available RNAseq samples across all the most frequently collected tissues (skin, nerve, adipose, artery, lung, skeletal muscle, heart, blood and thyroid). A $44,781 \times 9$ data matrix was created for each subject, with rows corresponding to genes, columns corresponding to the samples from the nine tissues and entries corresponding to the gene-expression levels. We focused on two important inferential aspects: i) study of the dependence structure among the nine tissues and ii) corroboration of the gene signatures when the dependence between tissues is accounted for.

TABLE 1. The estimated variances with a two letter abbreviations for the tissues.

SK	NE	AD	AR	LU	SM	HE	BL	TH
19123	16882	17523	17071	18244	27828	53881	757519	20445

To study the tissue specific variability and correlation pattern, we first estimated the corresponding covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathrm{C}}$. The estimated variances diplayed in Table 1 indicate that blood was by far the most variable tissue ($\widehat{SE} = 870.4$), with $\widehat{SE}$ at least four times that of the other tissues. We also observed that lung, heart, skeletal muscle and thyroid were mildly correlated with each other (see Table 2), while the remaining tissues showed weaker correlation. Although we rejected the spericity hypothesis $\mathrm{H}_0 : \mathbf{\Sigma}_{\mathrm{C}} = \sigma^2 \mathbf{I}_c$ (p-value<0.001) for all the tissues, we failed to reject the hypothesis that skin, adipose and lung tissue are uncorrelated with the same variation. There seems to exist a mild tissue-wise correlation which Melé *et al.* (2015) did not consider in their analysis.

Melé *et al.* (2015) generated lists of genes that showed tissue-specific expression. For a given tissue, we tested the hypotheses of conservation of the overall mean gene-expression levels of the corresponding genes-list between this tissue and any of the other eight, leading to a total of eight $p-$values, to which we applied an FDR correction. Failure to reject all hypotheses means that we do not have enough evidence these genes to be tissue-specific in their expression. After performing this analysis, we confirmed the validity of the tissue-specific gene-lists for skin, nerve, lung, skeletal muscle, heart and blood tissue. However, we failed to confirm that the overall mean gene-expression levels of the thyroid-specific gene-list is different than in the skeletal muscle (p-value $= 0.782$); that of adipose-specific gene-list different in the skin (p-value $= 0.105$), and that of artery-specific gene-list is simultaneously different from that of the skin, adipose and blood tissues (p-value$= 0.412$). The difference in our conclusions compared to those in Melé *et al.* (2015) presumably arises because the approaches used herein account for the presence of the tissue-wise correlation.

4 Discussion

In future works, we aim to extend the mean matrix hypothesis testing procedures to unbalanced designs i.e., when gene expression data are not collected across the same number of tissues for each subject, to develop a test statistic for assessing the Kronecker product assumption for the dependence structure and consider additional flexible but still parsimonious ways to model the dependence structure.

TABLE 2. The estimated tissue-wise correlation matrix.

	SK	NE	AD	AR	LU	SM	HE	BL	TH
SK	1.00	0.03	0.01	0.03	0.03	0.05	0.09	0.00	0.06
NE	0.03	1.00	0.03	0.05	0.04	0.09	0.09	0.00	0.06
AD	0.01	0.03	1.00	0.03	0.02	0.03	0.04	0.00	0.03
AR	0.03	0.05	0.03	1.00	0.05	0.09	0.08	0.00	0.06
LU	0.03	0.04	0.02	0.05	1.00	0.13	0.18	0.03	0.10
SM	0.05	0.09	0.03	0.09	0.13	1.00	0.34	0.00	0.15
HE	0.09	0.09	0.04	0.08	0.18	0.34	1.00	0.01	0.29
BL	0.00	0.00	0.00	0.00	0.03	0.00	0.01	1.00	0.04
TH	0.06	0.06	0.03	0.06	0.10	0.15	0.29	0.04	1.00

Acknowledgments: This research was based on collaborations with Prof. Simon Tavaré and Dr. John C. Marioni.

References

Melé *et al.* (2015). The human transcriptome across tissues and individuals. *Science* **348**, 660 – 665.

Piccirillo *et al.* (2015). Contributions to drug resistance in glioblastoma derived from malignant cells in the sub-ependymal zone. *Cancer Research* **75**, 194 – 202.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**, Article 32.

Sottoriva *et al.* (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *PNAS* **110**, 4009 – 4014.

Touloumis, A. (2015). Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Computational Statistics & Data Analysis*, **83**, 251 – 261.

Touloumis, A., Marioni, J.C. and Tavaré, S. (2016a). HDTD: Analyzing multi-tissue gene expression data. *To appear in Bioinformatics.*

Touloumis, A., Marioni, J.C. and Tavaré, S. (2016b). Hypothesis testing for a row or column covariance matrix in high-dimensional transposable data with a Kronecker product dependence structure. *Submitted.*

Touloumis, A., Tavaré, S. and Marioni, J.C. (2015). Testing the mean matrix in high-dimensional transposable data. *Biometrics*, **71**, 157 – 166.

Statistical shape analysis in a Bayesian framework for shapes in two and three dimensions.

Thomai Tsiftsi [1]

[1] University of Bath, Department of Mathematical Sciences, Bath, UK.

E-mail for correspondence: t.tsiftsi@bath.ac.uk

Abstract: In this paper, we describe a novel shape classification method which is embedded in the Bayesian paradigm. We discuss the modelling and the emerging shape classification algorithm for two and three dimensional data shapes. We conclude by evaluating the efficiency and efficacy of the proposed algorithm on the Kimia shape database for the two dimensional case.

Keywords: shape analysis; classification; planar shape model

1 Introduction

Shape is an important feature of objects; it can be used in many applications such as the recognition and classification of objects in images. In the approach we take we represent such objects and their boundaries as continuous planar curves (i.e. one-dimensional lines which denote the outline of the object) and study their shapes. Our goal is to develop shape models, statistical procedures and classification methods of continuous planar shape curves and establish the statistical framework needed for their classification. In particular, we study how to classify shapes that are generated by such curves and how we can probabilistically assign them into their respective categories; given a set of pre-determined classes we would like to classify the observed data shapes – we here define a **data shape** y to be one of the shapes that we observed i.e. an ordered set of points in $\mathbb{R}^2$ or $\mathbb{R}^3$. These questions occur in many applications of shape modelling and image analysis and thus are of broad interest.

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 Modelling and classification

The problem of classification can be mathematically formulated as the posterior probability of the class in question given the observed data; that is by $\mathbb{P}(C|\boldsymbol{y})$ where $C \in \mathcal{C}$ the set of all classes of the object and $\boldsymbol{y} \in Y$ the set of all the observed data shapes. In a Bayesian framework, classification is performed by maximising the posterior probability of the class which by Bayes' theorem is: $\mathbb{P}(C|\boldsymbol{y}) \propto \mathbb{P}(\boldsymbol{y}|C)\mathbb{P}(C)$. For simplicity we choose the prior $\mathbb{P}(C)$ over the classes to be uniform although it can be freely chosen. The major task is then to calculate the likelihood which we partition over nuisance parameters that correspond to the data formation process. For this, the model assumes that any data shape y has arisen by a representative shape curve $\beta \in \mathcal{B} \equiv R^{m\times n}/(R^m \times (R^+ \times SO(m)))$, which has been translated, scaled and rotated by $g \in G \equiv \mathbb{R}^m \ltimes (\mathbb{R}^+ \times SO(m))$ and has been sampled by a sampling function $s \in \mathcal{S}$. The inherent observational noise σ has perturbed the points from their original position and thus a bijection $b : [1, ...n] \to [1, ...n]$ compares the data points of y uniquely to the data points of β. The likelihood will be marginalised and finally be invariant to all the mentioned transformations. For our applications we choose the observational model to represent errors in shape point collection as additive Gaussian white noise so that the likelihood function for the complete data is given by:

$$\mathbb{P}(\boldsymbol{y}|C) = \sum_{b\in\mathcal{B}} \int \mathcal{D}\beta\ \mathcal{D}s\ \mathcal{D}g\ d\sigma\ \mathbb{P}(b)\mathbb{P}(s)\mathbb{P}(g)\mathbb{P}(\sigma)\mathbb{P}(\beta|C) \\ \times \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}|\boldsymbol{y_i} - g\circ\boldsymbol{\beta}(s(b_i^{-1}))|^2\right) \quad (1)$$

with a number of simplifying independence assumptions made. In order to estimate the posterior probability of a class, one should evaluate the sums and integrals over the nuisance parameters. In the next sections we discuss our computational strategies for dealing with these evaluations for both two and three dimensional shape data.

3 The two dimensional case

Our main goal is to evaluate the integrals in expression (1) and thus perform Maximum a Posteriori (MAP) classification to a certain class given the data. In previous work, *e.g.* Dryden and Mardia (1998) and Srivastava and Jermyn (2009) the integrations over the nuisance parameters were evaluated numerically by a zeroth order Laplace approximation. In Tsiftsi et al. (2014) we introduced an analytic way of carrying out the group integrations and the integrations over σ resulting in a closed form expression.

To achieve that, we made an appropriate and statistically significant choice of priors. Initially, we used Jeffreys' joint prior for $g \in G \equiv \mathbb{R}^2 \ltimes (\mathbb{R}^+ \times SO(2))$ and σ which preserves the invariance of the posterior under similarity transformations. However due to induced divergences a regularized version was employed. Although this broke the invariance of the original posterior, the result of this integration was found to be:

$$\mathbb{P}(y|b,\beta,s) = \frac{1}{Z} \sum_{b \in \mathcal{B}} \int \mathcal{D}\beta \; \mathcal{D}s \; \left[\tilde{n}\widetilde{\mathrm{Var}(\boldsymbol{y})} - \frac{\tilde{n}^2 \left|\widetilde{\mathrm{Cov}(\boldsymbol{v},\boldsymbol{y})}\right|^2}{\tilde{n}\widetilde{\mathrm{Var}(\boldsymbol{v})} + 1/B^2} + 2\zeta \right]^{-n-\alpha} \times \mathbb{P}(b)\mathbb{P}(s)\mathbb{P}(\beta|C) \tag{2}$$

where $\tilde{n} = \frac{nD}{n+D}$, whilst B, α, ζ, D are appropriate regulators, Z is the normalisation constant, n the number of sample points and $\widetilde{\mathrm{Cov}(\boldsymbol{v},\boldsymbol{y})} = \frac{1}{\tilde{n}}\left[\sum_i v_i \bar{y}_i - \frac{1}{\tilde{n}}\sum_i \sum_j v_i \bar{y}_j\right]$. For details regarding the priors and the calculations refer to Tsiftsi et al. (2014). The proposed algorithm returns high classification rates. To demonstrate its ability and power we present an example on two shape databases in section 5.

4 The three dimensional case

Another problem of interest is the generalisation of the previous case to its three dimensional equivalent by assuming that $y \in \mathbb{R}^3$. The three dimensional case is treated in a similar way as the two dimensional case: our goal is to classify a shape by performing MAP on the class C. We follow the same steps as in the two-dimensional case and we marginalise the likelihood over the nuisance parameters that take part in the data formation process; however similarity transformations are now represented by $g \in G \equiv \mathbb{R}^3 \ltimes (\mathbb{R}^+ \times SO(3))$ since $y \in \mathbb{R}^3$.

The challenge is the analytic evaluation of the integrals of the marginalised likelihood and especially the integration over three-dimensional rotations. Initially, translations were integrated against the uniform Haar measure. The result, as expected, was analogous to the two dimensional case and had to be integrated with respect to rotations. For this integration, we chose to represent rotations as unit quaternions; the full quaternionic space is described by: $\mathbb{H} = \{a + bi + cj + dka, b, c, d \in R\}$ with i, j, k the three special unit imaginary quaternions. The basis quaternions anti-commute and they provide a representation of $SU(2)$.

For the integration over quaternions, we choose to integrate over the full quaternionic space $\mathbb{R}^4$, imposing the constraint $\delta(|q|^2 - 1)$ which restricts us to unit quaternions that live on the surface of the unit 4-sphere. This δ-function is invariant under the action of $SU(2)$ on the parameters since rotations act by isometries. Carrying out this integration one has:

$$P(y|b,\beta,s,\sigma)=\int d^4q\,\delta(|q|^2-1)\exp\left(\frac{|\sum_i^N \boldsymbol{Y}_i|^2}{2n\sigma^2}-\frac{\sum_i^N |\boldsymbol{Y}_i|^2}{2\sigma^2}\right) \tag{3}$$

where $\boldsymbol{Y} = y - q\beta q^*$ with q a unit quaternion and q^* its quaternionic conjugate. To perform the integration, we represent the δ-function via its Fourier transformation which introduces a second integration that can be simplified to:

$$\begin{aligned} P(y|b,\beta,s,\sigma) &= \iint dk\,d^4q \exp\left(ik(|q|^2-1)\right)\exp\left(\frac{|\sum_i^N \boldsymbol{Y}_i|^2}{2n\sigma^2}-\frac{\sum_i^N |\boldsymbol{Y}_i|^2}{2\sigma^2}\right) \\ &= \frac{1}{2\pi}\iint dk\; d^4q\;\; \exp\left(-ik\right)\exp\left(-4n\;[q^TM(k)q]\right) \end{aligned} \tag{4}$$

where

$$M_{ij}(k)=-ik\delta_{ij}-\delta_{0i}(\overline{\underline{\hat{\boldsymbol{v}}}^T\times\underline{\hat{\boldsymbol{y}}}})_i-(1-\delta_{0j})(1-\delta_{0i})\left[(\overline{\underline{\hat{\boldsymbol{y}}}\otimes\underline{\hat{\boldsymbol{v}}}})_{ij}-\delta_{ij}(\overline{\underline{\hat{\boldsymbol{v}}}^T\times\underline{\hat{\boldsymbol{y}}}})\right]$$

is the symmetrised, positive definite 4 × 4 covariance matrix of the q components. We now discuss the integral over the quaternionic parameters that generate the $SO(3)$ rotations. We followed Wood (1993) and calculated the appropriate Haar measure for the quaternionic representation which was proven to be related to the Bingham distribution. The result of integrating over k will supply the Haar measure on the space of unit quaternions and restrict our parameters q to this surface. Indeed, we could rewrite the integration over rotations by diagonalising M to give:

$$\int_{SO(3)} \exp\left(-4n\;[q^TM(0)q]\right) = \int_{S^3} \exp\left(-4n\;\sum_i \lambda_i\tilde{q}_i^2\right) d[\tilde{q}] \tag{5}$$

where λ_i the eigenvalues of the covariance matrix $M(0)$. Here, the $\tilde{q}_i$ generate rotations in $SO(3)$ which are uniformly distributed **if and only if** the $\tilde{q}_i$ are uniform on a unit hemisphere in $\mathbb{R}^4$ so that the usual uniform measure on S^3 for $d[\tilde{q}]$ induces the Haar measure on the space of rotations. This ensures that the chosen measure in (4) is the appropriate one that favours no particular rotation over another.

Returning to expression (4), it is easy to see that the integral with respect to q refers to a multivariate Gaussian distribution. Assuming that the eigenvalues of matrix M are positive the evaluation of the quaternionic integral of this multivariate Gaussian distribution is:

$$\begin{aligned} P(y|b,\beta,s,\sigma) &= \frac{1}{2\pi}\iint d^4q\ dk\ \ \exp\left(-ik\right)\exp\left(4n\ [q^TM(k)q]\right) \\ &\propto \frac{1}{2\pi}\int dk\ \ \exp\left(-ik\right)\frac{4\ n\ \pi^2}{\sqrt{\det(\mathrm{M})}} \end{aligned} \tag{6}$$

with det(M) the determinant of matrix M which has k dependence, the result of which integration is invariant to rotations of y since it has been written in a manifestly rotationally invariant way. One may hope to evaluate integrals of this form by contour integration. For this we would have to promote k to the complex plane and choose an appropriate path in the k-plane. Since the expression of the determinant is not a perfect square, the presence of the square root in the denominator implies that the roots of the determinant produce branch cuts and we were thus unable to compute the integral over k analytically.

We were therefore forced to expand the square root in the denominator of expression (6) in order to analytically approximate the integral. However, this represents an important step towards generalising our work on planar shapes to three-dimensional curves. The calculations of the integration of the remaining nuisance parameters are challenging, although positive developments have been made towards a series solution. We leave the remaining calculation for future consideration as an extension of the analysis presented here. This work is still in progress but shows promising signs of improving upon the current shape classification methods in three-dimensions.

5 Example in two-dimensions

In order to test and verify the algorithm's efficacy on the classification of two-dimensional data shapes examples from two shape databases were considered: the Kimia and a simulated letter database. In the latter case the application of our algorithm comes with a warning; ordinarily the orientation of letters is crucial (for example W versus M and C versus U) whereas our likelihood has been constructed to be invariant under rotations of the data. The tests on this database should therefore be understood as a general test of our algorithm which is used for demonstrational purposes.

Both databases were comprised of binary images which were used for training and testing purposes. The shapes' boundaries were extracted in MATLAB and simulated shapes played the role of the observed data sets. The proposed algorithm was tested on the simulated data sets and its classification results are very positive, as is illustrated in Figure 1.

For the Kimia database we found that for 10 simulations of 10 shapes each, the average classification level was $\hat{\mu} = 59\% \pm 7\%$ with the average success rate being $\hat{\mu} = 80\% \pm 5\%$. From these experiments we concluded that the number of sampled points is crucial since as soon as the number of points increases to more than 50 the confidence levels become almost 90 percent. For the alphabet database, the results for the average classification level were $\hat{\mu} = 77\% \pm 5\%$ with the average success rate $\hat{\mu} = 73\% \pm 6\%$. The evaluation of the performance of the algorithm in three-dimensions could be tested by using examples from 3D geological sand formations as previously discussed in Tsiftsi et al (2014).

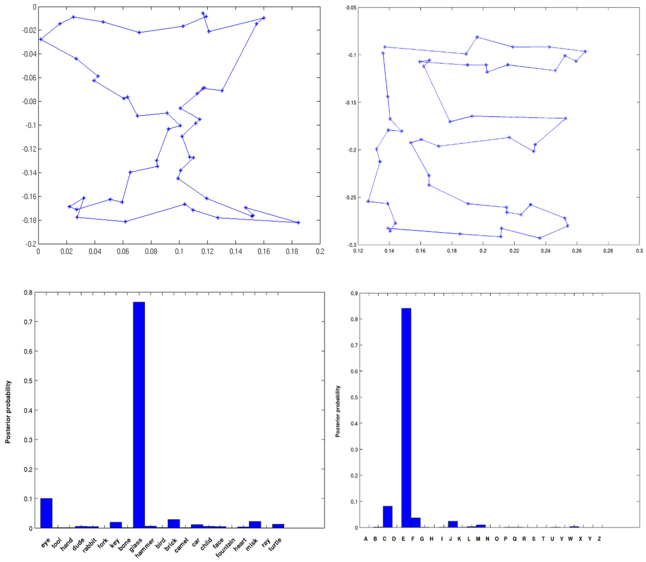


FIGURE 1. Classification results for a Kimia shape and the letter E.

References

Dryden, I.L. and Mardia, K (1998). *Statistical shape analysis.* J. Wiley.

Srivastava, A. and Jermyn, I.H. (2009), Looking for shapes in 2D cluttered point clouds, *IEEE Trans. Patt. Anal. Mach. Intell.*, **31(9)**, 1616–1629.

Tsiftsi, T. , Jermyn, I. and Einbeck, J. (2014) Bayesian shape modelling of cross-sectional geological data, *in 29th International Workshop on Statistical Modelling, 14-18 July 2014, Goettingen, Germany; proceedings*, Amsterdam: Statistical Modelling Society, 161–164.

Wood, A.T.A. (1993), Estimation of the concentration parameters of the Fisher matrix distribution on $SO(3)$ and the Bingham distribution on $S_q, q \geqslant 2$, *Australian Journal of Statistics*, **35 (1)**, 69–79.

Boosting Joint Models for Longitudinal Data and Survival Times

Elisabeth Waldmann[1], David Taylor-Robinson[2], Thomas Kneib[3], Matthias Schmid[4], Andreas Mayr[1,4]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] University of Liverpool, England
[3] Georg-August-Universität Göttingen, Germany
[4] Rheinische Friedrich-Wilhelms-Universität Bonn, Germany

E-mail for correspondence: elisabeth.waldmann@fau.de

Abstract: Joint Models for Longitudinal and Time-to-Event (JM) data have gained a lot of attention in the last few years as they are a helpful technique to approach a data structure very common in life sciences. Commonly JMs are estimated in likelihood based expectation maximization approaches or in a Bayesian framework. In this paper we propose a boosting algorithm to simultaneously estimate and select predictors for JMs.

Keywords: Joint Modeling; Survival Analysis; Longitudinal Data; Boosting.

1 Introduction

Joint Models (JM) as a term has been used in various contexts to describe modeling of a combination of different outcomes. This contribution deals with JMs for longitudinal and survival model outcomes like investigated by Rizopolous (2012). We stumbled across this problem when analysis lung function decline in cystic fibrosis patients. It has been shown that the onset of pulmonary infections implies the acceleration of the loss of lung function, when using the onset of infection as a covariate in a longitudinal model (Qvist et al. 2015). This onset however, could be seen as a process influenced by the same covariates as the lung function decline itself and hence should be modeled as a related yet separate process. Since the standard approaches to JM do not include variable selection, we suggest a boosting algorithm to tackle this problem. Boosting as statistical estimation approach has gained much attention since it is not only able to estimate a

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).

range of different effects but also includes shrinkage and variable selection. For an introduction and overview see Mayr et al. (2014).

2 Methods

2.1 Joint Modeling

The type of JM we are presenting is composed of two parts: the longitudinal and the survival outcome. In the following we will describe predictors and associated likelihoods. The predictor for the longitudinal outcome y_{it} is divided into two parts

$$y_{it} = \eta_{li}(t) + \eta_{lsi}(t) + \varepsilon_{it},$$

where $i = (1, \dots, n)$ refers to the i-th individual, $t = (t_1, \dots, t_{n_i})$ to the time of the observation and ε_{it} is the model error, which is assumed to follow a normal distribution. The two functions $\eta_{li}(t)$ and $\eta_{lsi}(t)$, which will be referred to as the longitudinal and the shared predictor in the following, are functions of two separate sets of covariates. In the boosting context they are constructed as sums of base learners, which are functions of potentially influential variables x_i and the time t. The shared predictor $\eta_{lsi}(t)$ reappears in the survival part of the model:

$$\lambda(t|\alpha, \eta_{lsi}(t)) = \lambda_0(t)\exp(\alpha\eta_{lsi}(t)),$$

where $\lambda_0(t)$ is the baseline hazard which will be chosen to be constant $(\lambda_0(t) = \lambda_0)$. The predictor with subscript ls refers to both, longitudinal and survival part of the model, whereas we assume that η_l only has an impact on the longitudinal structure. The relation between both parts of the models is quantified by the association parameter α. Those two models can be summarized in one likelihood:

$$\prod_{i=1}^{n}\left[\int_{-\infty}^{\infty}\left\{\prod_{j=1}^{n_i} f(z_{ij}|\eta_{li}(t), \eta_{lsi}(t), \sigma_\epsilon^2)\right\} f(T_i, \delta_i|\eta_{lsi}(T_i), \alpha)\right],$$

where T_i is the time of event for individual i and where the distribution for the longitudinal part the Gaussian. The likelihood for the survival part is:

$$f(T_i, \delta_i|\alpha, \eta_{lsi}(T_i), \lambda_0) =$$
$$[\lambda_0(T_i)\exp(\alpha(\eta_{lsi}(T_i)))]^{\delta_i}\exp\left[-\int_0^{T_i}\lambda_0(u)\exp(\alpha\eta_{lsi}(u))du\right].$$

Here, δ is the censoring indicator, taking the value 0 in the case of censoring and 1 in the case of an event.

2.2 Boosting Joint Models

The above described situation differs in one key aspect from the problems solved by boosting traditionally. The predictors for the different dependent variables are neither entirely different nor completely identical. One includes a function of the other. We suggest an update scheme at predictor stage rather than at the level of the dependent variables. We hence need an outer loop including three steps:

(step1) updating $\eta_l(t)$ in a boosting iteration

(step2) updating $\eta_{ls}(t)$ in a boosting iteration

(step3) updating α and λ by maximizing the likelihood.

3 Simulation

3.1 Setup

Three different simulation setups were constructed. Two of them included 20 covariates per predictor, of which only two were informative. As an additional parameter a time impact was entered into the shared predictor. The association parameter was set to 0.5 in the first setup (S1) and to -0.5 in the second setup (S2). The third setup (S3) was constructed to mimic the data situation more closely and to thus get a better sensation for the ability of the algorithm to do variable selection in a lower dimension. S3 had only four non informative covariates in each predictor but did not differ in any other features from S1. All three setups include random intercept and slope:

$$\boldsymbol{\eta}_l = \boldsymbol{X}_l \boldsymbol{\beta}_l \quad \text{and} \quad \boldsymbol{\eta}_{ls} = \boldsymbol{X}_{ls} \boldsymbol{\beta}_{ls} + \beta_t \boldsymbol{t} + \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_1 \boldsymbol{t}$$

The matrices $\boldsymbol{X}_l$ and $\boldsymbol{X}_{ls}$ are the collections of the standardised covariates, $\beta_l = (2, 1, -2)^\top$ and $\beta_{ls} = (1, -2)^\top$ the corresponding linear effects, $\beta_t = 1$ is the impact of time $\boldsymbol{t}$, γ_0 the random intercept and γ_1 the random slope. The time points were drawn in a way that mimics yearly examinations. The longitudinal outcome $\boldsymbol{y}$ was drawn from a Gaussian distribution with mean $\eta_l + \eta_{ls}$. The survival times were simulated based on the joint predictor multiplied by the association parameter $\alpha\eta_{ls}$ and the survival probabilities resulting from the above described hazard function. Stopping parameters for the boosting algorithm were selected by a tenfold cross-validation on a ten times ten grid for the two sub-predictors.

Setup	β_l(sd)	$\beta_{ls}(sd)$	P($\beta_l = 0$) (sd)	P($\beta_{ls} = 0$) (sd)
S1	β_0=1.982(0.116) 0.994(0.009) -1.995(0.009)	1.000(0.048) -2.000(0.053) β_t=1.044(0.267)	0.643(0.005)	0.782(0.015)
S2	β_0=1.977(0.080) 0.993(0.010) -1.994(0.011)	1.001(0.048) -2.002(0.052) β_t=1.058(0.173)	0.435(0.039)	0.655(0.047)
S3	β_0=2.061(0.102) 0.995(0.010) -1.996(0.009)	0.995(0.048) -1.984(0.048) β_t=0.854(0.233)	0.631(0.05)	0.897(0.044)

TABLE 1. Estimations and selection proportions of the parameters in the three setups. The estimation of the informative parameters are displayed individually, the non informative parameters in an overall average. Also selection probability is displayed in an overall average for the non informative parameters.

3.2 Results

Our algorithm performs good on our simulation studies and the predictors $\boldsymbol{\eta}_l$ and $\boldsymbol{\eta}_{ls}$ are captured just as well as in standard methods. The informative parameters were selected in 100% of the runs in both sub predictors in all three setups. Non informative parameters were selected in more than 50% but less than 75% in all setups, for the exact proportions see TABLE 1. The exact values of β_l, β_{ls} and β_t are estimated very well, for an overview see TABLE 1 and FIGURE 1 for the exemplary display of the results from S1. The estimation of the association parameter was less accurate for the S2, than for S1 and S3, yet still close (see FIGURE 2).

4 Cystic Fibrosis

After choosing the patients that have at least two observations, before the infection the data set contained a total of 6268 of 489 patients of which 53 were infected with PA in the course of the study. The covariates for the longitudinal predictor were height and weight of the patient as well as three binary covariates indicating, if the patient had one of three different additional lung infections. The covariates possibly having an impact on the shared part of the model were time, pancreatic insufficiency, sex, age at which CF was diagnosed and birth year. The stopping iterations ($m_l = 1100$ and $m_{ls} = 30$) were chosen based on tenfold cross validation. All parameters in the longitudinal predictor were selected. In the shared predictor birth year, time and pancreatic insufficiency were chosen as informative covariates. The association parameter $\alpha = -0.380$, the longitudinal process is hence having a negative impact on the risk of being infected.

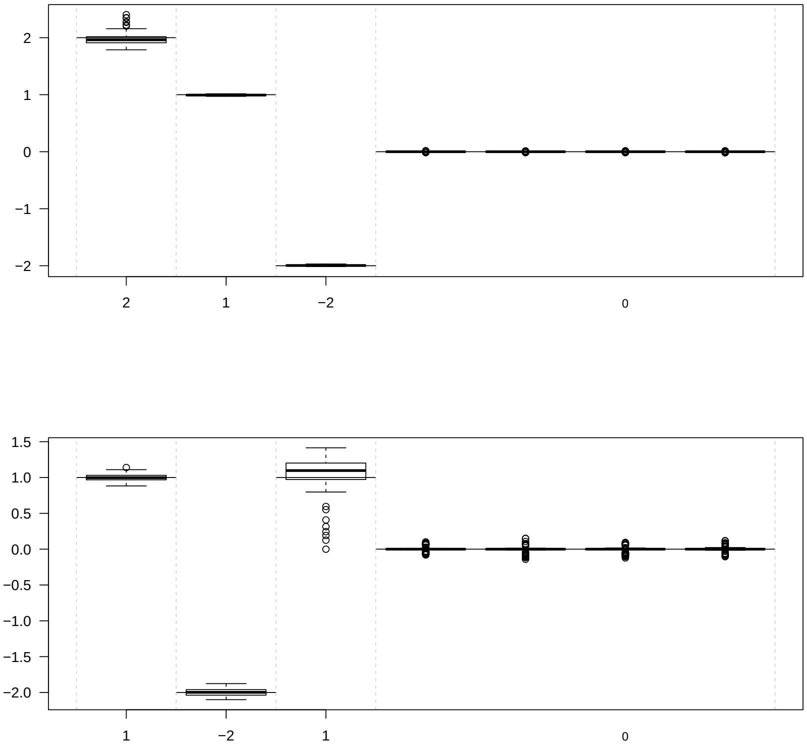


FIGURE 1. Boxplot of the parameter estimates for S1

5 Conclusion and Outlook

The presented approach is to our knowledge the first variable selection algorithm implemented in the joint modelling framework and hence an important step especially for prediction. Simulations show a tendency to incorporate also non informative variables with effects close to zero which could be by stability selection. Furthermore we plan to complete the model by incorporating a predictor η_s, i.e. including covariates which only have an impact on the survival time and are independent of the longitudinal structure. Once this is incorporated, we can make even better use of the features of boosting and implement variable selection and allocation between the predictors (i.e. we give the same variables to all three predictors η_l, η_s and η_{ls} and let boosting decide, to which predictor the covariates belong).

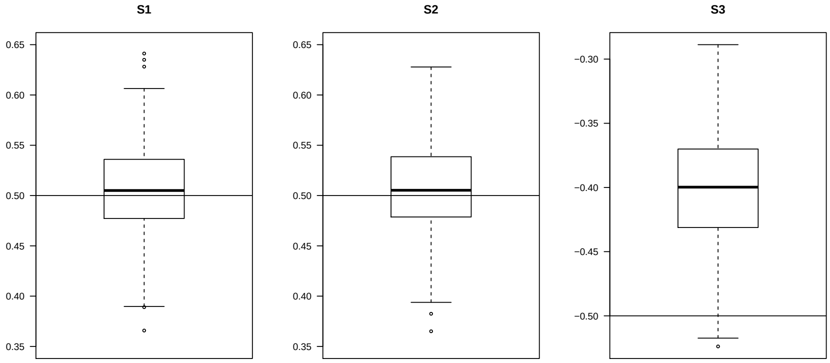


FIGURE 2. Boxplot of the estimates of α over all three setups

References

Mayr, A., Binder, H., Gefeller, O. and Schmid, M. (2014). The Evolution of Boosting Algorithms - From Machine Learning to Statistical Modelling. *Methods of Information in Medicine*, 53(6): 419−427.

Qvist, T., Taylor-Robinson, D., Waldmann, E., Olesen, H., Hansen, C., Mathiesen, I., Hoiby, N., Katzenstein, T., Smyth, RL, Diggle, P. and Pressler, T. (2015). Comparing the harmful effects of nontuberculous mycobacteria and Gram negative bacteria on lung function in patients with cystic fibrosis. *Journal of Cystic Fibrosis.* (In Press)

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data.* London: Chapman & Hall /CRC Biostatistics Series.

Projections of health indicators for chronic disease: illness death model and semi-Markov assumption.

Mathilde Wanneveich[12], Hélène Jacqmin-Gadda[12],
Jean-François Dartigues[12], Pierre Joly[12]

[1] University of Bordeaux, ISPED, Centre Inserm U1219, Bordeaux FRANCE
[2] Inserm, ISPED, Centre Inserm U1219, Bordeaux FRANCE

E-mail for correspondence: mathilde.wanneveich@isped.u-bordeaux2.fr

Abstract: Chronic diseases are a growing public health problem and their economic, social and demographic burden is alarming in years to come. Up to now, the method used to make projections and assess the future disease burden assumes a non-homogeneous Markov assumption in an illness-death model. Both age and calendar time have been taken into account in all parameter estimations, nevertheless the time spent with the disease was not considered. This work develops the method with a semi-Markov assumption to model mortality among the diseased and considering the time spent with the disease. The method is applied to estimate several health indicators for dementia in France in 2030.

Keywords: Multi-states model, Semi-Markov, Projection, Chronic disease.

1 Introduction

Currently, the high prevalence of some chronic diseases is a serious public health problem (Brayne 2007). Furthermore, population ageing foreshadows increases in their burden in the coming years. Dementia like Alzheimer's disease is one of these diseases of concern. Indeed, Jacqmin-Gadda et al. (2013) suggest that the prevalence of dementia will reach 1.700 million in 2030 in France (rise of 75% in 20 years), and Brookemeyer et al. (1998) project a 3.7 fold increase for the United State population between 1997 and 2047.

Up to now, multistate models based on Markov processes have been a well-established method for modeling incidence and mortality for dementia (Joly 2002), and macro-simulation is often used to provide health indicators

projections as prevalence (Hebert 2003; Wanneveich 2016). The transition probabilities used in these models may depend on age and calendar time. However, for many chronic diseases, mortality is likely to rise with disease duration.
In this work, we propose an alternative to the model developed by Joly et al.(2013) by introducing a semi-Markov approach to model the mortality among diseased subjects and thus, to consider the time spent with the disease to provide projections of several health indicators.

2 Methods

2.1 Illness-death model & semi Markov assumption

Figure 1 represents a three-state model called illness-death.

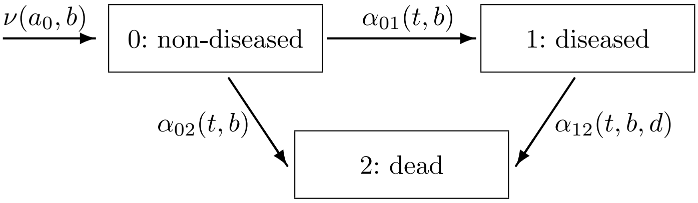


FIGURE 1. The illness-death model.

Initially all subjects are non-diseased (state "0"), then either they die (state "2") or they become diseased (state "1") and then die. Thus, the transition intensity α_{01}, is interpreted as the incidence rate, and α_{02} and α_{12} represent the mortality among non-diseased and diseased subjects respectively. These transition intensities depend on t the calendar time and b the year of birth (so $t-b$ is the age at time t) but α_{12} also depends on d the time spent with the disease. Also, as input to the model, we define $\nu(a_0, b)$ the size of the population at risk to develop the disease at age a_0 and born in b. Finally, $\alpha_2(t, b)$ denotes the overall mortality rate at time t for subjects born in b. We focus on incurable chronic diseases and thus assume no reversible transition from diseased to non-diseased.

2.2 Assumptions

The method is based on several assumptions. First, the incidence is supposed null before age a_0 (the value of a_0 depends of the disease). Then, for mortality among non-diseased subjects, in most cases it is relevant to distinguish it from the overall mortality: $\alpha_{02}(t, b) \neq \alpha_2(t, b)$, especially for diseases with high prevalence and high mortality. Lastly, for mortality among diseased subjects, we propose an additive model that takes into account age and time spent with the disease. Then, we write it as the sum of mortality among non-diseased subjects added to an over-risk of death depending on the time spent with the disease:

$$\alpha_{12}(t,b,d) = \alpha_{02}(t,b) + \alpha^d(d) \tag{1}$$

2.3 Estimation of transition intensities

Cohort data are used to estimate the functions $\alpha_{01}(t,b)$ and $\alpha^d(d)$ by fitting a non parametric illness death model with a penalized likelihood approach and spline approximation for each function. This method handles semi-competing risks of disease and death and interval censoring of age at disease onset. Demographic national projections provide values of $\alpha_2(t,b)$ and $\nu(a_0,b)$, that will allow to consider the time trends for each mortality. Then, α_{02} is estimated by solving a differential equation of the type: $y'(s) = c(s) - e^{-y(s)}a(s)$, involving α_2, α_{01} and α^d. The resolution (for each year of birth b) is analytic and allows to express mortality among non diseased subjects with known quantities. Finally, α_{12} is estimated using the equation (1).

2.4 Health indicators

Once the transition intensities have been estimated, we use the cumulative transition intensities (between two ages) A_{01}, A_{02}, A_{12} to calculate probabilities (by year of birth b):

- P_{00}, the probability for subjects born in b and alive at age a_0 to be alive and non-diseased at age $t-b$:

$$P_{00}(a_0, t-b|b) = e^{-A_{01}(a_0,t-b|b)-A_{02}(a_0,t-b|b)}$$

- P_{11}, the probability for subjects born in b and diseased at age a_d (with $a_d > a_0$) to be alive at age $t-b$:

$$P_{11}(a_d, t-b|b) = e^{-A_{12}(a_d,t-b|b,d)}$$

- P_{01}, the probability for subjects born in b, alive and non diseased at age a_0 to be alive and diseased at age $t-b$:

$$P_{01}(a_0, t-b|b) = \int_{a_0}^{t-b} e^{-A_{01}(a_0,u|b)-A_{02}(a_0,u|b)}\alpha_{01}(u|b)e^{-A_{12}(u,t-b|b,d)}du$$

Then, these probabilities are used to calculate several relevant health indicators for a given time t for projections (by year of birth b). We note:

- $LE_{00}(x|b)$, *the life expectancy without the disease* at age x, which is the remaining number of years that someone non-diseased at x can expect to live without the disease.
- $LE_{..}(x|b)$, *the overall life expectancy* which is the weighted mean of life expectancy for diseased and non-diseased subjects at age x.
- $F_{01}(x|b)$,*the life-long probability of the disease* which is the overall risk of developing the disease before death for subjects of age x alive and non-diseased.
- $Prev(a_0|t)$, *the prevalence of the disease*, which is the number of diseased subject between a_0 and 99 at time t.

3 Application on dementia in France

The French cohort PAQUID initiated in 1988 to study the aging population, allows to estimate the dementia incidence (α_{01}) and the over-risk of death for demented subjects (α^d). The sample consists of 3777 subjects aged 65 years and older and it is representative of the French population in terms of age and gender (Dartigues 1991). Based on this data, we hypothesized that the incidence is homogeneous over calendar time (it depends only on age) and null before $a_0 = 65$. The French National Institute of Statistics, INSEE, provides French demographic projections, including the age- and sex-specific overall mortality (α_2), and the population alive (ν) at age 65 by gender for each year of birth. Lastly, mortality among the non-demented (α_{02}) depending on age and calendar time is computed by analytic resolution of a differential equation for each years of birth b. Then, mortality among the demented (α_{12}) may be estimated.

Results, in particular on mortality among the demented depending on age and time spent with dementia are presented as well as the changes of the health indicators between 2015 and 2030. The discussion concern the choice of the model and the perspectives.

References

Brayne, C. (2007). The elephant in the room - healthy brains in later life, epidemiology and public health. *Nature Review Neuroscience*, **8**, 233-239.

Brookmeyer, R., Gray, S., Kawas, C. (1998). Forecasting the global burden of Alzheimer's disease. *American journal of public health*, **88**, 1337-1342.

Dartigues, J.F., Gagnon, M., Michel, P., et al. (1991). Le programme de recherche Paquid sur l'épidémiologie de la démence. Méthodologie et résultats initiaux. *Revue Neurologique*, **147**, 225-230.

Hebert, L.E., Scherr, P.A., Bienias, J.L., Bennett, D.A., Evans D.A. (2003). Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology*, **60**, 1119-1122.

Jacqmin-Gadda, H., Alpérovitch, A., Montlathuc, C., et al. (2013). 20-years prevalence projections for dementia and impact of preventive policy about risk factors. *European Journal of Epidemiology*, **28**, 493-502.

Joly, P., Commenges, D., Helmer, C., Letenneur, L. (2002). A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*, **3**, 433-443.

Joly, P., Touraine, C., Georget, A., Dartigues, J.F., Commenges, D. et al. (2013). Prevalence projections of chronic disease and impact of public health intervention. *Biometrics*, **69**, 109-117.

Wanneveich, M., Jacqmin-Gadda, H., Dartigues, J.F., Joly, P. (2016). Impact of intervention targeting risk factors on chronic disease burden. *Statistical Methods in Medical Research*, in press.

On statistical testing and mean parameter estimation for zero–modification in count data regression

Paul Wilson[1], Jochen Einbeck[2]

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

[2] Department of Mathematical Sciences, Durham University, UK

E-mail for correspondence: `pauljwilson@wlv.ac.uk`

Abstract: For the problem of testing for zero–modification in Poisson regression, a simple and intuitive test can be constructed by computing directly confidence intervals for the number of 0's under the Poisson assumption. This requires the ability of estimating the mean function accurately even if the data are in fact zero–inflated or deflated. A novel hybrid estimator is introduced for this purpose, which is of interest beyond the scope of the motivating test problem.

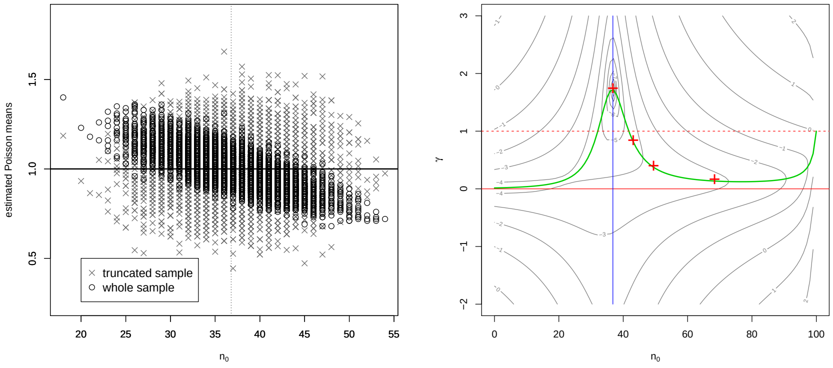
Keywords: Zero–modification; zero–truncated model; hypothesis testing

1 Introduction

Commonly used tests for zero–inflation/modification are likelihood ratio, score and Wald tests. Whilst these tests are all viable, they are not readily understood by non–statisticians, they do not distinguish between zero–inflation and zero–deflation (at least, not without adjustments), and they rely upon asymptotic results. Wilson and Einbeck (2015) proposed a new family of tests to test zero–modification in count data regression. Consider data $(y_i, \boldsymbol{x}_i^T), i = 1, \ldots, n$, where y_i are discrete counts and $\boldsymbol{x}_i \in \mathbb{R}^d$ a predictor vector. Let $p_i = P(y_i = 0)$. In the special case of (possibly zero–modified) Poisson regression, this test can be summarized as follows. For given significance level α: (i) fit the Poisson regression model, yielding Poisson means $\hat{\mu}_i = \hat{E}[y_i|\boldsymbol{x}_i]$; (ii) for each each y_i estimate $\hat{p}_i = \exp(-\hat{\mu}_i)$; (iii) use a Poisson–Binomial distribution with parameters $(n, \hat{p}_1, \ldots, \hat{p}_n)$ to determine a $1-\alpha$ confidence interval for the number of 0's.

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).

FIGURE 1. Left: Estimation from the zero-truncated and whole sample; right: Function $\gamma^*_{100}(n_0, 1)$ (thick curve) with $MSE(T|n_0)$ contours. In both plots, $\mu = 1$ and $n = 100$.



The challenging part in this procedure is the estimation of the Poisson means $\mu_i = E[y_i|\boldsymbol{x}_i]$ in the absence of the knowledge whether the Poisson assumption is correct. This problem has attracted attention earlier; Dietz and Böhning (2001) observed that ML estimation of the zero–modified Poisson model can be obtained by ML estimation of the zero–truncated Poisson (ZTP) model. For additional insight, consider Figure 1 (left), which shows the estimates of the Poisson means obtained when $n = 100$ observations are sampled from a Pois(1) distribution. The black circles indicate whole sample mean (Poisson) estimates $\hat{\mu}_P$, and the grey crosses the means $\hat{\mu}_T$ obtained from the positive observations. The horizontal axis gives the number of zeros, n_0, with the expected number of zeros under the Poisson model, $100e^{-1} \approx 37$, highlighted by a dotted line. It is clear that the Poisson estimator has smaller variance but is possibly biased if the observed number of zeros is far from their expected number. On the other hand, the ZTP–derived mean estimator does not demonstrate a noticeable bias, at the expense of a large variance.

2 A hybrid mean estimator

The illustrated bias–variance trade–off motivates the definition of the hybrid estimator

$$T = \gamma\hat{\mu}_P + (1 - \gamma)\hat{\mu}_T \qquad (1)$$

which is a weighted sum of the usual Poisson mean estimator $\hat{\mu}_P$ and an estimator of the zero–truncated mean, $\hat{\mu}_T$. The latter is based on the mean of the zero–truncated data only, to which we refer from now on as ζ. Note that the mean μ of a Poisson distribution and the mean ζ of the ZTP distribution are related by $\zeta = \frac{\mu e^{\mu}}{e^{\mu}-1} \equiv h(\mu)$. The MLE of μ under the ZTP assumption is then given by the inverse mapping $\hat{\mu}_T = h^{-1}(\hat{\zeta})$. Of course, all terms used in this section can be equipped with the index i to account for the case of covariates as laid out in Section 1.

3 Selection of the hybrid parameter

For the choice of γ, we have initially carried out a detailed theoretical study. To give some idea of this, we provide here the result that, in the covariate–free case, and only assuming a ZTP distribution for the non–zero part, the $MSE(T|n_0)$ is minimized at

$$\gamma_n^*(n_0,\mu) = \frac{\frac{n}{n-n_0} - h'(\mu)}{\frac{1}{n}\frac{\mu(n-n_0e^{\mu})^2 h'(\mu)^2}{e^{\mu}(e^{\mu}-1-\mu)} + h'(\mu)^2\left(1-\frac{n_0}{n}\right) + 2h'(\mu) + \frac{n}{n-n_0}} \qquad (2)$$

Figure 1 (right) shows the curve γ^* for fixed $n = 100$ and $\mu = 1$. It is, firstly, interesting to note that in a small range close to the expected value (≈ 37) under the Poisson model, the optimal γ is in fact > 1. However, for the majority of values of n_0 the curve is between 0 and 1, and falls very quickly below 1 when deviating from the expected value. While this kind of result could motivate an iterative procedure, in which T and γ are updated in turns via (1) and (2), we found this approach practically less useful since the increased variance incurred by the iterative estimation of γ contravenes the purpose of the hybrid estimator. We therefore considered two considerably simpler schemes:

(i) a single fixed rule–of–thumb value; where we have chosen $\gamma = 2/3$.

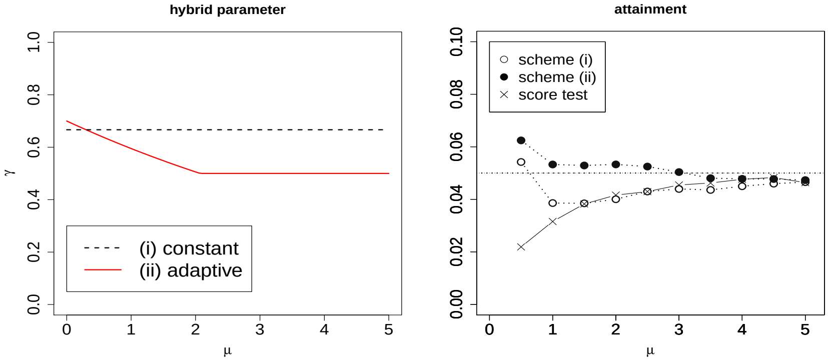
(ii) a parametric expression $\gamma = f(\hat{\mu}_P) = \begin{cases} 0.7\left(0.85^{\hat{\mu}_P}\right) & \hat{\mu}_P < \frac{\log(5/7)}{\log(17/20)} \\ \frac{1}{2} & otherwise \end{cases}$

The rationale of (ii) is to improve the attainment rate of the test by dereasing the weighting of the Poisson mean in the mixture for larger values of this estimator. The threshold $\frac{\log(5/7)}{\log(17/20)} \approx 2.07$ is chosen so that f is continuous. Figure 2 (left) compares settings (i) and (ii) graphically. Consider in this context the four crosses, from left to right in Figure 1 (right), which correspond to the optimal γ under zero–inflation parameter 0, 0.1, 0.2 and 0.5, respectively. We see that in the middle two cases (moderate zero–inflation) one has $\gamma^* \in [0.4, 0.8]$, so that we consider our suggested choices to be in harmony with our theoretical considerations.

4 Simulation

For the two–sided zero–modification test, Figure 2 (right) demonstrates, for a covariate–free simulation from Poisson data of varying μ, that (i) and

FIGURE 2. Left: choices (i) and (ii) for the selection of γ; right: attainment rate under mixture estimator (Two sided test of zero-modification)



(ii) both work well in terms of the nominal level attainment, with slight advantages for (ii). Focusing now on (ii), Figure 3 gives an impression of power as compared to the score test, as a function of sample size n. One sees that the powers are strong and very competitive to the score test, especially for smaller sample sizes. Note that here, and throughout this paper, the p-values reported for the proposed test are the mid p-values $\frac{1}{2}P_0[T \geq t+1] + \frac{1}{2}P_0[T \geq t]$ of Franck (1986).

FIGURE 3. Power under mixture estimator (Covariate–free Model, Two sided test of zero–modification)

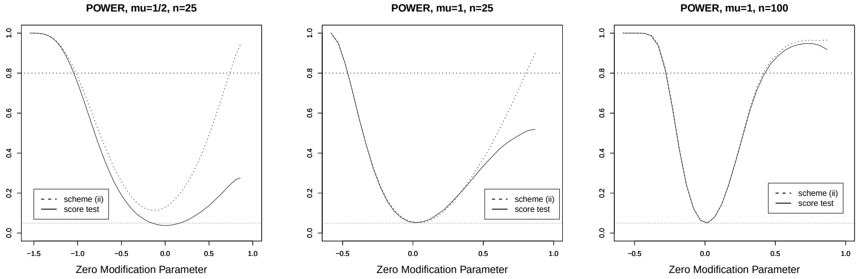
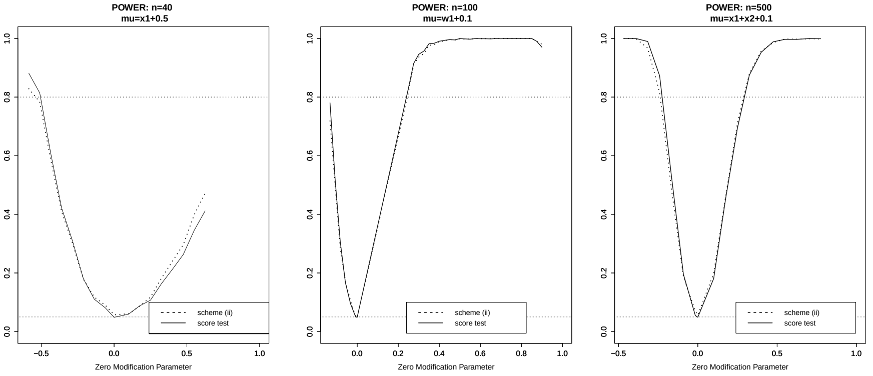


Figure 4 shows that the power and nominal attainment level of the proposed test also compares strongly to that of the score test in the presence of covariates. Here x_1 and x_2 are uniformly distributed on the interval $(0, 0.5)$, and w_1 is uniformly distributed on the interval $(1, 2)$. The adaptive mixing parameter is used, but the results remain similar for the constant estimator.

FIGURE 4. Power under mixture estimator (Covariate Model, Two sided test of zero–modification)



5 Examples

5.1 Biodosimetry Data

We consider four biodosimetry datasets consisting of chromosome aberration counts occuring after whole body exposure to ionising radiation. These datasets have previously been studied by Oliveira et al. (2016), detailed descriptions of the datasets are available in this paper.

Table 1 summarises the results obtained when the proposed test and a score test are used to test for zero-inflation relative to a quadratic Poisson model with log-link. We see that both tests fail to reject the Poisson model for the A3 data, but do not do so for the other datasets considered. For all the instances where the Poisson model was rejected we see that the observed number of zeros is greater than the upper limit of the 95% confidence interval, indicating that the data is zero-inflated.

TABLE 1. Analyses of Biodosimetry Data

	Proposed Test			Score Test	
Data	Obs. Zeros	95%CI	p-value	Statistic	p-value
A1	$14,430$	$(14204, 14329)$	$< 10^{-9}$	16.85	4.03×10^{-5}
A3	$2,747$	$(2719, 2823)$	0.368	1.01	0.317
B1	$7,280$	$(6707, 6829)$	$< 10^{-9}$	87.16	$< 10^{-9}$
C1	$6,786$	$(5031, 5164)$	$< 10^{-9}$	$1,996.10$	$< 10^{-9}$

5.2 Unwanted Pursuit Behaviour Data

Loeys et al. (2012) analysed data which concerns “separation trajectories”. Participants in a survey were assigned a score that theoretically ranges from

0 to 112, the maximum observed score was 34. This score is a measure of the participants experience of behaviour by their partner that contributed towards the breakup of a relationship. Two covariates were included in the model: a binary variable "education level" (0 = lower than bachelors degree, 1 = at least bachelors degree), and a continuous measurement for the level of anxious attachment in the former partner relationship. There are $n = 387$ data of which 246 are zeros. The proposed test shows that a 95% confidence interval for the number of observed zeros under the Poisson model is $(45, 72)$, and hence we may reject the Poisson model. Analysis of the data by a score test returns a statistic of 591.8, also indicating rejection of the Poisson model. Both tests return p-values $< 10^{-9}$.

6 Conclusion

The proposed test for zero-modification has power and attainment rates that compare very strongly to the score test. In addition to this it distinguishes between zero-inflation and zero-deflation and is a highly intuitive test that, unlike existing tests, is readily explainable to non-statistical specialists. The technique may be extended to compare any two count regression models, and may be used as the basis of a diagnostic plot for assessing model fit. See Einbeck and Wilson (2016).

References

Dietz, E. and Böhning, D. (2000). On estimation of the Poisson parameter in zero–modified Poisson models. *Computational Statistics & Data Analysis* **34**, pages 441 – 459.

Einbeck, J. and Wilson, P. (2016). A Diagnostic Plot for Assessing Model Fit. Proc's of the 31st IWSM, Rennes, France, *to appear.*

Franck, W. (1986). P-values for Discrete Test Statistics. *Biometrical Journal* **4**, pages 403 – 406.

Loeys,T., Moerkerke, B., De Smet, O., and Buysse, A. (2012) Expert Tutorial: The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology* **65**, pages 163 – 180.

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

Wilson, P. and Einbeck, J. (2015). A simple and intuitive test for number–inflation or number–deflation. In: Wagner, H. and Friedl, H. (Eds). Proc's of the 30th IWSM, Linz, Austria, Vol 2, pages 299 – 302.

Sparse relative risk survival modelling

Ernst C. Wit[1], Hassan Pazira[1], Fentaw Abegaz [2], Javier Gonzalez [3], Luigi Augugliaro[4]

[1] University of Groningen, Netherlands
[2] University of Liege, Belgium
[3] University of Sheffield, United Kingdom
[4] University of Palermo, Italy

E-mail for correspondence: e.c.wit@rug.nl

Abstract: Cancer survival is thought to be closely linked to the genomic constitution of the tumour. Discovering such signatures will be useful in the diagnosis of the patient and may be used for treatment decisions and perhaps even the development of new treatments. However, genomic data are typically noisy and high-dimensional, often outstripping the number of patients included in the study. Regularized survival models have been proposed to deal with such scenarios. These methods typically induce sparsity by means of a coincidental match of the geometry of the convex likelihood and (near) non-convex regularizer. The disadvantages of such methods are that (i) they are typically non-invariant to scale changes of the covariates, (ii) they struggle with highly correlated covariates and (iii) the have a practical problem of determining the amount of regularization. In this manuscript we propose a principled method for sparse inference in relative risk regression models based only on the likelihood. The method is computationally fast and is implemented in the R-package `dglars`.

Keywords: Relative risk regression models; survival analysis; gene expression data; high dimensional data; sparsity; differential geometry; dgLARS.

1 Introduction

Sparse inference in the past two decades has been dominated by methods that penalize typically convex likelihoods by functions of the parameters that happen to induce solutions with many zeros. The lasso (Tibshirani, 1996) and other penalization approaches are all examples of methods that depending on some tuning parameter conveniently shrink estimates to exact zeroes. Also in survival analysis these methods have been introduced.

Tibshirani (1997) applied the lasso penalty to Cox proportional hazards model. Although the lasso penalty induces sparsity, it is well known to suffer from possible inconsistent selection of variables.

In this paper, we will approach sparsity directly from a likelihood point of view. The angle between the covariates and the tangent residual vector within the likelihood manifold provides a direct and scale-invariant way to assess the importance of the individual covariates. The idea is similar to the least angle regression approach proposed by Efron *et al.* (2004) and Augugliaro *et al.* (2013). Moreover, the method extends directly beyond the Cox proportional hazard model. In fact, we derive all the results for general relative risk survival models.

2 Sparse relative risk regression

The dgLARS method (Augugliaro *et al.*, 2013) introduced sparse inference for generalized linear models.

2.1 Relative risk regression

To extend the method to survival models, Thomas (1977) observed that the partial surival likelihood

$$\mathcal{L}_p(\beta) = \prod_{i \in \mathcal{D}} \frac{\psi(\mathbf{x}_i(t_i);\beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i);\beta)}. \tag{1}$$

can arise from a multinomial sample scheme. Consider an index $i \in \mathcal{D}$ and let $\mathbf{Y}_i = (Y_{ih})_{h \in \mathcal{R}(t_i)}$ be a multinomial random variable with sample size equal to 1 and cell probabilities $\pi_i = (\pi_{ih})_{h \in \mathcal{R}(t_i)} \in \Pi_i$, i.e. $p(\mathbf{y};\pi_i) = \prod_{h \in \mathcal{R}(t_i)} \pi_{ih}^{y_{ih}}$. Assuming independence and the following model for the conditional expected value of the random variable Y_{ih}, i.e. $E_\beta(Y_{ih}) = \pi_{ih}(\beta) = \frac{\psi(\mathbf{x}_h(t_i);\beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i);\beta)}$, then our model space is the set

$$\mathcal{M} = \left\{ \prod_{i \in \mathcal{D}} \prod_{h \in \mathcal{R}(t_i)} \left(\frac{\psi(\mathbf{x}_h(t_i);\beta)}{\sum_{j \in \mathcal{R}(t_i)} \psi(\mathbf{x}_j(t_i);\beta)} \right)^{y_{ih}} : \beta \in \mathcal{B} \right\}. \tag{2}$$

The partial likelihood (1) is formally equivalent to the likelihood function associated with the model space $\mathcal{M}$ if we assume that for each $i \in \mathcal{D}$, the observed y_{ih} is equal to one if h is equal to i and zero otherwise. Let $\ell(\beta) = \sum_{i \in \mathcal{D}} \sum_{h \in \mathcal{R}(t_i)} Y_{ih} \log \pi_{ih}(\beta)$ be the log-likelihood function associated to the model space $\mathcal{M}$ and let $\partial_m \ell(\beta) = \partial \ell(\beta)/\partial \beta_m$.

2.2 Differential geometric least angle regression

The likelihood of a relative risk survival models is clearly non-linear in the parameters. The geometry of the likelihood manifold can be defined locally, by considering the structure of tangent spaces. This differential geometric representation can be used to extend the least angle regression approach. The dgLARS estimator is based on a differential geometric characterization of the Rao score test statistic, which is obtained considering the inner product between the bases of the tangent space $T_\beta\mathcal{M}$ and the tangent residual vector $r_\beta = \sum_{i\in\mathcal{D}}\sum_{h\in\mathcal{R}(t_i)} r_{ih}(\beta)\partial_{ih}\ell(\beta)$, where $r_{ih}(\beta) = y_{ih} - \pi_{ih}(\beta)$. The dgLARS method is a sequential method developed to estimate a sparse solution curve embedded in the in the parameter space $\mathcal{B}$. To explore the sparse structure of a relative risk regression model, we can use the following differential geometric characterization characterization of the mth element of the score vector, i.e., $\partial_m\ell(\beta) = \langle\partial_m\ell(\beta); r_\beta\rangle_\beta = \cos(\rho_m(\boldsymbol{\beta}))\cdot I^{1/2}_{mm}(\beta)\cdot\|r_\beta\|_\beta$, where $I_{mm}(\beta)$ is the Fisher information for β_m, and $\rho_m(\boldsymbol{\beta})$ is a generalization of the Euclidean notion of angle between the mth column of the design matrix and the residual vector $\mathbf{r}(\beta)$. One can see that the signed Rao's score test statistic can be geometrically characterized as follows:

$$r^u_m(\beta) = I^{-1/2}_{mm}(\beta)\partial_m\ell(\beta) = \cos(\rho_m(\boldsymbol{\beta}))\cdot\|r_\beta\|_\beta,$$

then we shall say that two given predictors, say m and n, satisfy the generalized equiangularity condition at the point β when $|r^u_m(\beta)| = |r^u_n(\beta)|$. Inside the dgLARS theory, the generalized equiangularity condition is used to identify the predictors that are included in the active set. Formally, for a given value of the Rao score test statistic $\gamma\in\mathbb{R}^+$ the corresponding active set is denoted by $\hat{\mathcal{A}}(\gamma)$ and the dgLARS estimator, denoted by $\hat\beta(\gamma)$, is such that the following conditions are satisfied:

$$\forall\, m\in\hat{\mathcal{A}}(\gamma) \quad\Rightarrow\quad r^u_m(\hat\beta(\gamma)) = s_m\gamma, \tag{3}$$

$$\forall\, n\notin\hat{\mathcal{A}}(\gamma) \quad\Rightarrow\quad |r^u_n(\hat\beta(\gamma))| < \gamma. \tag{4}$$

where $s_m = \text{sign}(\hat\beta_m(\gamma))$.

2.3 Estimation of the DgLARS solution path

Using the differential geometrical structure of a relative risk regression model and the previous conditions, it is possible to use the dgLARS method to explore the sparse structure of a relative risk regression model. Formally, the dgLARS method computes a finite sequence of transition points, say $0\le\gamma^{(K)}\le\ldots\le\gamma^{(2)}\le\gamma^{(1)}$, such that for each $\gamma^{(k)}$ one of the following two conditions can occur:

(i) $\exists\, n \notin \hat{\mathcal{A}}\gamma^{(k-1)}$ such that

$$\left| r_n^u(\hat{\beta}(\gamma^{(k)})) \right| \quad = \quad \gamma, \tag{5}$$

and therefore $\hat{\mathcal{A}}\gamma^{(k)} = \hat{\mathcal{A}}\gamma^{(k-1)} \cup \{n\}$;

(ii) $\exists\, m \in \hat{\mathcal{A}}(\gamma^{(k-1)})$ such that

$$\mathrm{sign}(r_m^u(\hat{\beta}(\gamma^{(k)}))) \quad \neq \quad \mathrm{sign}(\hat{\beta}_m(\gamma^{(k)})), \tag{6}$$

and therefore $\hat{\mathcal{A}}\gamma^{(k)} = \hat{\mathcal{A}}\gamma^{(k-1)} \setminus \{m\}$,

which means that a new predictor is included in the active set when the generalized equiangularity condition is satisfied, namely condition (5), or an active predictor is removed from the active set if the sign of the corresponding signed Rao's score test statistic is not in agreement with the sign of the estimated coefficient, i.e. condition (6). In order to simplify our notation, in the following of this section we shall assume that $\hat{\mathcal{A}}\gamma = \{1, 2, \ldots, k\}$. Observing that for each $\gamma \in (\gamma^{(k+1)}; \gamma^{(k)}]$ the signs of the estimated coefficients do not change, condition (3) tells us that, for a fixed value of the tuning parameter γ, the dgLARS estimator can be defined as the Z-estimator implicitly defined by the following system of estimating equations:

$$\left\{ \begin{array}{lcl} r_1^u(\hat{\beta}(\gamma)) - s_1\gamma & = & 0 \\ r_2^u(\hat{\beta}(\gamma)) - s_2\gamma & = & 0 \\ \vdots & & \vdots \\ r_k^u(\hat{\beta}(\gamma)) - s_k\gamma & = & 0. \end{array} \right. \tag{7}$$

3 Sparse Cox proportional hazards model

Let Z, C and $\mathbf{x}(t)$ denote the survival time, the censoring time and their associated p-dimension vector of covariates which can depend on time t, respectively. Further denote by $T = \min\{Z, C\}$ the observed time and $Y = \mathcal{I}\{Z \leq C\}$ the censoring indicator. For simplicity, we assume that Z and C are conditionally independent and the censoring mechanism is non-informative.

The proportional hazards model is very popular in survival data analysis partially due to its simplicity and its convenience in dealing with censoring. The proportional hazards model assumes that the hazard function is

$$\lambda(t; \mathbf{x}) = \lambda_0(t) \exp(\beta^T \mathbf{x}(t)), \tag{8}$$

where $\lambda_0(t)$ is the baseline hazard function is unspecified and needs to be estimated nonparametrically and β is a p-dimensional vector of unknown fixed parameters of interest. The proportional hazards model is an example

of relative risk regression models where the hazard function in ((8)) is given by $\psi(\mathbf{x}(t);\beta) = \exp(\beta^T \mathbf{x}(t))$.
Inference about β can be made using the partial likelihood originally introduced in Cox (1972), which is given via

$$\mathcal{L}_p(\beta) \quad = \quad \prod_{i\in\mathcal{D}} \frac{\exp(\beta^T \mathbf{x}_i(t_i))}{\sum_{j\in\mathcal{R}(t_i)} \exp(\beta^T \mathbf{x}_j(t_i))}. \tag{9}$$

where $\mathcal{D}$ is the set of indices corresponding to failed subjects and $\mathcal{R}(t_i)$ denotes the risk set, that is the set of indices corresponding to the subjects who have not failed and still under observation just prior to time t. The Cox partial likelihood is a special case of the likelihood function defined in general for relative risk regression models. Because the Cox proportional hazards model can be expressed as a relative risk regression model, variable selection can be performed using dgLARS as formulated and discussed in the previous sections.

4 Finding genetic signatures in cancer survival

We apply dgLARS relative risk regression to the identification of genes involved in the regulation of colon (Loboda *et al.*, 2011) and skin (Jonsson *et al.*, 2010) cancers. The set-up of the both studies was similar. After cancer was detected the patients started to follow some specific treatment. In all cases, the expression of some genes was measured in the affected tissue together with the survival times of the patients, which is assumed to be censored if the patients were alive when they left the study.

TABLE 1. Description of the two high-dimensional cancer experiments studied in this section. Datasets are available at http://www.ncbi.nlm.nih.gov/.

Cancer	n	uncen.	p	p sel.	GW test	Reference
Colon	125	70	23698	62	0.0224	Loboda *et al.* 2011
Skin	54	47	30807	21	0.025	Jonsson *et al.*, 2010

Table 1 shows that for both scenarios p is much larger than n. In genomic studies it is a common hypothesis to assume that just a few number of genes affect the dependent variable of interest. To identify such genes in our survival data analysis context, we estimate a relative hazard risk model using the dgLARS algorithm. To this end, we randomly select a training sample that contains the 60% of the patients and we save the remaining data to test the models. We calculate the paths coefficients in the four scenarios and we select the optimal number of components by means of the GIC criterion. The number of selected genes in each case is detailed in Table 1 ranging from 21 genes in the skin cancer data set to 62 in the colon dataset.

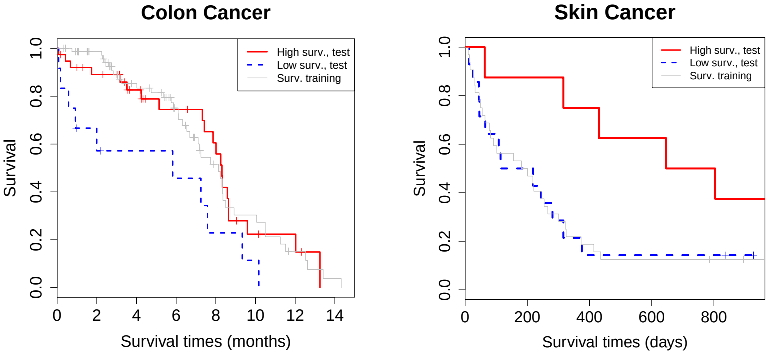


FIGURE 1. The Kaplan-Meier survival curves estimates for training data are shown together with the curves associated to the two groups obtained in the test sample by means of the predicted excess risk.

References

Augugliaro, L., Mineo, A., Wit, E.C. (2013). Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models. *JRSS-B*, **75**(3), 471 – 498.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**(2), 187 - 220.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**(2), 407 – 499.

Jonsson, G, *et al.* (2010). Gene expression profiling-based identification of molecular subtypes in stage iv melanomas with different clinical outcome. *Clin Cancer Res*, **16**(13), 3356 – 3367.

Loboda, A, *et al* . (2011). Emt is the dominant program in human colon cancer. *BMC Med Genomics*, **20**, 4 – 9.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *JRSS-B*, **58**(1) 267 – 288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, **16**, 385 – 395.

Thomas, D.C. (1977). Addendum to the paper by Liddell, McDonald, Thomas and Cunliffe. *JRSS-A*, **140**(4), 483 – 485.

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