IWSM 2020

Proceedings of the 35th International Workshop on Statistical Modelling

July 20-24, 2020 - Bilbao, Basque Country, Spain

Itziar Irigoien, Dae-Jin Lee, Joaquín Martínez-Minaya,
María Xosé Rodríguez-Álvarez (Editors)

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

CIP. Biblioteca Universitaria

International Workshop on Statistical Modelling (35º. 2020. Bilbao)

Proceedings of the 35th International Workshop on Statistical Modelling : July 20-24, 2020 Bilbao, Basque Country, Spain / Itziar Irigoien … [et al.] (Editors). – Datos. - Bilbao : Universidad del País Vasco / Euskal Herriko Unibertsitatea, Argitalpen Zerbitzua = Servicio Editorial, [2020]. – 1 recurso en línea : PDF (ix, 457 p. : il.)
Modo de acceso: World Wide Web
ISBN: 978-84-1319-267-3.

1. Estadística matemática – Congresos. 2. Modelos econométricos – Congresos. I. Irigoien, Itziar, coed.

(0.034)519.2(063)
(0.034)519.86/.87(063)

Due to COVID-19 pandemic, the IWSM2020 was cancelled. However, all extended abstracts submitted were reviewed and scored by the members of the Scientic Programme Committee. This document contains those papers selected for oral and poster presentation and whose authors gave their consent for their work to be included in the proceedings.

Editors:

Itziar Irigoien
University of the Basque Country UPV/EHU, Donostia, Spain
itziar.irigoien@ehu.eus

Dae-Jin Lee
BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
dlee@bcamath.org

Joaquín Martínez-Minaya
BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
jomartinez@bcamath.org

María Xosé Rodríguez- Álvarez
BCAM - Basque Center for Applied Mathematics
& IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
mxrodriguez@bcamath.org

ISBN: 978-84-1319-267-3

Scientific Programme Committee

- Carmen Armero
  Universitat de València (Spain)
- Marcelo Bourguignon Pereira
  Universidade Federal do Rio Grande do Norte (Natal, Brasil)
- Jochen Einbeck
  University of Durham (UK)
- Christel Faes
  Hasselt University (Belgium)
- Gillian Heller
  Macquarie University (Australia)
- Helmut Küchenhoff
  Ludwig-Maximilians-Universität München (Germany)
- Joseph Lang
  University of Iowa (USA)
- Dae-Jin Lee
  BCAM–Basque Center for Applied Mathematics (Spain)
- Luís Filipe Meira Machado
  Universidade do Minho (Portugal)
- Claire Miller
  University of Glasgow (UK)
- Vicente Núñez-Antón
  University of the Basque Country UPV/EHU (Spain)
- María Xosé Rodríguez Álvarez (Chair)
  BCAM–Basque Center for Applied Mathematics & IKERBASQUE, Basque Foundation for Science (Spain)
- Laura M. Sangalli
  Politécnico di Milano (Italy)
- Sabine Schnabel
  Wageningen University and Research (The Netherlands)
- Cécile Proust-Lima
  Université de Bordeaux (France)

Local Organising Committee

- Inmaculada Arostegui Madariaga

  University of the Basque Country UPV/EHU & BCAM–Basque Center for Applied Mathematics.

- Irantzu Barrio Beraza

  University of the Basque Country UPV/EHU

- Itziar Irigoien

  University of the Basque Country UPV/EHU

- Dae-Jin Lee (Chair)

  BCAM–Basque Center for Applied Mathematics

- Joaquín Martínez-Minaya

  BCAM–Basque Center for Applied Mathematics

- Jesús Orbe

  University of the Basque Country UPV/EHU

- María Xosé Rodríguez-Álvarez

  BCAM–Basque Center for Applied Mathematics & IKERBASQUE, Basque Foundation for Science

- Jorge Virto

  University of the Basque Country UPV/EHU

Preface

The International Workshop on Statistical Modelling (IWSM) is a reference workshop in promoting statistical modelling, applications of Statistics for researchers, academics and industrialist in a broad sense. Unfortunately, the global COVID-19 pandemic has not allowed holding the 35th edition of the IWSM in Bilbao in July 2020. Despite the situation and following the spirit of the Workshop and the Statistical Modelling Society, we are delighted to bring you the proceedings book of extended abstracts.

First, we would like to thank all the authors for their scientific contributions and congratulate them for the high quality of the extended abstracts. To keep the spirit of the IWSM, we have compiled them into two parts: Part I with those extended abstracts selected for oral presentations and Part II for poster presentations. A total number of 135 extended abstracts were submitted, with 62 extended abstracts chosen for oral presentation and 73 for poster presentation. From those, a total of 97 authors have given their consent for their extended abstracts to be included in the proceedings.

The proceedings could not have been possible without the great work of the scientific committee who have evaluated and scored all extended abstracts. We are aware that the work of selecting the extended abstracts for oral or posters presentations is always arduous, and we would like to thank each of the members of the Scientific Committee for their incredible work!

A deep thanks to the Executive Committee of the Statistical Modelling Society for their support and the organisers of the IWSM 2021 in Natal in Brazil for deferring their edition and allowing us to organise the IWSM next year in Bilbao.

Also, we would like to thank María Durbán, Montserrat Fuentes, Yudi Pawitan, Virginie Rondeau, Stijn Vansteelandt and Virgilio Gómez-Rubio for having accepted (twice!) our invitation to participate in the workshop. We are looking forward to having you next year.

Last but not least, we thank the editorial work by Itziar Irigoien and Joaquín Martínez-Minaya as well as the whole local organising committee for their full support in the decision to cancel the workshop and their willingness to accompany us in 2021.

We hope to see you in Bilbao in 2021. Save the date: July 18–23, 2021!

Dae-Jin Lee and María Xosé (Coté) Rodríguez Álvarez
Bilbao, July 2020

Contents

Part I

Hidden Markov models for multi-scale time series: an application to stock market data

Timo Adam[1,2] and Lennart Oelschlger[2]

[1] University of St Andrews, St Andrews, UK
[2] Bielefeld University, Bielefeld, Germany

E-mail for correspondence: ta59@st-andrews.ac.uk

Abstract: Over the last decades, hidden Markov models have emerged as a versatile class of statistical models for time series where the observed variables are driven by latent states. While conventional hidden Markov models are restricted to modeling single-scale data, economic variables are often observed at different temporal resolutions: an economy's gross domestic product, for instance, is typically observed on a yearly, quarterly, or monthly basis, whereas stock prices are available daily or at even finer temporal resolutions. In this paper, we propose hierarchical hidden Markov models to incorporate such multi-scale data into a joint model, where we illustrate the suggested approach using 16 years of monthly trade volumes and daily log-returns of the Goldman Sachs stock.

Keywords: Hidden Markov models; Multi-scale data; Stock markets; Time series modeling; Temporal resolution.

1 Introduction

Hidden Markov models (HMMs) constitute a versatile class of statistical models for time series where the observed variables are driven by latent states (Zucchini *et al.*, 2016). While the observations can be multivariate, basic HMMs have the limitation that all variables need to be observed at the same temporal resolution. Specifically in economic applications, however, corresponding variables are often observed at different time scales, ranging from yearly data such as economic indices to high-frequency stock market data. Incorporating multiple such variables, with differing sampling rates, into a joint model may help to draw a more comprehensive picture of stock market dynamics, in particular by explicitly distinguishing short-term and long-term variation in volatility. In this paper, we propose hierarchical

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

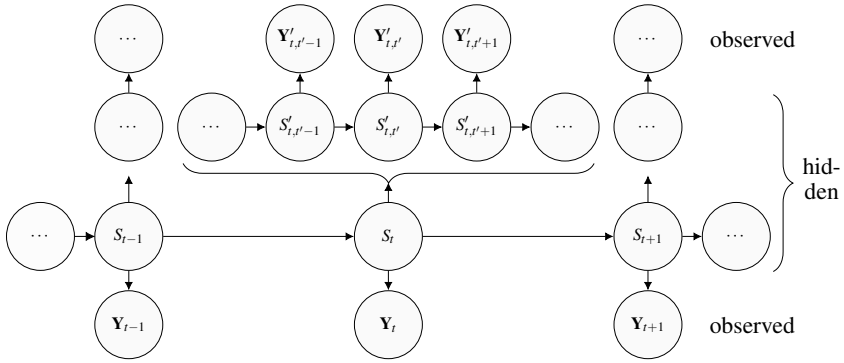


FIGURE 1. Dependence structure of an hierarchical HMM.

HMMs, which originate from the field of machine learning (Fine *et al.*, 1998) and have later been applied in ecology (Leos-Barajas *et al.*, 2017; Adam *et al.*, 2017; Adam *et al.*, 2019), to incorporate such multi-scale data into a joint model. The suggested approach is illustrated by jointly modeling 16 years of monthly trade volumes and daily log-returns of the Goldman Sachs stock.

2 Model formulation and likelihood evaluation

A basic HMM comprises two stochastic processes: a hidden state process $\{S_t\}_{t=1,\ldots,T}$ and an observed state-dependent process $\{Y_t\}_{t=1,\ldots,T}$. The state process is typically modeled as a discrete-time, N-state Markov chain with initial distribution $\boldsymbol{\delta} = (\delta_i)$, $\delta_i = \Pr(S_1 = i)$, and transition probability matrix (t.p.m.) $\boldsymbol{\Gamma} = (\gamma_{i,j})$, $\gamma_{i,j} = \Pr(S_{t+1} = j | S_t = i)$. The state at time t, $S_t = i$, selects one of N possible distributions, which are denoted by $f(y_t|S_t = i)$, that generates the outcome of the state-dependent process (cf. Zucchini *et al.*, 2016).
By exploiting this relatively simple dependence structure, the likelihood can be written as a matrix product,

$$\mathcal{L}^{\text{HMM}}(\boldsymbol{\theta}|y_1, \ldots, y_T) = \boldsymbol{\delta}\mathbf{P}(y_1) \prod_{t=2}^{T} \boldsymbol{\Gamma}\mathbf{P}(y_t)\mathbf{1}, \tag{1}$$

where $\mathbf{P}(y_t) = \text{diag}\big(f(y_t|S_t = 1), \ldots, f(y_t|S_t = N)\big)$ and $\mathbf{1}$ denotes a column vector of ones (cf. Zucchini *et al.*, 2016).
Hierarchical HMMs extend the model structure outlined above in that they distinguish between processes operating at different time scales (cf. Figure 1 for an illustration of the model structure): the coarse-scale state at time t, $S_t = i$, selects among N possible distributions for the coarse-scale observations, which are denoted by y_t (e.g. the trade volume observed for month

t), and N possible HMMs (each of which has its own t.p.m. $\mathbf{\Gamma}_i'$) for the fine-scale observations, which are denoted by $\mathbf{y}_t'$ (e.g. all daily log-returns observed during month t). The likelihood then follows as

$$\mathcal{L}^{\text{HHMM}}(\boldsymbol{\theta}|y_1,\ldots,y_T,\mathbf{y}_1',\ldots,\mathbf{y}_T') = \boldsymbol{\delta}\mathbf{P}(y_1,\mathbf{y}_1')\prod_{t=2}^{T}\mathbf{\Gamma}\mathbf{P}(y_t,\mathbf{y}_t')\mathbf{1}, \quad (2)$$

where $\mathbf{P}(y_t,\mathbf{y}_t') = \text{diag}\big(f(y_t|S_t=1)\mathcal{L}^{\text{HMM}}(\boldsymbol{\theta}|\mathbf{y}_t',S_t=1),\ldots,f(y_t|S_t=N)\mathcal{L}^{\text{HMM}}(\boldsymbol{\theta}|\mathbf{y}_t',S_t=N)\big)$. Estimation of the model parameters is typically carried out by numerical likelihood maximization (cf. Adam *et al.*, 2019).

3 Application to stock market data

To investigate stock market dynamics at different time scales, we jointly model 16 years of monthly trade volumes and daily log-returns of the Goldman Sachs stock. The data cover 4,026 working days (192 months) between January 1, 2004, and December 31, 2019. For the trade volumes, we assumed gamma distributions, while for the log-returns, scaled t-distributions (as preferred over Normal distributions by Akaike's information criterion) were considered.

The estimated state-dependent distributions of monthly trade volumes, displayed in the top left panel of Figure 2, reveal three different market regimes: while coarse-scale states 1 and 2 capture low and moderate trade volumes (inactive and moderately active market), respectively, state 3 relates to high trade volumes (active market).

The t.p.m. associated with the coarse-scale state process was estimated as

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.984 & 0.016 & 0.000 \\ 0.043 & 0.900 & 0.057 \\ 0.000 & 0.282 & 0.718 \end{pmatrix},$$

which implies the stationary distribution $(0.687, 0.261, 0.053)$, indicating that about 69 %, 26 %, and 5 % of the monthly trade volumes were generated in coarse-scale states 1, 2, and 3, respectively. Notably, in 2007, when a sudden increase in interest rates for inter-bank credits marked the beginning of the global financial crisis, the decoded time series displayed in the top right panel of Figure 2 reveals a switch from coarse-scale state 1 (inactive market) to 2 (moderately active market). In September 2008 (when the Lehman Brothers collapse marked the peak of the global financial crisis), we observe a switch from coarse-scale state 2 to 3 (active market).

The estimated state-dependent distributions of daily log-returns are displayed in the middle panel of Figure 2: depending on the coarse-scale state that is active in month t, the log-returns' volatility is determined by the fine-scale HMM associated with the two distributions displayed in either the left, the middle, or the right panel, respectively.

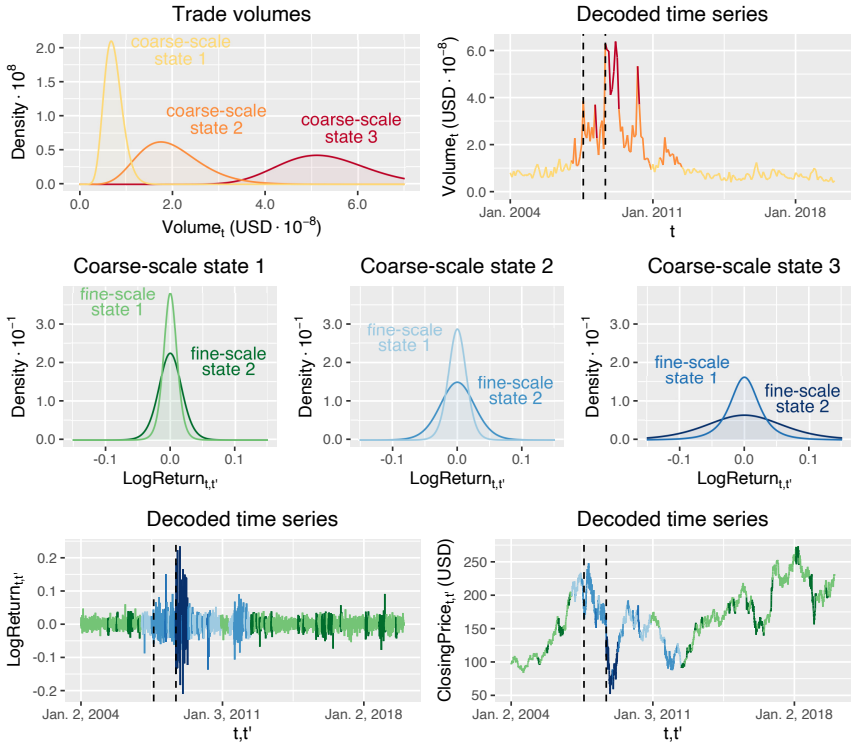


FIGURE 2. Estimated state-dependent distributions and decoded time series of monthly trade volumes, daily log-returns, and closing prices of the Goldman Sachs stock. Dashed lines in the top-right and the bottom panel indicate important events associated with the global financial crisis.

The t.p.m.s associated with the fine-scale state processes were estimated as

$$\hat{\mathbf{\Gamma}}_1' = \begin{pmatrix} 0.993 & 0.007 \\ 0.034 & 0.966 \end{pmatrix}, \; \hat{\mathbf{\Gamma}}_2' = \begin{pmatrix} 0.993 & 0.007 \\ 0.024 & 0.976 \end{pmatrix}, \; \hat{\mathbf{\Gamma}}_3' = \begin{pmatrix} 0.915 & 0.085 \\ 0.029 & 0.971 \end{pmatrix},$$

which imply the stationary distributions $(0.823, 0.177)$, $(0.779, 0.221)$, and $(0.255, 0.745)$. According to the fitted model, when coarse-scale state 1 (inactive market) is active (about 67 % of the time) then the marginal distribution of the log-returns under the fitted model has standard deviation 0.013. When coarse-scale state 3 (active market) is active (about 5 % of the time), then the log-returns' volatility is about five times higher: the corresponding marginal distribution has standard deviation 0.065.
Quantile-quantile-plots and sample autocorrelation functions (ACFs) of ordinary normal pseudo-residuals for monthly trade volumes and daily log-returns are displayed in Figure 3. While, in principle, more flexible state-

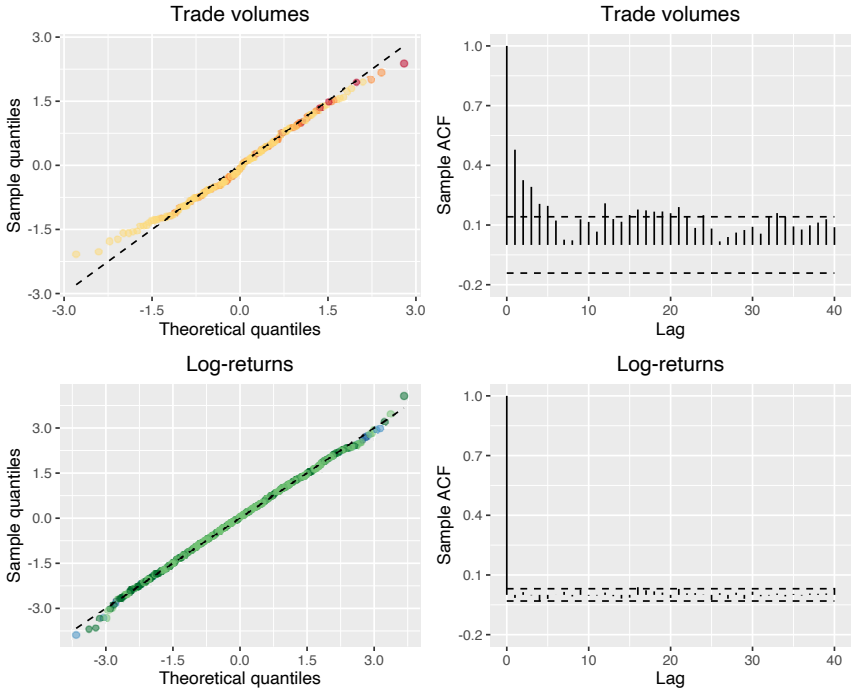


FIGURE 3. Quantile-quantile-plots (left panel) and sample ACFs (right panel) of normal ordinary pseudo-residuals for monthly trade volumes (top panel) and daily log-returns (bottom panel). Overall, the plots indicate some minor lack of fit with regard to the marginal distribution of the trade volumes and the serial correlation in the trade volumes' series.

dependent distributions (especially for the trade volumes) could be used to improve the model fit (cf. Langrock *et al.*, 2018), we consider the goodness of fit to be satisfactory and trade some minor lack of fit against a more complex model formulation, which facilitates the interpretation of the fitted model.

4 Conclusions

The results presented in this paper indicate that coarse-scale market dynamics strongly affect the stochastic properties of other processes operating at finer time scales. By explicitly modeling such multi-scale processes, hierarchical HMMs may help to draw a more comprehensive picture of stock market dynamics, to more accurately quantify risks conditional on the coarse-scale market regime, and ultimately to improve our understanding of the market agents' behavior.

References

Adam, T., Leos-Barajas, V., van Beest, F.M., and Langrock, R. (2017). Using hierarchical hidden Markov models for joint inference at multiple temporal scales. *Proceedings of 32nd International Workshop on Statistical Modelling*, **2**, 181 – 184.

Adam, T., Griffiths, C.A., Leos-Barajas, V., Meese, E.N., Lowe, C.G., Blackwell, P.G., Righton, D., and Langrock, R. (2019). Joint modeling of multi-scale animal movement data using hierarchical hidden Markov models. *Methods in Ecology and Evolution*, **10**(9), 1536 – 1550.

Fine, S., Singer, Y., and Tishby, N. (1998). The hierarchical hidden Markov model: analysis and applications. *Machine Learning*, **32** (1), 41 – 62.

Langrock, R., Adam, T., Leos-Barajas, V., Mews, S., Miller, D.L., and Papastamatiou, Y.P. (2018). Spline-based nonparametric inference in general state-switching models. *Statistica Neerlandica*, **72**(3), 179 – 200.

Leos-Barajas, V., Gangloff, E.J., Adam, T., Langrock, R., van Beest, F.M., Nabe-Nielsen, J., and Morales, J. (2017). Hierarchical generalized linear models. *Journal of Agricultural, Biological and Environmental Statistics*, **22**(3), 232 – 248.

Zucchini, W., MacDonald, I.L., and Langrock, R. (2016). *Hidden Markov models for time series: an introduction using R*. Boca Raton: Chapman and Hall/CRC.

Multivariate conditional dependence. The effect of institutional quality on competitiveness indicator relations

Jone Ascorbebeitia[1], Eva Ferreira[1], Susan Orbe[1]

[1] University of the Basque Country UPV/EHU, Spain

E-mail for correspondence: jone.ascorbebeitia@ehu.eus

Abstract: Nonparametric estimators for multivariate conditional copulas as well as for a multivariate conditional Kendall's tau are proposed in a random design context. We also propose a flexible Wald type statistic based on Kendall's tau estimator to test for the influence of a conditioning variable outcome in the joint distribution between two or more variables. Asymptotic properties of the estimators are derived together with a simulation study, and a data-driven smoothing parameter selection is also provided. A second simulation study presents different models to check the size and power of the test and runs comparisons with previous proposals when appropriate. For the empirical illustration, we study the relationship between some indicators from the European Regional competitiveness index (RCI). We find interesting results, such as weaker links between innovation and higher education in regions with lower institutional quality. Analyzing this kind of comovements is very useful for regulatory purposes to measure the impact of economic policies.

Keywords: Conditional copula; Nonparametric estimation; Multivariate dependence; Kendall's tau.

1 Introduction

Since the last financial crisis, economic growth has been an important concern all over the world. Particularly, the European Commission constructs the RCI index as an indicator of economic progress every three years since 2010. This index comprises more than 70 indicators that enable measuring the ability of the regions to offer an attractive and sustainable environment for firms and residents to live and work and they are grouped into 11 pillars to help to identify the strengths and weaknesses of each region. In this

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

context, we analyze the relationship of indicators related to the efficiency and the innovation of a region. Additionally, since institutional quality is a factor that can disturb the economic growth, we find interesting to study if the indicators dependence structure is affected by the quality level.
We estimate the conditional joint dependence with nonparametric conditional copulas, which is a flexible way of modeling the dependence structure and better that elliptic distributions when variables are not normal (Embrechts et al., 2002). We consider a random design context, suitable for many economic applications, and derive the asymptotic results for the multivariate conditional copula estimator.
As an overall measure of conditional dependence, we use a multivariate rank correlation measure beyond the linear correlation: The conditional Kendall's tau. We take as a reference the multivariate Kendalls tau proposed by Joe (1990) beyond the average pairwise tau (Kendall and Smith, 1940) to extend it to the conditional case and we derive the main asymptotic results. A bandwidth selection algorithm for Kendall's tau estimator and a simulation study to check for its robustness are also provided.
Moreover, we want to test for the structure of conditional dependence according to Kendall's tau. Tests for unconditional multivariate independence are well known in the literature, but we are interested in testing for general restrictions in the conditional rank correlation. Related works introduce tests of the so-called simplifying assumption, which assumes that the conditional copula coincides with the partial copula. Moreover, Gijbels et al. (2017) propose tests based on Kendall' tau and compare their performance with some tests based on conditional copulas. Nevertheless, our objective is to test for a broader type of restrictions that include conditional independence or constant dependence but also constant conditional dependence, linear restrictions between different Kendall's taus, and equality of conditional Kendall's tau between different samples. Results from a simulation study show that our proposal performs well in many different scenarios, such as mixtures of copulas as joint distributions, and entails a low computational cost.
The results in the application show that the dependence between pillars such as higher education and innovation increases as the institutional quality improves. Actually, this relationship is strong and significantly affected by low quality levels. Thus, the results support that the institutions quality perception has a clear significant effect on regional competitiveness factors. It seems that in regions with low institutional quality, the investment in efficiency does not have the expected results on improving the innovative capability of the region.

2 Multivariate conditional copulas

Let $\boldsymbol{Y} = \{Y_j\}_{j=1}^{p}$ be a set of p variables. To estimate the dependence structure we propose a nonparametric estimator for the multivariate copula

conditional on $Z = z$ defined as

$$\hat{C}_z(\boldsymbol{u}) = \sum_{i=1}^{n} w_i(z,h) I\{Y_{1i} \leq \hat{F}_{1z}^{-1}(u_1), ..., Y_{pi} \leq \hat{F}_{pz}^{-1}(u_p)\}, \qquad (1)$$

where Y_{ji} sets for a generic sample value i of the variable Y_j, $\{w_i(z,h)\}$ is a sequence of weights, and $h > 0$ is the bandwidth that decreases to zero as the sample size increases. $I\{\cdot\}$ is the indicator function and $\hat{F}_{jz}(y) = \sum_{i=1}^{n} w_i(z,h) I\{Y_{ji} \leq y\}$ denotes the nonparametric conditional estimator of the j-marginal distribution. It is noteworthy that this nonparametric estimator does not have a smoothing role as is usual in regressions. In fact, when the bandwidth h gets large enough, $\hat{C}_z$ approaches the empirical distribution function, which is not smooth.

To measure the degree of dependence, we estimate Kendall's tau coefficient, which measures the monotonicity instead of linearity. We extend the version proposed by Joe (1990) for multivariate Kendall's tau to conditional copulas as $\tau_z = (2^{p-1} - 1)^{-1} \left(2^p \int_{I^p} C_z(\boldsymbol{u}) dC_z(\boldsymbol{u}) - 1\right)$ and we define the corresponding nonparametric estimator to be

$$\hat{\tau}_z = \frac{1}{2^{p-1}-1} \Big(\frac{2^p}{1 - \sum_{i,j=1}^{n} w_i(z,h)^2} \sum_{i,j=1}^{n} w_i(z,h) w_j(z,h) I\{\mathbf{Y}_i < \mathbf{Y}_j\} - 1 \Big).$$

The following theorem establishes the asymptotic results of the proposed estimators in a random design context.

Theorem. Consider the usual assumptions in nonparametric estimation and $h = o(n^{-1/5})$. Then, when $n \to \infty$,

i) $(nh)^{1/2} \left(\hat{C}_z(\boldsymbol{u}) - C_z(\boldsymbol{u}) \right) \xrightarrow{d} C_z^L$,

where C_z^L is a Gaussian variable with zero mean and asymptotic variance $\sigma^2(C_z^L) = d_k f(z)^{-1} C_z(\boldsymbol{u}) \big(1 - C_z(\boldsymbol{u})\big)$. Moreover, if G_z is a Gaussian variable given by

$$G_z = \frac{2^p}{2^{p-1} - 1} \left(\int_{I^p} C_z(\boldsymbol{u}) dC_z^L(\boldsymbol{u}) + \int_{I^p} C_z^L(\boldsymbol{u}) dC_z(\boldsymbol{u}) \right),$$

ii) $(nh)^{1/2} (\hat{\tau}_z - \tau_z) \xrightarrow{d} G_z$,

where the asymptotic variance of $\hat{\tau}_z$ is given by $\sigma^2(G_z)$.

Moreover, we propose a method to select the bandwidth for the nonparametric conditional Kendall's tau estimator based on minimizing its mean squared error and we conduct a simulation study to analyze the performance of the conditional rank correlation in the bivariate context for a linear and a non-linear model. The results show that the bandwidth selection works well for large sample sizes and variables coming from a linear model but the results are still better bias-variance balanced for non-linear models. Moreover, the results remark the importance of the accuracy in the parameter selection when analyzing the dependence between variables.

2.1 Testing for restrictions in conditional dependence

We propose a test for a null hypothesis that can be expressed as

$$H_0 : R\,\boldsymbol{\tau_z} = \boldsymbol{r}, \tag{2}$$

where $\boldsymbol{\tau_z}$ is an m order column vector of estimated Kendall's tau, R is a $q \times m$ order matrix and $\boldsymbol{r}$ a q order column vector with q being the number of linear restrictions to be tested. The alternative is $H_a : R\,\boldsymbol{\tau_z} \neq \boldsymbol{r}$. We define the statistic

$$\mathcal{J}_n = (R\,\hat{\boldsymbol{\tau}}_{\boldsymbol{z}} - \boldsymbol{r})'(RV_{\hat{\tau}_z}R')^{-1}(R\,\hat{\boldsymbol{\tau}}_{\boldsymbol{z}} - \boldsymbol{r}),$$

where $V_{\hat{\tau}_z}$ denotes the variance and covariance matrix of $\hat{\boldsymbol{\tau}}_{\boldsymbol{z}}$. Taking into account the asymptotic results in the Theorem above, $\mathcal{J}_n$ has an asymptotic χ^2 distribution with q degrees of freedom under H_0.

The null hypothesis in expression (2) accounts for many possible situations. In particular, it enables tests to be run for conditionally constant dependence in a random design context. In this particular case, we use a permutation procedure to estimate $V_{\hat{\tau}_z}$. Another application of the $\mathcal{J}_n$ statistic is to test linear restrictions across different samples, as equal conditional dependence structure. In this case, the permutation procedure has been adapted into an appropriate resampling procedure.

To show the performance of the test we run a second simulation study and we consider several cases to calculate size and power under different sample sizes. We also compare the results with the proposal made by Gijbels et al. (2017) for situations where their test can be directly applied. The results show that the statistics perform well for different kinds of restriction, even when quite complex joint distributions are considered. Moreover, it is easy to compute.

3 Empirical application

We apply the previous methodology to detect whether institutional quality helps to increase the relationship between pillars as it does between higher education and innovation. As suggested in previous studies (see e.g. Lucas (1988) and Maradana et al. (2017)), higher education and innovation are directly related to economic growth. In this sense, we are interested in analyzing if the low institutional quality hinders transfers between higher education and innovation results.

Motivated by this, we study the relationship between higher education and innovation, conditional on the institutional quality (INST) level with 2019 regional data from the European Commission. For the sake of illustration and to provide further empirical results, we also consider the dependence between the Efficiency and Innovation groups and between higher education and the other pillars in the Efficiency (Labor market efficiency, L and

Market size, M) and Innovation groups (Business sophistication, BS and Technological Readiness, TR).

In particular, we are interested in four different hypotheses. *i*) Is there concordance between movements in higher education and innovation? *ii*)Is the concordance between higher education and innovation fully explained by institutional quality? ($H_0^{(2)} : \tau_{ij|z} = 0$) *iii*) Does institutional quality explain part of the concordance between higher education and innovation? ($H_0^{(3)} : \tau_{ij|z} = \tau_{ij}$). *iv*) Does the concordance between higher education and innovation depend on the level of quality of institutions? In this case $H_0^{(4)} : \tau_{ij|z_k} = \tau_{ij|z_l}$ is tested. A rejection would provide evidence of a relationship that varies according to institutional quality, which is especially interesting since it provides a starting point for studying how institutional quality affects relationships between pillars making them stronger or weaker. Analyzing this kind of conditional comovements is very useful for policy makers intending to control the impact of their interventions.

First, we remark that there is a significant unconditional relationship between the pillars, as expected.

TABLE 1. Kendall's tau coefficients based on 2019 data.

	τ	$\tau_{q_{0.05}}$	$\tau_{q_{0.15}}$	$\tau_{q_{0.5}}$	$\tau_{q_{0.85}}$	$\tau_{q_{0.95}}$	$H_0^{(2)}$	$H_0^{(3)}$	$H_0^{(4)}$
Eff.-Inn.	0.678	0.097	0.623	0.484	0.204	0.733	***	***	***
HE-L	0.447	0.036	0.248	0.448	0.054	0.357	***	***	*
HE-M	0.197	-0.039	-0.053	-0.294	-0.009	-0.108		***	
HE-TR	0.405	0.077	0.215	0.382	0.052	-0.093	**	***	*
HE-BS	0.308	0.253	0.073	-0.056	0.154	-0.154		***	
HE-I	0.528	0.091	0.307	0.406	0.456	0.530	***	***	**
HE-L-M	0.363	0.025	0.155	0.020	0.151	0.029		***	
TR-BS-I	0.518	0.366	0.517	0.245	0.255	0.069	***	***	**
HE-L-M-TR-BS-I	0.387	0.112	0.225	0.171	0.025	0.118	***	***	

Table 1 contains the estimated Kendall's tau coefficients and summarizes tests results. The results are quite interesting and encourage further study. The joint relationship between pairs such as innovation and higher education is only partially explained by the quality of institutions. Moreover, there is evidence in favor of a joint dependence that increases with the quality of institutions. This may be an interesting starting point for studying whether there is a causal link between institutional quality and the ability of regions to transfer human capital to innovation results. For the other pillars, the results are different. Actually, no conditional dependence is found between HE and BS and for the relationship between HE and TR, INST provides a further contribution for regions with medium institutional quality, which might be linked with specific regions.

For multivariate relationships, there is a low dependence with a constant quality effect between the Efficiency pillars. However, the Innovation group pillars are significantly affected by the institutional quality and show a

higher dependence on lower quality levels. As the results in the last row of Table 1 show, there is no significant effect on the six pillars relationship. Regardless of whether or not conditional dependence is constant, the quality perception effect can vary from one period to another. To test for the consistency of the quality effect over a three-year period we compare the behavior patterns with RCI index data for 2016. The test reveals that in general there are no significant changes in the dependency between pillars from 2016 to 2019 conditional on the quality of institutions.

Acknowledgments: This work is supported by *MEC ECO2014-51914-P, BETS-UFI11/46 and MACLAB-IT93-13, BiRTE-IT1336-19 and PIF16 from UPV/EHU.* We also thank Thorsten Schmidt and the colleges from the Mathematical Institute of the University of Freiburg for all the comments.

References

Embrechts, P., McNeil, A. and Straumann, D. (2002). Correlation and Dependence in Risk Management: Properties and Pitfalls. In: *Risk Management: Value at Risk and Beyond*,Cambridge, M.A.H Dempsted (Ed.), pp 176 – 223.

Gijbels, I., Veraverbeke, N. and Omelka,M. (2011). Conditional Copulas, Association Measures and their Applications. *Computational Statistics & Data Analysis*, **55**, 1919 – 1932.

Gijbels, I., Veraverbeke, N. and Omelka,M. (2017). Nonparametric Testing for no Covariate Effects in Conditional Copulas. *Statistics*, **51**, 475 – 509.

Joe, H. (1990). Multivariate Concordance. *Journal Multivariate Analysis*, **35**, 12 – 30.

Kendall, M. and Smith, B. (1940). On the Method of Paired Comparisons. *Biometrika*, **31**, 324 – 345.

Lucas, R.E. (1988). On the Mechanics of Economic Development. *Journal of Monetary Economics*, **22**, 3 – 42.

Maradana, R.P., Pradhan, R.P., Dash, S., Gaurav, k., Jayakumar, M. and Chatterjee, D. (2017). Does Innovation Promote Economic Growth? Evidence from European Countries. *Journal of Innovation and Entrepreneurship*, **6**, 1 – 23.

Tree-Based Modeling of Discrete Subdistribution Hazards

Moritz Berger[1], Matthias Schmid[1]

[1] University of Bonn / University Hospital Bonn, Germany

E-mail for correspondence: moritz.berger@imbie.uni-bonn.de

Abstract: Subdistribution hazard models are a popular tool for competing risks analysis. The classical approach in discrete time consists of fitting parametric models, which focuses on main effects. An alternative tree-based method is proposed that allows for more flexibility, in particular when interactions between the covariates are present. The method is illustrated by an analysis of age-related macular degeneration among elderly people.

Keywords: Competing risks; Recursive partitioning; Subdistribution hazards.

1 Discrete Subdistribution Hazard Model

Assume that the interest is in the analysis of the observation time to the occurrence of one out of J competing events measured on a discrete time scale $t = 1, 2, \ldots, k$. Let T_i be the event time and C_i the censoring time of individual i with covariate vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, $i = 1, \ldots, n$. For right-censored data, the observation time is defined by $\tilde{T}_i = \min(T_i, C_i)$. It is further assumed that T_i and C_i are independent random variables and that C_i is non-informative for T_i. A key quantity to describe competing risks data is the discrete *cumulative incidence function*, which is defined by $F_j(t|\mathbf{x}_i) := P(T_i \leq t, \epsilon_i = j|\mathbf{x}_i)$, where the event type is represented by the random variable $\epsilon_i \in \{1, \ldots, J\}$. In the following, w.l.o.g., the event of interest and its cumulative incidence function are defined by $\epsilon_i = 1$ and $F_1(t|\mathbf{x}_i)$, respectively. The function F_1 is bounded between 0 and $F_1(k|\mathbf{x}_i) = P(\epsilon_i = 1|\mathbf{x}_i) \leq 1$.
A popular modeling approach for the cumulative incidence function is the proportional subdistribution hazard model (Fine and Gray, 1999), which has been extended to the discrete-time case by Berger et al. (2018). The

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

discrete subdistribution hazard model links $F_1(t|\mathbf{x}_i)$ to the subdistribution time

$$\vartheta_i := \begin{cases} T_i, & \text{if } \epsilon_i = 1 \\ \infty, & \text{if } \epsilon_i \neq 1\,. \end{cases}$$

By definition, ϑ_i measures the time to the occurrence of the type 1 event. Specifically, it assumes that the type 1 event will never be the first event to be observed once a competing event has occurred (Fine and Gray, 1999), implying that there is no finite event time for the occurrence of a type 1 event if $\epsilon_i \neq 1$. Accordingly, the *discrete subdistribution hazard*, which is defined by

$$\begin{aligned} \lambda_1(t|\mathbf{x}_i) &:= P\left(T_i = t, \epsilon_i = 1 | (T_i \geq t) \cup (T_i \leq t-1, \epsilon_i \neq 1), \mathbf{x}_i\right) \\ &= P(\vartheta_i = t | \vartheta_i \geq t, \mathbf{x}_i), \quad t = 1, \ldots, k\,, \end{aligned}$$

represents the discrete hazard function of the subdistribution time ϑ_i. With a little algebra it can be shown that the subdistribution hazard $\lambda_1(t|\mathbf{x}_i)$ is linked to $F_1(t|\mathbf{x}_i)$ by $F_1(t|\mathbf{x}_i) = 1 - \prod_{s=1}^{t}(1 - \lambda_1(s|\mathbf{x}_i)) = 1 - S_1(t|\mathbf{x}_i)$, where $S_1(t|\mathbf{x}_i) = P(\vartheta_i > t|\mathbf{x}_i)$ is the discrete survival function for a type 1 event. Consequently, there is a one-to-one relationship between $\lambda_1(t|\mathbf{x}_i)$ and $F_1(t|\mathbf{x}_i)$. Thus, the effects of the covariates on $\lambda_1(t|\mathbf{x}_i)$ have a direct interpretation in terms of the cumulative incidence of a type 1 event. A parametric model for $\lambda_1(t|\mathbf{x}_i)$ is given by

$$\lambda_1(t|\mathbf{x}_i) = h(\gamma_{0t} + \mathbf{x}_i^\top \boldsymbol{\gamma}),\ t = 1, \ldots, k-1\,, \tag{1}$$

where $h(\cdot)$ is a strictly montone increasing distribution function. The predictor function in (1) is composed of the baseline coefficients $\gamma_{01}, \ldots, \gamma_{0,k-1}$ and the regression coefficients $\boldsymbol{\gamma} \in \mathbb{R}^p$. Estimates of the model parameters in (1) can be derived from a weighted maximum likelihood estimation scheme (Berger et al., 2018) with binary outcome values

$$(y_{i1}, \ldots, y_{i,\tilde{T}_i}, \ldots, y_{i,k-1}) = \begin{cases} (0, \ldots, 0, 1, 0, \ldots, 0), & \text{if } \Delta_i \epsilon_i = 1\,, \\ (0, \ldots, 0, 0, 0, \ldots, 0), & \text{if } \Delta_i \epsilon_i \neq 1\,, \end{cases} \tag{2}$$

indicating if a type 1 event occurred at t or not ($\Delta_i := I(T_i \leq C_i)$). The maximization of the weighted log-likelihood is based on an estimate of the censoring survival function $\hat{G}(t) = \hat{P}(C_i > t)$ and on a vector of weights

$$w_{it} := \frac{\hat{G}(t-1)}{\hat{G}(\min(\tilde{T}_i, t) - 1)} \cdot \left(\mathrm{I}(t \leq \tilde{T}_i) + \mathrm{I}(\tilde{T}_i \leq t-1, \Delta_i \epsilon_i > 1)\right), \tag{3}$$

that equals estimates of the individual-specific conditional probabilities of being (still) at risk for a type 1 event at time t.

2 Recursive Partitioning for Discrete Subdistribution Hazards

The model in Equation (1) assumes that the predictor is a linear function of the covariates. When unknown interactions between covariates are present, an alternative strategy is to apply *recursive partitioning methods* or *trees*. Following the tree-based method by Schmid et al. (2016), which was designed for discrete hazard models with one single type of event, we propose a discrete subdistribution hazard model of the form

$$\lambda_1(t|\mathbf{x}_i) = f_1(t, \mathbf{x}_i)\,, \tag{4}$$

where the function $f_1(\cdot)$ is determined by a Classification and Regression Tree (CART) with binary outcome (Breiman et al., 1984). For tree building, the covariates $x_1, \ldots, x_p$ as well as the time t (coded as an ordinal variable) are considered as candidates for splitting.
When a tree has been constructed, the result is a set of Q terminal nodes that are represented by a set of binary outcome values $y_{i1}, \ldots, y_{i,k-1}$ and corresponding weights $w_{i1}, \ldots, w_{i,k-1}$. Note that because the time t is a candidate splitting variable, each terminal node of the tree corresponds to a subset defined by the covariates and to a time interval $T_q = [a_q, b_q]$, with $1 \leq a_q \leq b_q \leq k$. Therefore, we propose to estimate the subdistribution hazards in each terminal node by

$$\hat{\lambda}_{1q} = \frac{1}{\sum\limits_{i,t \in q} w_{it}} \sum_{i,t \in q} y_{it}\, w_{it}\,, \quad q = 1, \ldots, Q\,,$$

where the weights w_{it} are computed from Equation (3). Concerning the splitting criterion, the classical CART approach is based on impurity measures. Consider, for instance, the Gini impurity in one node m defined by

$$GI(m) = 2\,\hat{\lambda}_{1m}\,(1 - \hat{\lambda}_{1m})\,.$$

Then, in each step of the tree-building algorithm, one chooses the split (among all covariates and split points) that minimizes the weighted sum of the Gini impurities in the children nodes. An important tuning parameter of CART is the tree size, which can be optimized using pruning techniques. Controlling the tree size prevents the resulting subdistribution hazard estimates from having a too large variance (which is inversely related to the terminal node size). Consequently, we propose to consider the minimum number of observations that must exist in a node in order to perform further splits as the main parameter for pruning. This parameter can be optimized by either cross-validation of the log-likelihood or by information criteria such as AIC and BIC.

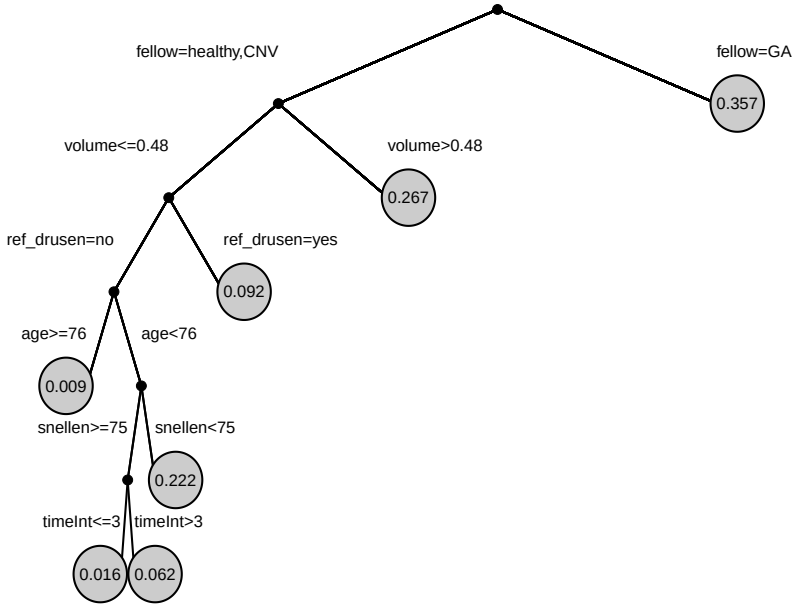


FIGURE 1. Analysis of the AMD data. Tree obtained from fitting the proposed model with log-likelihood based pruning. The numbers at the terminal nodes refer to the estimated subdistribution hazards for GA.

3 Age-Related Macular Degeneration

For illustration, we analyzed the database of the MODIAMD (Molecular Diagnostics of Age-related Macular Degeneration) study, which is an ongoing non-interventional study in patients at high risk for developing late-stage age-related macular degeneration (AMD, Steinberg et al., 2016). AMD either manifests by geographic atrophy (GA) or by choroidal neovascularization (CNV). GA is an advanced stage of AMD with irreversible loss of photoreceptors and severe loss of vision. Therefore, it is of high interest to develop intervention strategies for high-risk patients.

In total, 98 Patients were enrolled between November 2010 and September 2011 at the Department of Ophthalmology, University of Bonn, Germany. All patients were monitored at the time of their inclusion in the study (baseline visit) and subsequently monitored by annual study visits. For our analysis, the data up to and including the fifth annual study visit was available ($t = 1, \ldots, 5$). Exclusion of one patient with missing values in the analyzed risk factors resulted in an analysis data set of size $n = 97$. On completion of the fifth visit, 16 study eyes had developed GA, 25 study eyes had developed CNV, 26 patients were still in the study while 30 patients were censored (i.e., had dropped out at earlier visits). The risk factors incorporated in our analysis were age (years), visual acuity (*snellen*; measured

by the Snellen chart), drusen volume (mm^3), the presence of the natural crystalline lens, smoking, the presence of refractile drusen (*ref_drusen*), and the disease status of the fellow eye (*fellow*; healthy, CNV or GA).
Figure 1 visualizes the result when fitting the proposed tree-based subdistribution hazard model (4) for GA. There is a particularly high risk for the development of GA for patients with GA in the fellow eye ($\hat{\lambda}_{GA} = 0.357$). Without GA in the fellow eye, the risk is high for patients with a large drusen volume ($> 0.48\,mm^3$, $\hat{\lambda}_{GA} = 0.267$), but considerably lower for patients with a smaller drusen volume ($\leq 0.48\,mm^3$).

Acknowledgments: We thank the MODIAMD Study Investigators for providing us with the data.

References

Berger, M., Schmid, M., et al. (2018). Subdistribution hazard models for competing risks in discrete time. *Biostatistics, published online.*

Breiman, L., Friedman J.H., Olshen, R.A. and Stone, J.C. (1984). *Classification and Regression Trees.* Wadsworth, Monterey.

Fine, J.P. and Gray, R.J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**, 496 – 509.

Schmid, M., Kchenhoff, H., Hrauf, A. and Tutz, G. (2016). A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Statistics in Medicine*, **35**, 734 – 751.

Steinberg, J.S., Gbel, A.P., et al. (2016). Development of intraretinal cystoid lesions in eyes with intermediate age-related macular degeneration. *Retina*, **36**, 1548 –1556.

Variational Bayesian inference for sparse high–dimensional Graphical–VAR models

Nicolas Bianco[1], Mauro Bernardi[1], Daniele Bianchi[2]

[1] Department of Statistical Science, University of Padova, Padova, Italy.
[2] School of Economics and Finance, Queen Mary University of London, UK.

E-mail for correspondence: nicolas.bianco@phd.unipd.it

Abstract: We develop a variational approximation method to deal with sparse estimation of high–dimensional graphical vector autoregressive models. The purpose of the project is two–fold. First, we exploit the product density factorisation of the joint variational density that leads to the mean field paradigm, as well as, the representation of the problem as a sequence of auxiliary regressions that rely on the Cholesky factorisation of the precision matrix. A Normal–double–Gamma prior is imposed to shrink toward zero both the autoregression and the precision parameters. The second contribution concerns the solution of the lack–of–identification problem that relies on the employed Cholesky factorisation. We propose to approximate the marginal likelihood of each model permutation by the variational model evidence (ELBO) and to exploit it to get MaP estimates of the model parameters. To explore the space of permutations, when the dimension of the model is large, we develop a new parallel collapsed simulated annealing algorithm (PCSA).

Keywords: Vector autoregressive models; sparsity; variational inference.

1 Introduction

Approximate methods for statistical inference on the parameters of mathematical and statistical models are becoming a relevant issue as model complexity increases. Although variational methods are not confined within the Bayesian framework, the problem of approximating the posterior distribution is relevant and deterministic variational approximations represent viable alternatives to the widely employed simulation–based methods to alleviate the computational burden in high dimensional settings. Moreover, in many situations the marginal likelihood is either not available in closed–form or its computation is time–consuming, preventing the possibility to

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

evaluate and assess the proposed models and estimates using goodness–of–fit procedures. Previous considerations motivate the relevance of approximate inferential techniques to perform Bayesian analysis especially when high–dimensional data or complex models are considered (see, e.g., Robert and Casella 2011). Variational approximations (Ormerod and Wand, 2010) is a body of deterministic techniques for making approximate inference that tackle the problem of optimising a functional over a class of functions in order to minimise a given divergence between a target distribution and a given proposal. They certainly exploit their potentials to provide approximate inference in a likelihood–based context, but they are widely employed within the Bayesian paradigm, where they are also known as variational Bayes methods, (VB). As deterministic alternatives to stochastic approximation methods, such as MCMC methods, they can successfully be employed when the dimension of the problem is large or even in more involved situations where either the data or the model display complex dependence structures. Their major advantages over deterministic approximations rely on the possibility of arbitrarily increasing their accuracy at the expense of computational time. Unlike stochastic methods, deterministic variational algorithms are based on analytical approximation of the target distribution. As a consequence, these methods have limited approximation accuracy, but they offer a relevant gain in terms of computational cost. This paper is devoted to introduce new variational–based inferential procedures and algorithms for estimating high–dimensional vector autoregressive process (VAR) of dimension d. We further assume that $d \gg n$, where n denotes the number of observations, thereby leading to a problem that only admits either sparse or regularised solutions (Rothman et al., 2010). Appropriate prior should be imposed on the elements of the precision matrix to shrink their elements down to zero, thereby enforcing the invertibility of the matrix. We consider the Normal–Gamma prior recently introduced by Griffin and Brown (2010) as an alternative to the Bayesian–Lasso prior of Park and Casella (2008), in a regression context. As for the Bayesian–Lasso, the Normal–Gamma prior of Griffin and Brown (2010) penalises each parameter independently and it is highly inappropriate for the estimation of large–dimensional data. Therefore, the Normal–double–Gamma prior is adapted to the estimation of the precision matrix by adding a common latent factor that jointly penalises towards multiple directions, as in Bitto and Frühwirth-Schnatter (2019). To overcome the lack–of–identification problem, we develop a new parallel collapsed simulated annealing algorithm that maximises the ELBO over the space of permutations. The rest of this paper is organised as follows. Section 2 introduces the model while main applications are discussed in Section 3. Although the proposed methodology is quite general and can be applied in several contexts where appropriate dynamic models are needed, from biology to economics, in this paper, we consider the evolution of financial contagions using international stock indexes, and the analysis of functional magnetic resonance imaging (fMRI) data.

2 Model specification and inference

Let $Y_t = (Y_{1,t}, Y_{2,t}, \ldots, Y_{d,t})^\intercal$ a multivariate random variable, we define the $\mathsf{VAR}(1)$ process as the realisation of the following stochastic difference equation

$$Y_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi} Y_{t-1} + \mathbf{u}_t, \quad \text{for} \quad t = 2, 3, \ldots, \tag{1}$$

where $\boldsymbol{\phi}_0$ is a d–dimensional vector of intercepts, $\boldsymbol{\Phi}$ is a $d \times d$ matrix containing the autoregressive parameters and $\mathbf{u}_t \sim \mathsf{N}_d\left(0, \boldsymbol{\Omega}^{-1}\right)$ for $t = 1, 2, \ldots$ is a sequence of uncorrelated innovation terms such that $\mathbf{u}_{t-k} \perp \mathbf{u}_{t-j}$ for $k \neq j$ and $k, j = \pm 1, \pm 2, \ldots$ and cross–covariance matrix equal to $\boldsymbol{\Omega}^{-1}$, with $\boldsymbol{\Omega} \in \mathbb{S}^d_{++}$ being a positive and definite positive matrix. Exploiting the modified Cholesky factorization of the precision matrix $\boldsymbol{\Omega} = \mathbf{L}^\intercal \mathbf{V} \mathbf{L}$, we can write the process in equation (1) as a $\mathsf{VAR}(1)$ with orthogonal innovations:

$$\mathbf{L} Y_t = \boldsymbol{m} + \mathbf{L}\boldsymbol{\Phi} \mathbf{V}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \text{for} \quad t = 2, 3, \ldots, \tag{2}$$

where $\boldsymbol{\varepsilon}_t \sim \mathsf{N}_d(0, \mathbf{V}^{-1})$, $\mathbf{V}^{-1} = \mathsf{diag}\left\{1/\nu_j,\, j = 1, 2, \ldots, d\right\}$, $\boldsymbol{m} = \mathbf{L}\boldsymbol{\phi}_0$ and $\mathbf{L}$ lower–unitriangular. Leveraging the unit triangular matrix factorisation $\mathbf{B} = \mathbf{I}_d - \mathbf{L}$, allows to rewrite the process in (2) as:

$$Y_t = \boldsymbol{m} + \mathbf{B} Y_t + (\mathbf{I}_d - \mathbf{B})\boldsymbol{\Phi} Y_{t-1} + \boldsymbol{\varepsilon}_t, \quad \text{for} \quad t = 2, 3, \ldots, \tag{3}$$

which is a function of the vector of parameters $\boldsymbol{\vartheta} = (\boldsymbol{\nu}^\intercal, \boldsymbol{m}^\intercal, \boldsymbol{\beta}^\intercal, \boldsymbol{\phi}^\intercal)^\intercal$ with $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d) \in \mathbb{R}^d_+$, $\boldsymbol{m} = (m_1, \ldots, m_d) \in \mathbb{R}^d$, $\boldsymbol{\phi} = \mathsf{vec}(\boldsymbol{\Phi}^\intercal) \in \mathbb{R}^{d^2}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_d)^\intercal \in \mathbb{R}^{d(d-1)/2}$, where $\boldsymbol{\beta}_j$ for $j = 2, \ldots, d$ identifies the j–th row of $\mathbf{B}$ below the main diagonal. (3) represents the main ingredient of the Mean–Field variational Bayes (MFVB) algorithm here proposed. The fully Bayesian approach to the inference here adopted, requires the specification of a prior distribution for all the parameters involved in (3). We assume a Gamma distribution for ν_j and a Normal distribution for m_j, with $j = 1, \ldots, d$, while, following Bitto and Frühwirth–Schnatter (2019), a Normal–double–Gamma prior is imposed to the elements of $\boldsymbol{\beta}$ and for $\boldsymbol{\phi}$. Therefore, we impose joint shrinkage for $\boldsymbol{\beta}_j$ by assuming:

$$\begin{aligned} \beta_{j,k} | \tau_{j,k} \sim \mathsf{N}(0, \tau_{j,k}), \quad \tau_{j,k} | \eta_j, \lambda_j \sim \mathsf{Ga}\left(\eta_j, \frac{\eta_j \lambda_j}{2}\right), \\ \lambda_j \sim \mathsf{Ga}(e_1, e_2), \quad \eta_j \sim \mathsf{Exp}(e_3), \end{aligned}$$

where $k = 1, \ldots, j-1$ and e_1, e_2 and e_3 are fixed hyperparameters. For each row of $\boldsymbol{\Phi}$ we instead assume:

$$\begin{aligned} \phi_{j,s} \sim \mathsf{N}(0, \upsilon_{j,s}), \quad \upsilon_{j,s} | \xi_{j,m}, \kappa_{j,m} \sim \mathsf{Ga}\left(\xi_{j,m}, \frac{\xi_{j,m}\kappa_{j,m}}{2}\right), \\ \kappa_{j,m} \sim \mathsf{Ga}(h_1, h_2), \quad \xi_{j,m} \sim \mathsf{Exp}(h_3), \end{aligned}$$

where $s = 1, \ldots, d$, while $m = 1$ and $m = 2$ indicates whether we are considering an element on the diagonal (hence $s = j$) or off–diagonal respectively and h_1, h_2 and h_3 are fixed hyperparameters. The idea is to distinguish the amount of shrinkage induced on the diagonal and off–diagonal elements of $\boldsymbol{\Phi}$. The most relevant issue in developing the MFVB algorithm consists to define the factorisation of the variational density $q(\boldsymbol{\vartheta})$ which plays a central role in this approximation scheme. Here, we factorise $q(\boldsymbol{\vartheta})$ as follows:

$$q(\boldsymbol{\vartheta}) = q(\boldsymbol{\phi})q(\boldsymbol{\upsilon}, \boldsymbol{\kappa}, \boldsymbol{\xi})q(\boldsymbol{\nu}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\eta}) \tag{4}$$

$$q(\boldsymbol{\upsilon}, \boldsymbol{\kappa}, \boldsymbol{\xi}) = \prod_{j=1}^{d} \left[\prod_{s=1}^{d} q(\upsilon_{j,s}) \prod_{m=1}^{2} q(\kappa_{j,m}) q(\xi_{j,m}) \right] \tag{5}$$

$$q(\boldsymbol{\nu}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\lambda}, \boldsymbol{\eta}) = q(\nu_1) \prod_{j=2}^{d} \left[q(\nu_j) q(\boldsymbol{\beta}_j) \left(\prod_{k=1}^{j-1} q(\tau_{j,k}) \right) q(\lambda_j) q(\eta_j) \right], \tag{6}$$

where $\upsilon_{j,s}, \kappa_{j,m}, \xi_{j,m}, \tau_{j,k}, \lambda_j, \eta_j$ are the latent factors. One of the major novelties of the proposed variational approach relies on the factorisation of the variational distribution. Indeed, as emerges in (4), a joint distribution is imposed on the vector $\boldsymbol{\phi}$ accounting for the dependence among all the elements of $\boldsymbol{\Phi}$. Under the factorisation in (4)–(6) the MFVB algorithm is provided and all the inferential procedures are avaiable for $\boldsymbol{\vartheta}$ within a Bayesian framework. Given the variational densities computed for $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, it is possible to exploit the factorization $\boldsymbol{\Omega} = \mathbf{L}^{\mathrm{T}}\mathbf{V}\mathbf{L}$ to recover an approximation for the posterior distribution of $\boldsymbol{\Omega}$ through simulation from $q(\boldsymbol{\beta})$ and $q(\boldsymbol{\nu})$. It is worth noting that the MFVB approach strongly relies on the representation of the VAR process provided in (3) which is, in turn, obtained by exploiting the modified Cholesky factorisation. This approach works well if there is a relatively clear ordering for the Cholesky decomposition, but such strong assumptions are unlikely to hold in many situations. Therefore, the provided MFVB algorithm suffers from the lack–of–identification problem that originates the non–invariance of the complete likelihood of model in (3) under permutation of the ordering of variables $\{Y_j,\ j = 1, 2, \ldots, d\}$. We deal with the lack–of–identification issue by providing a new parallel collapsed simulated annealing (PCSA) algorithm that leverages the marginal likelihood approximation provided by the ELBO as byproduct of the MFVB. The PCSA is a fast algorithm for optimisation of the non–smooth objective function provided by the ELBO that explores the non–convex space of permutations of the indexes $\{1, 2, \ldots, d\}$.

3 Applications

3.1 International Stock Indexes data

In this section we present an application to financial data. We consider the time series of stock indexes returns for $d = 37$ countries observed for

$T = 189$ months. The estimated conditional dependence graph is provided in Figure 1 (left) where vertexes belonging to the same continent are plotted with the same colour, while Figure 1 (right) displays the estimated partial correlations. The main finding is a stronger correlation among neighbouring countries: as we can see in Figure 1 we are able to identify two main clusters of countries which correspond to Europe (in blue) and Asia (in yellow) and we can notice a sort of block structure of the partial correlation matrix highlighted by red boxes which group together European and Asian countries. However, there are also cross–continent partial correlations: the most evident one concerns the role of USA which seems to be very central in the network with strong connections with some European countries.

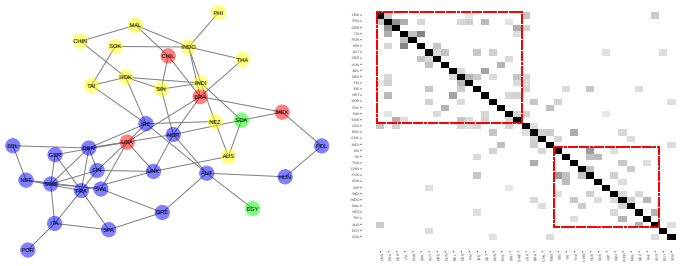


FIGURE 1. Conditional dependence graph (left) and partial correlation matrix (right).

3.2 fMRI data

In this section we consider $d = 68$ time series of length $T = 404$ of fMRI data recorded in particular brain locations according to Desikan atlas. The data are the same used in Gasperoni and Luati (2018). The aim of the analysis is to estimate how the connectivity of the brain's areas changes in different patients affected by different dysfunctions. Results are depicted in Figure 2 where the patient on the left side is healthy, while the one on the right side has a clinical history of alcohol, cannabis and cocaine abuse. It is possible to notice a different pattern and strength of connections according to different characteristics of the subject. We compute the weighted degree index (WDI) as a measure of importance of each area. This index is computed as the sum of the weights of its edges:

$$\mathsf{WDI}_i = \sum_{j \neq i} \widehat{\boldsymbol{R}}_{i,j}, \tag{7}$$

where $\widehat{\boldsymbol{R}}$ is the estimated partial correlation matrix which generates the conditional independence graph. The importance indexes are aggregated

by lobe and cerebral hemisphere. Our results suggest that drugs and alcohol abuse causes a reduction in brain connectivity especially in right frontal (orange points) and right parietal (violet points). Our findings are in accordance with previous studies in neurological science.

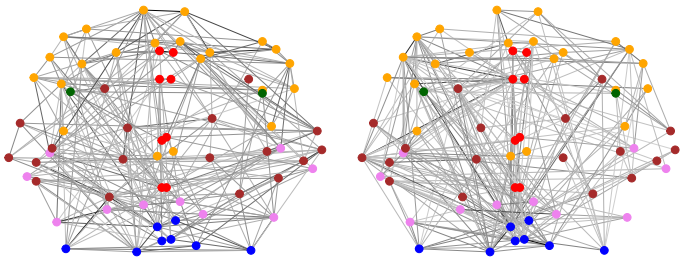


FIGURE 2. Conditionally dependence graph estimated on two patients with different dysfunctions. Greyscale indicates whether the connection is week or strong and different nodes colour refers to different lobes.

References

Bitto, A. and Frühwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, **210**, 75 – 97.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Amer. Statist. Assoc.*, 112(518):859–877.

Gasperoni, F. and Luati, A. (2018). Robust Methods for Detecting Spontaneous Activations in fMRI Data. *Springer Proceedings in Mathematics and Statistics*, **257**.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, **5**, 171 – 188.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *American Statistician*, **64**, 140 – 153.

Robert, C. P. and Casella, G. (2011). A short history of Markov chain Monte Carlo: subjective recollections from incomplete data. *Statistical Science*, **26**, 102 – 115.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.*, **19**, 947 – 962.

Modelling the effect of rural electrification on employment via component-wise boosted causal distributional regression

Guillermo Briseño Sanchez[1], Andreas Groll[1]

[1] Department of Statistics, TU Dortmund University, Germany.

E-mail for correspondence: briseno@statistik.tu-dortmund.de

Abstract: This work is concerned with addressing the issue of variable selection in high-dimensional distributional regression models employed for causal inference. We regularise a Two-Stage Generalised Additive Model for Location, Scale, and Shape (2SGAMLSS) using component-wise gradient boosting in order to obtain a sparse model to assess the causal effect of rural electrification on female and male employment rates using socio-demographic data from South Africa.

Keywords: Causal inference; Boosting; Instrumental variable; Variable selection; GAMLSS.

1 Motivation

We aim at modelling the causal effect of a treatment on two response variables, the change in both female and male employment rates between 1996 and 2001, using a boosted instrumental variable distributional regression approach. The data consists of socio-demographic information on communities located in the KwaZulu-Natal province of South Africa. Each observation is uniquely located in one of ten different districts, suggesting a spatial structure in the data. Additional covariates include demographic control characteristics of each community, such as poverty rate, household density, as well as geographic information like average land inclination, and distance to the nearest road and town. Table 1 displays the preeminent variables in our analysis. A key issue of determining the causal effect of electrification on employment is the selection bias that occurs at the time of assigning an electricity project to an observational unit. Due to political patronage, communities were not randomly targeted for electrification, i.e.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

the treatment is endogenous due to unobserved confounders. Hence, any direct estimation of the treatment effect on the responses will yield biased results.

TABLE 1. Summary statistics for the main variables.

Variable	Description	Mean	S.D.
Δp_female_emp	Diff. share of female employment	-0.00	0.07
Δp_male_emp	Diff. share of male employment	-0.04	0.09
Eskom	Electricity project (Yes=1, No=0)	0.20	0.40
Gradient	Mean land gradient / inclination	10.10	4.89
Number of communities $N = 1816$. Number of districts $G = 10$.			

In order to draw causal inferences from a treatment (Eskom) that exhibits selection bias, we employ Instrumental Variable (IV) methods. An IV or instrument is a regressor that fulfils three key assumptions: (i) It is independent of the unobserved confounders. (ii) It has explanatory power on the endogenous covariate. (iii) It only affects the outcome through the endogenous regressor. Our instrument is given by the average land inclination (Gradient) of a community. The idea behind this particular IV is the fact that a higher land inclination will result in higher costs for an electricity project. Communities that had a lower propensity to receive electricity, but at the same time had a very small or no land inclination (or vice versa) could potentially offset the selection bias. This setting was originally analysed in Dinkelman (2011) using classical IV techniques, which consist of a two-step estimation using ordinary least-squares. We can set up two equations to represent the analysed scenario:

$$\texttt{Eskom}_i = \beta_0^{[1]} + f^{[1]}(\texttt{Gradient}_i) + \sum_{j=1}^{p} f_j^{[1]}(x_{ij}) + \upsilon_i\,, \qquad (1)$$

$$\Delta\texttt{p\_gender\_emp}_i = \beta_0^{[2]} + \texttt{Eskom}_i\beta_1^{[2]} + \sum_{j=1}^{p} f_j^{[2]}(x_{ij}) + \epsilon_i\,, \qquad (2)$$

where Equation (1) is the *treatment equation*, and Equation (2) is the *outcome equation* for either females or males. The superscripts denote the first and second estimation step. We define the *Average Marginal Effect* (AME) on a distributional quantity θ_i of the outcome Y_i given some covariates $X = x_i$ as

$$AME_\theta = \frac{1}{n}\sum_{i=1}^{n}\big(\theta_i(Y_i|d=1, X=x_i) - \theta_i(Y_i|d=0, X=x_i)\big)\,,$$

where $\hat{\theta}_i$ denotes the distributional quantity, e.g. the outcome's standard deviation. The indicator d is the status of the treatment variable, i.e. $d = 1$

for receiving treatment, or $d = 0$ otherwise. Based on Equations (1) and (2), two obstacles arise. One is the choice of a suitable modelling approach with enough flexibility that accommodates the functional form between covariates, treatment, and outcomes, i.e. how should the functions $f_j(\cdot)$ be modelled. The second is the assignment or selection of which regressors impact different characteristics of treatment and outcomes, i.e. which covariates should enter the model. The first obstacle was addressed in Briseño Sanchez et al. (2019), where IV estimation was combined with the high flexibility of the Generalized Additive Models for Location, Scale and Shape (GAMLSS; Stasinopoulos et al., 2018) framework, resulting in the Two-Stage GAMLSS (2SGAMLSS) estimator, hereon referred to as *causal distributional regression*. However, the second issue remains unaddressed. Although highly flexible, causal distributional regression is prone to overfitting due to the lack of an automated variable selection mechanism, i.e. there is no straightforward manner of assigning a specific subset of covariates and their respective representation to any of the response distribution parameters. In order to carry out data-driven variable selection in high-dimensional regression models we turn to component-wise gradient boosting (CGB).

2 Boosting causal distributional regression

Let y_i be a sample of $i = 1, \ldots, n$ responses that are conditionally independent, conditioned on vectors $\mathbf{x}_i$ collecting $j = 1, \ldots, p$ covariates. The responses are assumed to follow a distribution that consists of $k = 1, \ldots, K$ parameters $\vartheta_{k,i}$, i.e.

$$y_i \sim f(y_i|\vartheta_{1,i}, \ldots, \vartheta_{K,i}).$$

In the GAMLSS framework, we can model each distribution parameter $\vartheta_{k,i}$ by relating it to a structured additive predictor $\eta_{k,i}$ via a link function $g_k(\cdot)$:

$$\vartheta_{k,i} = g_k^{-1}(\eta_{k,i}) \quad \Leftrightarrow \quad g_k(\vartheta_{k,i}) = \eta_{k,i} = \beta_{k0} + \sum_{j \in L_k} f_{k,j}(x_{ij}) \,,$$

where $L_k \subseteq \{1, \ldots, p\}$ indicates that each distribution parameter can be modelled by a subset of the covariates in the data. The functions $f_j(\cdot)$ can feature different specifications to accommodate e.g. linear, non-linear or spatial functional forms of the considered regressors. This high flexibility of the GAMLSS framework comes at the cost of increased model complexity, as well as the possibility of inducing bias into the model via an *inappropriate* set of covariates in the predictor of the k-th distribution parameter. CGB allows for a regularised regression framework that retains the flexibility of the GAMLSS approach, but allows us address the inquiry of variable selection for causal distributional regression. Within the CGB famework a univariate regression function (base-learner) $b_{k,j}(\mathbf{x}_{\bullet j}, \boldsymbol{\theta}_{kj})$ is specified for

each considered covariate $\mathbf{x}_{\bullet j} = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ and depends on a vector of unknown parameters $\boldsymbol{\theta}_{kj}$. The index k again emphasizes that the base-learner belongs to the k-th predictor. Each base learner can be specified in order to accommodate a suitable functional form of the regressor. The employed `gamboostLSS` algorithm (Mayr et al., 2014) can be sketched as follows:

(1) Set the boosting iteration counter $m = 0$, initialize the predictors $\hat{\eta}_{k,i}^{[m]}$ with offset values, e.g. using intercept-only models. Increase m by one.

For each distribution parameter $k = 1, \ldots, K$ proceed as follows:

(2) Compute the partial derivative of the log-likelihood w.r.t. η_k: $\frac{\partial \ell(y_i, \boldsymbol{\vartheta})}{\partial \eta_k}$. Plug-in the current estimates $\hat{\boldsymbol{\vartheta}}_i^{[m-1]} = (g_k^{-1}(\hat{\eta}_{k,i}^{[m-1]}))_{k=1,\ldots,K}$. Then set

$$u_{k,i}^{[m-1]} = \left. \frac{\partial \ell(y_i, \boldsymbol{\vartheta})}{\partial \eta_k} \right|_{\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}_i^{[m-1]}}, \quad i = 1, \ldots, n.$$

(3) Fit each of the L_k base-learners in the k-th predictor to the gradient vector $\mathbf{u}_k^{[m-1]}$ and select the base-learner j^* that fits best the gradient vector according to the residual sum of squares. Update the additive predictor

$$\hat{\eta}_k^{[m-1]} = \hat{\eta}_k^{[m-1]} + \nu \cdot b_{k,j^*}(\cdot)\,,$$

where ν is a (weak) learning rate ($0 < \nu \ll 1$).

(4) Set $\hat{\eta}_k^{[m]} = \hat{\eta}_k^{[m-1]}$ and increase m by one.

The algorithm then iterates between steps (2)-(4) until an m_{stop} is reached. Using this algorithm for causal distributional regression, the treatment and outcome equations are boosted in order to obtain a parsimonious, regularised causal model. An optimal value for m_{stop} is determined via cross-validation for the treatment and outcome equation, respectively.

3 Empirical results and discussion

We assume a Bernoulli distribution for the treatment, and a logistic distribution for both outcomes. The logistic distribution depends on a location ϑ_1 and scale ϑ_2 parameter; it is a leptokurtic distribution that accommodates this particular property of both responses. The expectation of a logistically-distributed random variable is given by the parameter ϑ_1 (i.e. we employ the identity link function), whereas the variance is given by $\vartheta_2^2 \pi^2/3$ (i.e. we use a log-link function for the scale parameter). Estimation of the treatment equation via CGB suggests a downward sloping, non-linear effect of `Gradient` on the expectation of `Eskom` as shown in Figure 1 (a).

TABLE 2. Estimated regression coefficients, AME on the mean and standard deviation (s.d.), as well as optimal stopping iterations.

Response	$\hat{\beta}^{\vartheta_1}_{\texttt{Eskom}}$	$\hat{\beta}^{\vartheta_2}_{\texttt{Eskom}}$	$\widehat{AME}_{mean}$	$\widehat{AME}_{s.d.}$
`Male`	−0.021	−0.123	−0.021	−0.005
Treatment equation m_{stop} ($\vartheta_1^{\texttt{Eskom}}$): 3725.				
Male outcome equation m_{stop} ($\vartheta_1^{\texttt{Male}}, \vartheta_2^{\texttt{Male}}$): (195, 1241).				

Such a curve was also observed in Briseño Sanchez et al. (2019) where estimation was carried out via penalized Maximum Likelihood (ML). This is a remarkable result, since the treatment equation model recovers essentially the same functional form of `Gradient` on the treatment and is more parsimonious compared to ML estimation. Figure 1 (b) displays the spatial heterogeneity in the estimated probability of receiving an `Eskom` project. The north-eastern districts of KwaZulu-Natal exhibit the lowest predicted probability of receiving the treatment compared to the rest of the province's districts. The optimal boosting iterations of the treatment and male outcome equation were obtained using 25-fold bootstrap. Results for the female response are not listed because here the `Eskom` treatment was not selected by CGB. Table 2 lists the boosted coefficients for the location (ϑ_1) and scale (ϑ_2) parameters, as well as the AME on the expectation and s.d. of the male response. Note that we decided to report the AME on the s.d. due to the scale of the response variables, otherwise the variance would have been numerically quite small. Due to the identity link function employed on the location parameter ϑ_1, the AME on the mean and $\hat{\beta}^{\vartheta_1}$ coincide. These estimates of the male outcome equation indicate that the treatment induces a reduction in male employment rates of 2.1% on average *cet. par.*, across the communities of KwaZulu-Natal. The `Eskom` treatment has a multiplicative effect on ϑ_2 of $\exp(-0.123) = 0.88$ on average *cet. par.*, i.e. a reduction of the outcome's variance. The AME on the s.d. of the male outcome suggests that not only the expected employment rates decreased, but also the observed employment rates are now closer to the mean male employment rate across communities, i.e. male employment rates become more homogeneous. Figure 1 (c) displays the AME on the mean, whereas Figure 1 (d) shows the AME on the standard deviation (s.d.). Male employment rates of communities located on the southern districts become more homogeneous (s.d. of outcome is reduced) compared to those in the northern districts. We would like to point out that the optimal stopping iteration suggests once again that the CGB model is a more parsimonious version of the ML models previously mentioned. These results indicate that data-driven variable selection can be beneficial in causal distributional regression models via CGB, since the specification of covariate effects is not trivial and could lead to biased treatment effect estimates.

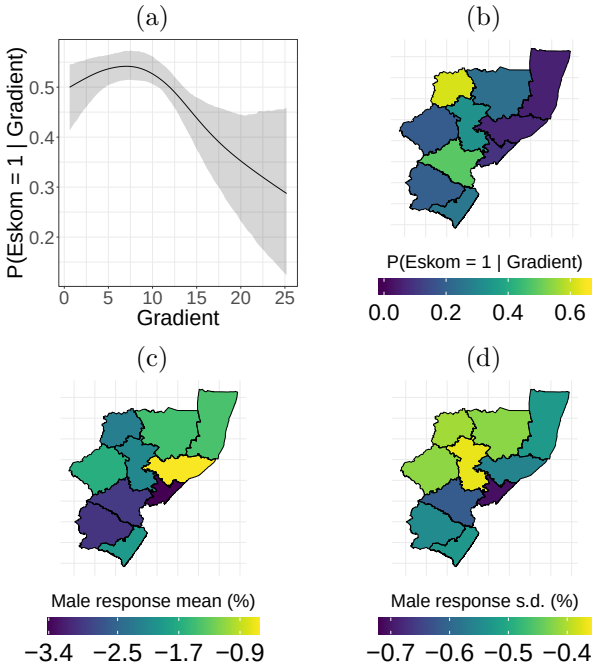


FIGURE 1. Smooth effect of `Gradient` on `Eskom` (a). Probability of receiving an `Eskom` project per `district` (b). AME on mean (c) and standard deviation (s.d.) (d) of the male response.

References

Briseño Sanchez, G., Hohberg, M., Groll, A., and Kneib, T. (2019). Flexible Instrumental Variable Distributional Regression. In: *Proceedings of the 34th International Workshop on Statistical Modelling: Volume II.*, Guimarães, Portugal, 299 – 305.

Dinkelman, T. (2011). The effects of rural electrification on employment: New evidence from South Africa. *American Economic Review*, **101 (7)**, 3078 – 3108.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2014). Generalized additive models for location, scale and shape for high dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61 (3)**, 403 – 427.

Stasinopoulos, M. D., Rigby, R. A., and Bastiani, F. D. (2018). GAMLSS: a distributional regression approach. *Statistical Modelling*, **18 (3-4)**, 248 – 273.

Hazard smoothing along two time scales

Angela Carollo[1,2], Hein Putter[2], Paul H. C. Eilers[3], Jutta Gampe[1]

[1] Max Planck Institute for Demographic Research, Rostock, Germany
[2] Leiden University Medical Center, Leiden, The Netherlands
[3] Erasmus Medical Center, Rotterdam, The Netherlands

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

Abstract: The occurrence of events is mostly modelled by the hazard, and usually one considers only one preferred time scale. Other time scale(s) that may influence the event of interest are incorporated as a (time-varying) covariate(s). Here we propose an approach to estimate the hazard as a smooth bivariate function over two time scales using P-splines. We illustrate the model by analyzing the transition from cohabitation to marriage where the age of the individual and the duration of the cohabitation are relevant. Data come from the German Family Panel (pairfam) and we demonstrate that considering the two time scales jointly provides additional insights about the transition from cohabitation to marriage.

Keywords: Time scales; Multidimensional hazard; P-splines; Cohabitation; Marriage.

1 Introduction

Survival analysis models the time until the occurrence of an event of interest. In many applications, time-to-event data can be measured along several time scales. Clinical examples include the time since disease onset and the time since treatment. Two time scales are also naturally present in the social sciences when modelling the life course. For example, in the transition from cohabitation to marriage age certainty has an important role, but previous research has also pointed out the role of the duration of the cohabitation for the event (marriage).
Usually, time-to-event data are described by means of hazard models. Most popular methods for the estimation of such models, like Cox's proportional hazards model, require the choice of a single time scale and the inclusion of

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

potential other time scale(s) as covariates. However, through the assumption of proportional hazards the possible interplay between different time scales is limited in such an approach. If understanding the joint role of several time dimensions is of interest a different approach is needed.
Here we propose to model the hazard of an event as a bivariate function of two time scales, avoiding the need to choose one preferred time axis. This bivariate hazard is assumed to be smooth and we choose to estimate it by a P-spline approach. Such a smooth hazard surface can capture the interplay between the two time dimensions in a flexible way.
We will present an application of the model to study transitions from cohabitation to marriage by age and by duration of the cohabitation, using data from the German Family Panel (pairfam).

2 Smoothing the hazard over two time scales

We denote with t and s the two time scales. The hazard of an event at (t, s) is denoted by $\lambda(t, s)$, with the log-hazard defined as $\eta(t, s) = \log[\lambda(t, s)]$. A common approach for a flexible hazard model is to assume that it is piecewise constant. The support of the hazard is divided into a grid of cells, mostly rectangles, and for each cell the total exposure and the number of events are calculated. Standard MLE of the hazard is obtained by dividing the number of events by the total exposure time in each cell.
The same estimates are obtained if we view the number of events $E(t, s)$ in each cell as a realization of a Poisson variable with expected value $\mu(t, s) = R(t, s)\lambda(t, s)$ that is, $E(t, s) \sim \text{Pois}(\mu(t, s))$. Here (t, s) denotes the coordinates of a cell (usually represented by its center) and $R(t, s)$ denotes the total at-risk time in the cell.
With a fine grid, we obtain a very flexible hazard, but at the price of an erratic behaviour where fewer individuals are observed. To obtain a smooth surface we use a combination of B-splines bases and difference penalties on the estimated coefficients, known as P-splines. P-splines have been efficiently used to smooth hazards in two dimensions by Currie et al. (2004). Here we will extend this approach to the case of hazards with two time scales.
In most applications the two time scales t and s move at the same speed. So if individual i enters at (t_i, s_i) and exits after Δ time units, the exit point $(\breve{t}_i, \breve{s}_i)$ is given by $\breve{t}_i = t_i + \Delta$ and $\breve{s}_i = s_i + \Delta$. In this way, individuals move along diagonal lines with slope 1 in a Lexis diagram (Keiding, 1990). Consequently, possible combinations of t and s are restricted to the positive half-plane where $s < t$. (For example, the duration of a cohabitation s always is shorter than the age t of the individual.)
In order to overcome this restriction we propose to transform the data into

new points (u, s) by:

$$\begin{pmatrix} u \\ s \end{pmatrix} = \begin{pmatrix} t-s \\ s \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} t \\ s \end{pmatrix}. \tag{1}$$

The transformed data points are now scattered over the whole positive plane, where $u \geq 0$ and $s \geq 0$. (In the example, if t is age and s is the duration of the cohabitation, then $u = t - s$ denotes the age at entry into cohabitation.)

The transformed (u, s)-plane is split into a large grid of small squares, by dividing the u-axis into J intervals and the s-axis into K intervals. Then we compute the $J \times K$ matrix of exposure times $\mathbf{R}$ and the event matrix $\mathbf{E}$. We denote with $\mathbf{B}_u$ and $\mathbf{B}_s$ the two B-spline matrices built over the u- and s-axis with c_u and c_s columns, respectively. The bi-dimensional regressor matrix is obtained as the tensor product of these two B-spline matrices:

$$\mathbf{B} = \mathbf{B}_s \otimes \mathbf{B}_u \tag{2}$$

with dimension $JK \times c_u c_s$. Correspondingly, the vector of coefficients for the B-spline basis is denoted by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{c_u c_s})$ whose elements need to be estimated from the data.

A penalty matrix $\mathbf{P}$ is introduced to tune the amount of smoothing. We denote by $\mathbf{I}_u$ and $\mathbf{I}_s$ the identity matrices of dimension c_u and c_s, respectively, and by $\mathbf{D}_u$ and $\mathbf{D}_s$ differences matrices of order d along the rows (u) and and columns (s). A common choice is a second or a third order penalty. The penalty matrix has two terms, one for the row coefficients and one for the columns, and it is constructed as:

$$\mathbf{P} = \rho_u(\mathbf{I}_s \otimes \mathbf{D}_u^{\mathrm{T}}\mathbf{D}_u) + \rho_s(\mathbf{D}_s^{\mathrm{T}}\mathbf{D}_s \otimes \mathbf{I}_u), \tag{3}$$

where ρ_u and ρ_s are the smoothing parameters. To choose the optimal values of the smoothing parameters AIC is minimized. For this the values of ρ_u and ρ_s are varied over a grid of combinations, on $\log_{10}$-scale, and the model is estimated repeatedly for each combination of values. The optimal smoothing parameters $\hat{\rho}_u$ and $\hat{\rho}_s$ are the ones which minimize the AIC of the model.

3 Application: Marriage after cohabitation

We use data from the first ten waves of the German Family Panel (pairfam), release 10.0, to study the hazard of marriage after having cohabited. The two time dimensions are the age of the individual (t) and the duration of cohabitation (s). Therefore u (the transformed time scale) is the age at which the individual started cohabiting.

Germany is an interesting case study because marriage is very prevalent but attitudes toward marriage differ significantly between West and East

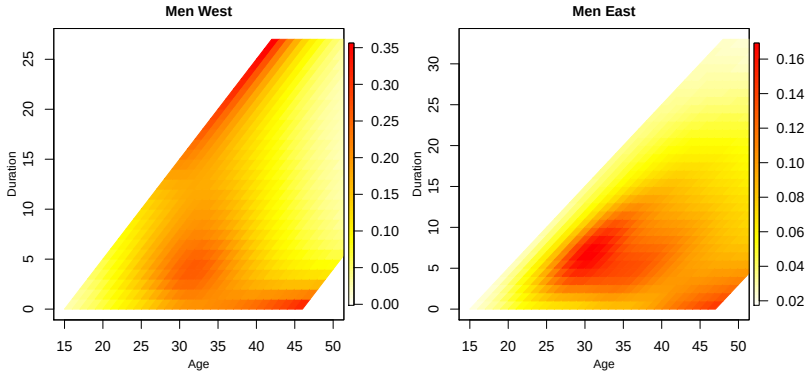


FIGURE 1. Smooth hazard of marrying after a cohabitation by age and duration of the cohabitation.

Germany. We, therefore, estimate the smooth hazard for West and East German men and women separately.
Figure 1 presents the estimated hazards of marriage after cohabiting for West and East German men, plotted in the original plane and only for the observed ages and durations. The hazard of marrying after a period of cohabitation not only shows different levels in East and West Germany, but also quite different age-duration patterns. Also, the interplay between age and duration of cohabitation is rather complex, which would be difficult to capture unless both time dimensions are considered simultaneously.

Acknowledgments: This paper uses data from the German Family Panel pairfam, coordinated by Josef Brüderl, Sonja Drobnič, Karsten Hank, Bernhard Nauck, Franz Neyer, and Sabine Walper. pairfam is funded as long-term project by the German Research Foundation (DFG).

References

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

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **332**, 487 – 509.

Non-stationary spatial model for the distribution of *Xylella fastidiosa* in Alicante.

Martina Cendoya[1], Ana Hubel[12], Antonio Vicent[1], David Conesa[2]

[1] Centre de Protecció Vegetal i Biotecnología, Institut Valencià d'Investigacions Agràries (IVIA). Valencia, Spain.
[2] Departament d'Estadística i Investigació Operativa, Universitat de València. Valencia, Spain.

E-mail for correspondence: cendoya_mar@externos.gva.es

Abstract: Describing the effect of climatic and spatial factors on the geographic distribution of the plant pathogenic bacterium *Xylella fastidiosa* has been the main aim since the moment that it was discovered its presence in Alicante (Spain). This work started with the analysis of the presence/absence data of the pathogen using Bayesian hierarchical models through the integrated nested Laplace approximation methodology and the stochastic partial differential equation approach. Spatial models usually assume stationarity, however, this may be not applicable when physical barriers are present in the study area. Taking into account the irregularities of the terrain and what this may entail in the spread of the disease, higher altitude areas have been considered as possible barriers in the area of interest. The results show that the spatial effect had a strong effect in the model and also that there was no great influence of the barriers due to their reduced extension. Future work will be focused in using these barriers models with theoretical phytosanitary barriers.

Keywords: *Xylella fastidiosa*; INLA; SPDE; Barriers.

1 Introduction

Species distribution models (SDMs) are useful tools to establish which conditions are potentially suitable for the expansion of populations, to evaluate the associations of biotic and abiotic factors with the geographic extent of the species, as well as to predict the species distribution in space and time. These types of models can be developed through different methodologies. However, in many cases, they ignore the spatial dependence which usually

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

exists among the geographical locations of the observations. This can lead to an overestimation of the parameters and to establish erroneus relationships between observations and covariates. Spatial Bayesian hierarchical models allow the inclusion of spatial autocorrelation.

Spatial models are usually based on the fact that the spatial correlation between observations only depends on the Euclidean distance between locations, i.e. that assumes that is stationary and isotropic. Nevertheless, this assumption may lead to a bias in the prediction of species distribution when there are dispersal barriers in the study area (Bakka et al., 2019, Martínez-Minaya et al., 2019).

Xylella fastidiosa was detected in 2017 in Alicante (Spain), affecting mainly almond trees, although it has also been detected in other plant species. The first interest in this study was to analyze the effects of climatic and spatial factors on the distribution of the pathogen. But taking into account that the study region had a variable topography, with areas at sea level and mountains over than 1500 m of altitude, the areas with the highest altitude were considered as physical barriers.

2 Data and modeling

Data were considered as continuous locations that occur within a defined spatial domain (geostatistical data). Presence/absence data of *X. fastidiosa* from the official surveys in 2018 in Alicante were analyzed using a Bayesian hierarchical model through the Integrated Nested Laplace Approximation methodology (Rue et al., 2009). The spatial effect was included using the Matérn covariance function, approximated as a solution to a stochastic partial differential equation (Lindgren et al., 2011).

The mean of the response variable Y_i was linked to a structural predictor which included the effect of covariates and spatial effect in an additive way:

$$g(\pi_i) = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + u(s_i),$$

where β_0 is the intercept, β_m are the coefficients of the covariates x_m, π_i is the probability of presence at location i, and u_i represents the spatial random effect.

Following Bakka et al. (2019), taking into account a non-stationary process, in the areas with barriers the correlation was eliminated by introducing a different Matérn field, with the same variance (σ) but a range (r) close to zero. Thus, $u(\boldsymbol{s})$ is the solution to a system of stochastic differential equations that includes the normal area with the area of the barriers. In this case, the areas with highest altitude were established as barriers (above 1065 m).

Climatic variables for Alicante were obtained from the WorldClim v.2 database (Fick and Hijmans, 2017). Due to the high linear correlation found

among the climatic variables, a selection of variables was made prior to the modeling process, where the collinearity was evaluated by means of the variance inflation factor (VIF). Once the climatic variables to be included in the model were pre-selected and taking into account the spatial effect, a model selection was made based on two criteria: the Watanabe Akaike information criterion (WAIC) (Watanabe, 2010), which indicates the goodness of fit of the model; and the logarithmic conditional predictive ordinate (LCPO) (Roos and Held, 2011), which evaluates the predictive capacity.

3 Results and discussion

Based on the linear correlation and the value of the VIF, the pre-selected climatic variables were: mean diurnal range (*bio2*), mean temperature of wettest quarter (*bio8*) and precipitation of wettest month (*bio13*). The combination of these three climatic covariates and the spatial effect resulted in 16 models to evaluate. According to WAIC and LCPO criteria, the one that included the covariate *bio13* and the spatial effect was selected as the best model.

The probability that the posterior distribution of the parameter for *bio13* was less than zero was 0.94, therefore, it was considered relevant in the model. The effect of this covariate on the model would imply that areas with higher precipitation in the wettest month would have lower probability of the presence of *X. fastidiosa*.

Figure 1 shows the mean and standard deviation of the predictive posterior distribution. Athough the covariate *bio13* was considered relevant in the model, a strong influence of the spatial component in the model was observed. In this way, the highest probability of the presence of *X. fastidiosa* was found in those areas where the spatial effect had higher values.

Due to the small extent of the barriers considered in the study area of *X. fastidiosa*, they did not have a major impact on the spatial component, nevertheless, it was observed a smoothing effect around the areas with higher altitude. In the study of species distributions, the elements that are barriers for dispersal cannot be ignored, since it would be wrongly assumed that the species can be found in areas where it would be actually impossible to be present.

Acknowledgments: We thank Generalitat Valenciana for providing the survey data. The present work has received funding from the European Unions Horizon 2020 research and innovation programme under Grant Agreement No. 727987 - XF-ACTORS; from Project E-RTA 2017-00004-C06-01 FEDER INIA-AEI Ministerio de Ciencia, Innovacin y Universidades and Organizacin Interprofesional del Aceite de Oliva Espaol, Spain; and finally from grants MTM2016-77501-P TEC2016-81900-REDT from the

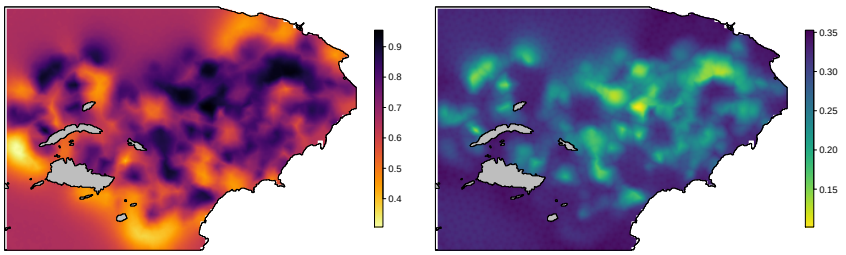


FIGURE 1. Mean (left) and standard deviation (right) of the posterior predictive distribution of the probability of presence of *Xylella fastidiosa*.

Spanish Ministry of Science, Innovation and Universities State Research Agency. MC held an IVIA grant partially funded by the European Social Fund Comunitat Valenciana 2014-2020.

References

Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D., and Rue, H. (2019). Non-stationary Gaussian models with physical barriers. *Spatial Statistics*, **29**, 268 – 288.

Fick, S. E. and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, **37**, 4302 – 4315.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423 – 498.

Martínez-Minaya, J., Conesa, D., Bakka, H., and Pennino, M. G. (2019). Dealing with physical barriers in bottlenose dolphin (Tursiops truncatus)distribution. *Ecological Modelling*, **406**, 44 – 49.

Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, **6**, 259 – 278.

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

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571 – 3594.

Bayesian Structured Antedependence Model Proposals for Longitudinal Data

Edilberto Cepeda-Cuervo[1], Vicente Núñez-Antón[2]

[1] Universidad Nacional de Colombia, Colombia
[2] University of the Basque Country UPV/EHU, Spain

E-mail for correspondence: ecepedac@unal.edu.co

Abstract: An important problem in Statistics is the study of longitudinal data taking into account the effect of explanatory variables such as treatments and time and, at the same time, incorporate into the model the time dependence between observations on the same individual. The latter is specially relevant in the case of having nonstationary correlation, as well as nonconstant variance for the different time point at which measurements are taken. Antedependence (AD) models constitute a well known commonly used set of models that can accommodate this behavior. In this paper, a new Bayesian approach for analyzing longitudinal data within the context of antedependence models is proposed. This innovative approach takes into account the possibility of having nonstationary correlations and variances, and proposes a robust and computationally efficient estimation method for this type of data. We consider the joint modelling of the mean and covariance structures for the general AD model, estimating their parameters in a longitudinal data context. Our Bayesian approach is based on a generalization of the Gibbs sampling and Metropolis-Hastings by blocks algorithm, properly adapted to the AD models longitudinal data settings. Finally, we illustrate the proposed methodology by analyzing the race dataset.

Keywords: Antedependence models; Bayesian methods; Mean-covariance modelling; Nonstationary correlation.

1 Introduction

Continuous longitudinal data consist of repeated measurements on the same subject over time. These measurements are typically correlated and there have been several proposals in the literature to handle stationary or nonstationary correlations and variances, as well as balanced or unbalanced longitudinal data. A general fixed effects regression model for longitudinal

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

data can be defined by assuming that the response variable $\boldsymbol{Y_i}$ can be explained with the model given by $\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon_i}$, $i = 1, \ldots, m$, where $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{in_i})^T$ is the $n_i \times 1$ vector of responses for subject i, X_i is the $n_i \times q$ design matrix of rank q, which includes the covariates for the i-th subject; $\boldsymbol{\epsilon_i}$ is the $n_i \times 1$ vector of errors, assumed to follow a multivariate normal distribution with mean $\mathbf{0}$, and variance-covariance matrix $\Sigma_i(\boldsymbol{\theta}) = \sigma^2 V_{0i}$, whereas $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ are k and q-dimensional vectors of unknown parameters for the variance-covariance and mean model, respectively. Here, n_i represents the number of observations available for the i-th subject. If $n_i = n$, $\forall i$, we have a balanced data set. In addition, m represents the number of individuals in the study, which are assumed to be independent from one another.

Fitting longitudinal models can be carried out by using maximum likelihood estimation methods, such as the Newton Raphson, the EM algorithms, restricted maximum likelihood or alternative Bayesian methodological proposals. A Bayesian proposal with no specific variance-covariance structure assumes a multivariate normal prior distribution for the mean regression parameters and a Wishart prior distribution for the covariance structure. A second approach assumes regression structures in both the mean and the variance-covariance matrix of normal variables. This approach is based on the modelling proposal which uses the Cholesky's matrix decomposition.

In this paper we propose a Bayesian method for the joint estimation of the mean and covariance parameters in the regression longitudinal models settings under the normality assumption, and also allowing for the specification of several different variance-covariance structures. Our proposals start by considering variance-covariance models with stationary correlations and homogeneous variances, as is the case in the CS, AR(1) and ARMA(1,1) models, so that they are then generalized to consider nonstationary correlations and heterogeneous variances, such as is the case in the structured antedependence model of order one, or SAD(1) model. That is, we extend the previous proposal to consider parametric more parsimonious variance-covariance models that have been shown to be more useful in longitudinal data settings than those of the unstructured AD model previously considered therein. In order to illustrate the performance of the proposed methodology, it was applied to fit longitudinal models with structured AD of order one, SAD(1), covariance structures to the 100-km race dataset.

1.1 Bayesian estimation proposals

In longitudinal models, if $\boldsymbol{Y} = (\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_m})^T$ denotes the vector of responses for all of the m individuals in the study, having a design matrix $X = (X_1^T, X_2^T, \ldots, X_m^T)^T$, we have that $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon_1}, \ldots, \boldsymbol{\epsilon_m})^T$ is a vector of random errors associated to the corresponding component in the responses vector $\boldsymbol{Y}$, so that the $\boldsymbol{\epsilon_i}$'s are assumed to be independent from each other, $\boldsymbol{\epsilon} \sim MVN$ with mean $\mathbf{0}$ and block diago-

nal variance-covariance $\Sigma(\boldsymbol{\theta})$ with diagonal components $\Sigma_1(\boldsymbol{\theta}), \ldots, \Sigma_m(\boldsymbol{\theta})$. Thus, under mean and variance-covariance model assumptions, apart from a constant term, the likelihood function is given by:

$$\mathbb{L}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{Y}) \propto \prod_{i=1}^{m} |\Sigma_i(\boldsymbol{\theta})|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[(\boldsymbol{Y} - X\boldsymbol{\beta})^T \Sigma_i^{-1}(\boldsymbol{\theta})(\boldsymbol{Y} - X\boldsymbol{\beta})\right]\right\},$$

where the $\Sigma_i(\boldsymbol{\theta})'s$ are assumed to follow: (1) a compound symmetry (CS), equicovariance or equicorrelation model, with $\boldsymbol{\theta} = (\sigma^2, \rho)^T$, $\text{Var}(Y_{ij}) = \sigma^2$, $j = 1, \ldots, n_i$, and $\text{Cor}(Y_{ij}, Y_{il}) = \rho$; (2) a first order autoregressive regression structure AR(1) model, with $\boldsymbol{\theta} = (\sigma^2, \rho)^T$, $\text{Var}(Y_{ij}) = \sigma^2$ and $\text{Cor}(Y_{ij}, Y_{il}) = \rho^{|t_{ij} - t_{il}|}$, $j \neq l$; (3) an autorregressive with moving average ARMA(1,1) model, with $\boldsymbol{\theta} = (\sigma^2, \rho, \phi)^T$, $\text{Var}(Y_{ij}) = \sigma^2$, $\text{Cor}(Y_{ij}, Y_{il}) = \phi$ if $|t_{ij} - t_{il}| = 1$, and $\text{Cor}(Y_{ij}, Y_{il}) = \phi\rho^{|t_{ij} - t_{il}|-1}$ if $|t_{ij} - t_{il}| > 1$; and (4) a structured antedependence (SAD) model, where the components of $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\rho}^T, \lambda, \boldsymbol{\psi})^T$ are given by $\text{Cor}(Y_{ij}, Y_{i,j-k}) = \rho_{j,j-k} = \rho_k^{f(t_{ij},\lambda_k) - f(t_{i,j-k},\lambda_k)}$, $j = s+1, \ldots, n$, $k = 1, \ldots, s$, and $\sigma_j^2 = \sigma^2 G(t_{ij}, \boldsymbol{\psi})$, $j = 1, \ldots, n$, whit $f(t_{ij}, \lambda_k) = (t_{ij}^{\lambda_k} - 1)/\lambda_k$, if $\lambda_k \neq 0$ and $f(t_{ij}, \lambda_k) = \log(t_{ij})$, if $\lambda_k = 0$.

Thus, assuming independent prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the posterior parameter distribution is given by $\pi(\boldsymbol{\theta}) \propto \mathbb{L}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{Y}) p(\boldsymbol{\beta}) p(\boldsymbol{\theta})$, where $p(\boldsymbol{\beta})$ is the prior distribution of $\boldsymbol{\beta}$, assumed to be a multivariate normal distribution, and $p(\boldsymbol{\theta})$ the prior distribution of $\boldsymbol{\theta}$, defined according to the variance-covariance structure. Then, samples of $\boldsymbol{\beta}$ are obtained from their posterior conditional distribution, a multivariate normal distribution. For $\boldsymbol{\theta}$, assuming prior independence between their parameters components and that $\lambda_k = \lambda$, for all k, the following prior distribution were assumed: $p(\varphi) \equiv \text{Gamma}(g_0/2, g_0\sigma_0^2/2)$, where $\varphi = 1/\sigma^2$; $p(\lambda) = U(-a, a)$; $p(\rho) \equiv \text{Beta}(a, b)$, and a multivariate prior normal distribution $p(\boldsymbol{\psi}) \equiv MVN(\boldsymbol{\psi_0}, K_0)$ for $\boldsymbol{\psi} = (\psi_0, \psi_1, \ldots, \psi_r)^T$, in order to build a normal kernel transition function. Thus, samples of φ can be obtained from their full conditional posterior distribution, which is a gamma distribution. Samples of ρ, λ and $\boldsymbol{\psi}$ are obtained from the posterior conditional distributions by applying the Metropolis Hastings algorithms, defining appropriate kernels transition function. That is, appropriate kernel transition functions should be defined in order to attain a reasonable efficiency for the proposed algorithm. As for ϕ and λ, a kernel transition function, such as the one assumed for ρ and given in equation (1), is also assumed.

$$q(\rho^{(*)}|\rho^{(k)}) = \begin{cases} \rho^{(*)} \sim U(0, 2\rho^{(k)}) & \rho^{(k)} \leq 0.5 \\ \rho^{(*)} \sim U(2\rho^{(k)} - 1, 1) & \rho^{(k)} > 0.5 \end{cases} \tag{1}$$

For $\boldsymbol{\psi}$, we assume a kernel transition function given by the observational model obtained from $\tilde{Y}_j = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_{ij} - \bar{Y}_j\right)^2$, where $\bar{Y}_j = \frac{1}{m}\sum_{i=1}^{m} Y_{ij}$,

and by assuming, without loss of generality, that $n_i = n$, and that the working observational model

$$\widetilde{w}_j = \log(\tilde{Y}_j) = \psi_0 + \psi_1 X_{1j} + \cdots + \psi_p X_{pj} + \varepsilon_j \tag{2}$$

follows a normal distribution, where $\varepsilon_j \in N(0, \sigma^2)$, with σ^2 known, and such that $\tilde{\boldsymbol{X}}_j = (1, X_{1j}, \ldots, X_{pj})$ and $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}_1^T, \ldots, \tilde{\boldsymbol{X}}_n^T)^T$. Thus, the kernel transition function for $q(\boldsymbol{\psi})$ is obtained from the combination of the normal prior distribution and the observational model in (2). That is,

$$\pi_{q(\boldsymbol{\psi})} \equiv N(\mu_{\boldsymbol{\psi}}, K_{\boldsymbol{\psi}}), \tag{3}$$

where $\mu_{\boldsymbol{\psi}} = K_{\boldsymbol{\psi}}(K_0^{-1}\boldsymbol{\psi}_0 + \tilde{X}^T\tilde{\Sigma}^{-1}\tilde{W})$, with $\tilde{\Sigma} = \text{diag}(\sigma^2)$, $\tilde{W} = (\widetilde{w}_1, \ldots, \widetilde{w}_n)^T$ and $K_{\boldsymbol{\psi}} = (K_0^{-1} + \tilde{X}^T\tilde{\Sigma}^{-1}\tilde{X})^{-1}$.

2 Application: 100-Km Race Data

The 100-Km Race Data correspond to each of the partial times in minutes for each of the 80 competitors in each of the 10-kilometer sections of a 100-km race in the United Kingdom in 1984. The objective is to find a parsimonious model describing in the best possible way how competitor's performance on each 10-km section and performance on previous sections. Based on shape of the data set we assume a cubic in time mean regression model $Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \epsilon_{ij}$, $i = 1, \ldots, 80$, and, based on the sample correlation and variance values for this dataset, variances increase as the race progresses, and correlations are nonstationary, so that we propose the use of a variance-covariance structure having an SAD(1) model given by:

$$\rho_{j,j-k} = \rho^{f(t_{ij},\lambda) - f(t_{i,j-k},\lambda)}, \; j = s+1, \ldots, n; \; k = 1, \ldots, s \tag{4}$$

$$\sigma_j^2 = \exp(\psi_0 + \psi_1 t_{ij} + \psi_2 t_{ij}^2), \quad j = 1, \ldots, n, \tag{5}$$

where $f(t_{ij}, \lambda) = (t_{ij}^{\lambda} - 1)/\lambda$, if $\lambda \neq 0$ and $f(t_{ij}, \lambda) = \log(t_{ij})$, if $\lambda = 0$. The proposed Bayesian method shows a good performance, showing a small transient period and parameter estimates that agree with the dataset behavior. The AIC, DIC and BIC values are small for the proposed model compared with those obtained in previous analysis. Results include the regression parameter estimated mean values under the Bayesian proposal, together with their respective standard deviations, and including median values, as well as estimates obtained by restricted maximum Likelihood methods (REML). Table 1 presents the posterior mean parameter estimates, obtained under the Bayesian proposal for the Type 3 - SAD variance-covariance structure, together with their respective standard deviations within parentheses, including median values, and parameter estimates under REML-methods for the 100-km race dataset. In the Bayesian proposal

for the Type 3 - SAD model, there is a slight difference with the one assumed in previous analysis (i.e., those applying the REML methods), where the proposed variance function is $\sigma_j^2 = \sigma^2(1+\psi_1 t_{ij}+\psi_2 t_{ij}^2)$, $j = 1, \ldots, 10$.

TABLE 1. Mean parameter estimates for the Type 3 - SAD model.

Parameter	**Mean**	**Median**	**REML-estimates**
β_0	44.585 (1.632)	44.573	43.428
β_1	-2.410 (2.102)	-2.421	1.354
β_2	1.327 (0.752)	1.326	0.253
β_3	-0.097 (0.072)	-0.097	-0.017

Table 2 includes the estimated values for the variance-covariance parameters under the Bayesian proposal, together with their respective standard deviations, and including median values, as well as estimates obtained by restricted maximum Likelihood methods (REML), when available, where standard deviations for the variance-covariance parameters were not provided. In any case and in order to be able to compare the estimated variances at each split time, we also include their REML-estimates for the variance parameters: $\hat{\sigma}^2 = 16.952$, $\hat{\psi}_1 = 0.590$, and $\hat{\psi}_2 = 0.450$.

TABLE 2. Variance parameter estimates for the Type 3 - SAD model.

Parameter	**Mean**	**Median**	**REML-estimates**
ρ	0.918 (0.031)	0.924	0.929
λ	1.680 (0.261)	1.684	1.600
ψ_0	2.771 (0.308)	2.767	–
ψ_1	0.677 (2.128)	0.683	–
ψ_2	-0.034 (0.021)	-0.034	–

3 Conclusions

We have proposed alternative Bayesian longitudinal models for fitting compound symmetry, autoregressive of order one, autoregressive with moving averages, as well as structured antedependence models for nonstationary in variance and/or correlation longitudinal data settings. In this paper, we assume flexible prior distributions, specific methods to obtain samples of the conditional posterior distribution are proposed. The usefulness of the proposed method was illustrated with the analysis of the 100-km race dataset, and results were compared to those obtained by restricted maximum likelihood methods. Results suggested that the proposed methods behave well

under very general conditions, and estimated values were similar to those obtained by classic methods. However, classic methods require specific programming, whereas the proposed Bayesian methods can be easily adjusted to the data sets under study by using very flexible and easy programming, as well as general available software, such as R.

Acknowledgments: This work was supported by the Department of Statistics, Faculty of Sciences of the Universidad Nacional de Colombia, Ministerio de Economía y Competitividad, Agencia Estatal de Investigación (AEI), Fondo Europeo de Desarrollo Regional (FEDER), the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group), and Universidad del País Vasco UPV/EHU under research grants MTM2016-74931-P (AEI/FEDER, UE), IT1359-19 and UFI11/03.

References

Castillo-Carreno, E., Cepeda-Cuervo, E., and Núñez-Antón, V. (2020). Bayesian structured antedependence model proposals for longitudinal data. *SORT*. In press.

Cepeda-Cuervo, E. (2001). Modelagem da Variabilidade em Modelos Lineares Generalizados. Unpublished Math Ph.D. Thesis. Mathematics Institute, Universidade Federal do Río de Janeiro, Brazil.

Cepeda, E., and Gamerman, D. (2004). Bayesian modeling of joint regressions for the mean and covariance matrix. *Biometrical Journal*, 46(4), 430 – 440.

Cepeda, E., and Gamerman, D. (2005). Bayesian methodology for modeling parameters in the two parameter exponential family. *Revista Estadística*, 57, 168 – 169.

Cepeda-Cuervo, E., and Núñez-Antón, V. (2009). Bayesian modelling of the mean and covariance matrix in normal nonlinear models. *Journal of Statistical Computation and Simulation*, 79(6), 837 – 853.

Zimmerman, D.L and Núñez-Antón, V. (2010). *Antedependence Models for Longitudinal Data*. New York: CRC Press.

Bayesian concurrent functional regression for sparse data.

Beatrice Charamba[1], Andrew J Simpkin[1]

[1] National University of Ireland Galway, Ireland

E-mail for correspondence: b.charamba1@nuigalway.ie

Abstract: The recent abundance of wearable technology has led to a sharp rise in the availability of multivariate data streams. However, many functional data analysis (FDA) methods require such data to be measured regularly without missingness, with data being collected at the same fixed times for all individuals. In order to deal with irregular, concurrent, functional data including missing values, we developed the Bayesian model for function-on-function regression. This method is tested in a simulation study and applied to concurrently measured glucose (every 5 minutes for 1 week) and electrocardiogram (ECG) data (every 10 minutes for 1 week) in a cohort of $n = 17$ type 1 diabetics. The Bayesian model outperformed other models when the underlying relationship is complex and non-linear.

Keywords: Functional concurrent regression; Functional data analysis; Bayesian Models

1 Introduction

Functional data analysis (FDA)(Ramsay and Silverman 2006) assumes that the observed data (e.g. recorded over time for an individual) are a stochastic process and that the data can be represented as functions. Methods for FDA include functional principal components analysis (Yao et al. 2011), functional correlation and functional regression models (Goldsmith and Schwartz 2017) among others. In the early years of FDA, application focussed on data measured on a dense and regular grid. However, many data are collected irregularly over time with each subject having a different number of samples and different sampling points. Years ago, such data had few replicates and were analysed using longitudinal data techniques such as linear mixed models (Laird and

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

Ware 1982). With developments in sensor technology, functional data (e.g. from wearable devices) can have thousands of replicates measured per individual and sampling points which differ from person to person. These data maybe highly non-linear, and hence FDA techniques are best placed to maximize their value. In this paper, we focus on modelling the concurrent relationship between functional variables, where both the response and predictor variables are functional and measured on the same domain (i.e. the functional concurrent model or FCM).

The frequentist methods readily available to fit FCM to irregular data with missingness use only complete cases (Leroux et al. 2018). Other methods considered for such data are not available in software. Bayesian models can use all the data available in such cases. However, models considered so far in the Bayesian framework only fit data collected on a regular grid (Crainiceanu and Goldsmith 2010). In this paper we develop and test a Bayesian model that can fit FCM to irregular data including missing values.

2 Methods

2.1 Functional Concurrent model (FCM)

We consider both functional responses and covariates which are irregularly and sparsely measured. When both response and predictor are functions, the function-on-function model regression model is popular. A special type of function-on-function regression is the functional concurrent regression model which estimates the concurrent relationship between the response and predictors dynamically across the same domain t. The observed data are (Y_{ij}, X_{ij}) which denote the i^{th} individual's measurement at time j. The model with one covariate is given by the formula (Leroux et al. (2018))

$$Y_i(t) = f_0(t) + X_i(t)f_1(t) + b_i(t) + \epsilon_i(t) \tag{1}$$

where $Y_i(t)$ is the response at time i for the response and $f_0(t)$ is the constant function, $X_(t)$ is the functional predictor at at time j, $b_i(t)$ are subject specific deviations from the intercept function and $\epsilon_(t)$ are independent random errors. Ivanescu et al. (2018), Febrero-Bande et al.(2012) and Ramsay and Silverman (2006) developed models to fit the FCM. These models can fit sparse data measured on a regular grid or irregular data on a dense grid. For irregular and sparse data, Leroux et al. (2018) developed the fcr package. A concurrent relationship can also be estimated by the additive model using the *mgcv* R package. In the Bayesian framework, Goldsmith et al. (2017) developed an R package (*vbvs.concurrent* package) for variable selection in the FCM model which can estimate the parameter function. However, the fcr, *mgcv*, and *vbvs.concurrent* packages use complete cases only. In addition, the *vbvs.concurrent* does not produce inferences.

2.2 Bayesian Functional Concurrent Model

The functional parameters, $f_i(t)$ in Model 1 were expressed in terms of basis functions. Let $\mathbf{B}(t) = \{B_1(t), \cdots, B_c(t)\}'$ be a sequence of basis functions evaluated at t, where c is the number of basis functions. Then let $f_i(t) = \mathbf{B}(t)'\Theta$ where $\Theta = (\theta_1, \cdots, \theta_c)'$ is the vector of coefficients. The subject specific deviations are also modelled using basis functions $b_i(t) = \sum_{i=1}^{c} u_{ik}B_k(t)$ where, $\mathbf{u}_i = (u_{i1}, \cdots, u_{ic})'$. Let $\mathbf{y}_i = (Y_{i1}, \cdots, Y_{im_i})'$ and $\mathbf{X}_i = (X_{i1}, \cdots, X_{im_i})'$, $\mathbf{B}_i = [\mathbf{B}(t_{i1}), \cdots, \mathbf{B}(t_{im_i})]'$ and

$$\mathbf{Z}_i = [\mathbf{B}_i, diag(X_{1,i})\mathbf{B}_i, \cdots, diag(X_{p,i})\mathbf{B}_i].$$

With this notation, the model can be expressed as a mixed model

$$\mathbf{y}_i = \mathbf{Z}_i'\Theta + \mathbf{u}_i'\mathbf{B}_i$$

which can be fitted using Bayesian methods or any other software, and Bayesian methods can easily be applied. The parameter functions were then estimated separately as $f_i(t) = \mathbf{B}(t)'\Theta$.The model was fitted in R using the *R2jags* package. The response was assumed to be normally distributed. The design matrix $\mathbf{Z}$ was obtained from the data and the smoothness on Θ was imposed using the first order random walk. Subject specific $\mathbf{u_i}$ was assumed to follow normal with covariance Γ. Specifically; $Y_{ij} \sim N(\mathbf{X}\Theta, \sigma^2)$, $\theta_i \sim N(\theta_{i-1}, 10000)$, $i = 2, \cdots, c$, and $\mathbf{u_i} \sim N(0, \Gamma)$. The parameter functions were obtained as $f_j(t) = \sum_{i=1}^{c} \theta_i\mathbf{B}_i, j = 0, 1$, these were set as stochastic nodes in the model.

2.3 Simulation study

A simulation study was performed to compare the Bayesian model with the *fcr*, *vbvs.concurrent* and *mgcv* models to determine which performs best when there are sparse and irregular data. The Bayesian model was fitted using the *R2jags* package and convergence was checked using trace and history plots. The model simulated was

$$Y_i(t) = f_0(t) + f_1(t)W_i(t) + \epsilon_i(t)$$

with $W_i(t) = X_i(t) + \delta_i(t)$, $X(t) = 4cos(10t - 0.1) + 1.5sin(10t - 0.6) + 2cos(20t - 0.6)$,$f_0(t) = 0.75t - \exp(-6t)$ (to mimic glucose trend in our application below), $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and $\delta_i(t) \sim N(0, \sigma_\delta I)$. Four functions were considered for $f_1(t)$;1. Linear: $12t - 6$,, 2. Exponential: $1/(1+\exp(5-10t))$, 3. Polynomial: $1.5t-2t^2+1.6t^3-2t^5$ and 4. Sinusoidal: $0.2 - cos(t(1.5t - \pi)) + 1.2\exp(-8t^2)$.

Sparsity was induced to the data with (i) 10% missing completely at random - MCAR for all individuals and (ii) a 10% missing middle chunk of data for half of the population. Sample size was set to $n = 100$, with $n_i = 100$ sampling points and error variance $\epsilon = 1$. All the models were

fitted using equal basis splines for easier comparability since all the models are spline based. Models were fitted in R v3.6 to 100 datasets and comparisons were made using squared deviation, $\int (f_p(t) - \hat{f}_p(t))^2 dt,\ p = 1, 2, 3, 4$. The smaller the deviation, the better the model.

2.4 Application

In order to determine the relationship between extracellular glucose and heart functioning, $n = 17$ type 1 diabetes patients were recruited in a prospective observational study. They were fitted the continuous glucose monitor (CGM) which measured glucose in mmol/L every 5 minutes, and also wore a vest with sensors to measure ECG data every 10 minutes (in particular QT interval - see Figure 1). There were missing ECG data due to renewal of sensor gel or, for instance, to take a bath. The functional response variable of interest was the QTc (corrected QT) interval from the ECG vest and the functional predictor was glucose from the glucometer(mmol/L).

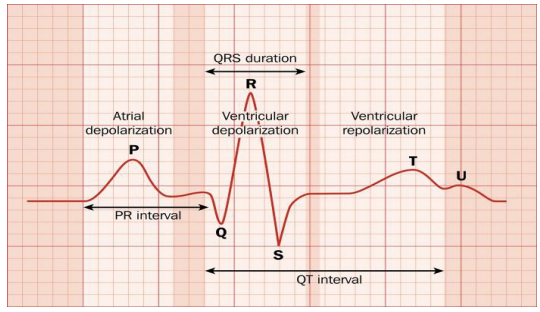


FIGURE 1. ECG tracing for one heart-beat

3 Results

3.1 Simulation study

The different models were fitted on 100 datasets for the two missingness scenarios. Boxplots for the 100 squared deviations for each function are shown in Figure 2. We can see that the Bayesian model provides stable estimates regardless of function and type of missingness, and gives the best approximation of the true function compared to other models. The time taken to run 100 observations model was 9.95, 44.56, 10.56 and 32.95 seconds using the *mgcv*, *fcr*, *vbvs.concurrent* R packages and Bayesian respectively. It can be seen that the Bayesian model is not as fast as the variational Bayes and *mgcv* but it's quicker than the *fcr* approach.

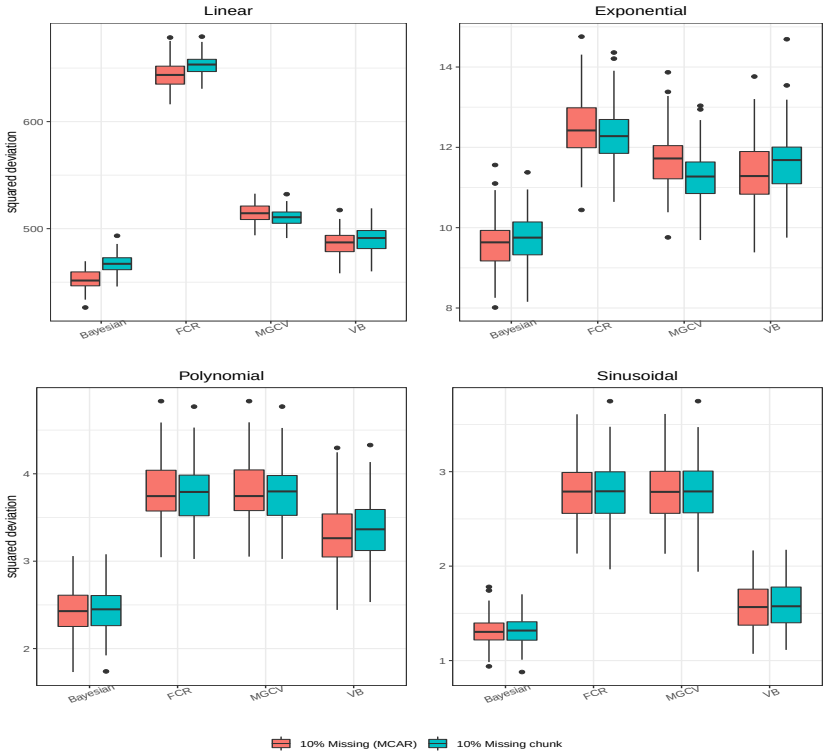


FIGURE 2. Boxplots for different missingness proportions for the four models

3.2 Application

The Bayesian FCM was fitted for the ECG-glucose data and the parameter function is shown in Figure 3. The plot provides evidence for a dynamic relationship over time. The fitted model finds a positive relationship between glucose and QTc at several times during the study. In these periods, as glucose increases, mean QTc also increases. This may be useful in the care of diabetes patients, although more work is required to investigate this complex dynamic relationship.

4 Conclusion

We have developed a novel Bayesian functional concurrent model which can be applied to sparse, irregular data. Our approach is competitive with other models regardless of the shape of the relationship between functional predictors and responses. It can deal with missing data as it uses all the available data and can impute missing observations along the way. In addition, it provides straightforward inferences through confidence bands. It

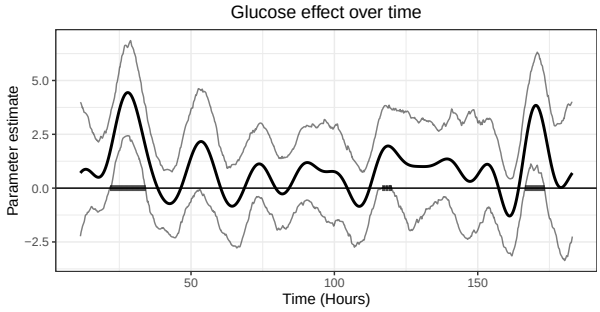


FIGURE 3. Parameter function for the effect of glucose on QTC

can be extended to include more covariates, both functional and scalar. However, there are a wide variety of functions for the relationship between functional variables which were not considered here. The Bayesian model takes more time to fit in comparison with some other models, but if better performance is valued, it should be used. In conclusion, we have seen that the novel Bayesian functional concurrent model outperforms other established models in the scenarios we have attempted.

Acknowledgments: This research was funded by Science Foundation Ireland and the Insight Centre for Data Analytics.

References

Crainiceanu, C. M., and Goldsmith, A. J. (2010). Bayesian Functional Data Analysis Using WinBUGS. *Journal of Statistical Software*, **32**.

Febrero-Bande, M., and de la Fuente, M.O. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software*, **51**, 1 – 28

Goldsmith, J., and Schwartz, J.E. (2017). Variable selection in the functional linear concurrent model. *Statistics in medicine*, **36**), 2237 – 2250.

Ivanescu, A.E., Staicu, A.M., Scheipl, F., and Greven, S. (2015). Penalized function-on-function regression. *Comput Stat* **30**, 539 – 568

Leroux, A., Xiao, L., Crainiceanu, C.M., and Checkley, W. (2018). Dynamic prediction in functional concurrent regression with an application to child growth. *Statistics in Medicine* **37**, 1376 – 1388.

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed., Springer,New York.

Invariance and the forecasting of mortality II: Standard errors

Iain Currie[1]

[1] Heriot-Watt University, UK

E-mail for correspondence: I.D.Currie@hw.ac.uk

Abstract: This is a companion paper to the paper we presented at IWSM 34 in 2019 on the modelling of human mortality. Many such models are not identifiable so parameter constraints are often used to obtain parameter estimates that are then used for forecasting. In the 2019 paper we considered the invariance of the central forecasts of the force of mortality with respect to the choice of constraints when an ARIMA model is used to forecast parameter estimates. In the present paper we consider the standard errors of these forecasts and show that these too are invariant when an ARIMA model is used. We illustrate our results with the same Portuguese data.

Keywords: Forecasting; Identifiability; Invariance; Mortality.

1 Introduction

The forecasting of human mortality is a central problem for the providers of pensions, annuities and other financial products which depend on the future duration of a human life. The usual approach is to build a model which depends on an individuals age, the current year and their date of birth. However, dependencies among these three determinants mean that most such models are not identifiable. As a consequence the forecasting of particular parameter estimates subject to an arbitrary set of constraints is problematic. In our earlier paper we showed that, while parameter estimates are not identifiable, forecast values are identifiable when an ARIMA model is used to forecast. In the present paper we extend these invariance results to the standard errors of the forecasts.

We use Portuguese mortality data downloaded from the Human Mortality Database on December 18, 2018. We have the number of deaths $d_{x,y}$ and the corresponding central exposed to risk $e_{x,y}$ for ages 50 to 90 and years

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

1970 to 2015. For simplicity we will index the ages by $\boldsymbol{x}_a = (1, ..., n_a)^{\mathrm{T}}$, the years by $\boldsymbol{x}_y = (1, ..., n_y)^{\mathrm{T}}$ and the years of birth by $\boldsymbol{x}_c = (1, ..., n_c)^{\mathrm{T}}$ where $n_c = n_a + n_y - 1$ is the number of distinct cohorts. We suppose that the number of deaths at age x in year y follows a Poisson distribution $P(e_{x,y}\lambda_{x,y})$ where $\lambda_{x,y}$ is the force of mortality at age x in year y.

2 Method

We consider a generalized linear model or GLM with model matrix $\boldsymbol{X}$, $n \times p$, $n > p$, rank $p - q$, $q \geq 1$, and vector of parameters $\boldsymbol{\theta}$. Since $\boldsymbol{X}$ is not of full rank $\boldsymbol{\theta}$ is not identifiable. However, there exists a matrix $\boldsymbol{H}$, $q \times p$, with rank q such that $\boldsymbol{H\theta} = \boldsymbol{0}$. Now, subject to the condition that $\boldsymbol{H\theta} = \boldsymbol{0}$, we do have a unique estimate of $\boldsymbol{\theta}$. We refer to $\boldsymbol{H}$ as a *constraints matrix* and we note that $\boldsymbol{H}$ is not unique.
Currie (2013) gave the following formula for the variance of the parameter estimates, $\hat{\boldsymbol{\theta}}$, in a constrained GLM with model matrix $\boldsymbol{X}$ and constraints matrix $\boldsymbol{H}$. We define $\boldsymbol{\Delta} = \boldsymbol{X}^{\mathrm{T}}\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$ then $\mathrm{Var}(\hat{\boldsymbol{\theta}})$ is given by

$$\boldsymbol{\Psi} = \boldsymbol{\Delta}^{-1} - \boldsymbol{\Delta}^{-1}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}^{\mathrm{T}})^{-1}\boldsymbol{H}\boldsymbol{\Delta}^{-1} \tag{1}$$

and the variance matrix of the fitted values $\boldsymbol{H}\hat{\boldsymbol{\theta}}$ follows as $\boldsymbol{X\Psi X}^{\mathrm{T}}$. We show that not only is $\boldsymbol{X\Psi X}^{\mathrm{T}}$ invariant with respect to the choice of $\boldsymbol{H}$ but so also are the standard errors of the central forecasts.

3 Example

We consider the age-period-cohort or APC model:

$$\log \lambda_{i,j} = \alpha_i + \kappa_j + \gamma_{n_a - i + j},\ i = 1, ..., n_a,\ j = 1, ..., n_y, \tag{2}$$

where i is the age at death, j is the year of death and $n_a - i + j$ is the year of birth. Let $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{n_a})^{\mathrm{T}}, \boldsymbol{\kappa} = (\kappa_1, ..., \kappa_{n_y})^{\mathrm{T}}, \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_{n_c})^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\kappa}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. We illustrate invariance with three constraint systems: a standard one found in the literature, a random one and one equivalent to Rs method of fitting a rank deficient regression model.
Currie (2019) showed that the parameter estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\kappa}}$ and $\hat{\boldsymbol{\gamma}}$ under the different constraint systems were strikingly different. Figure 1 shows that the same is true for their corresponding variances. However, $\boldsymbol{X\Psi X}^{\mathrm{T}}$, the variance matrix of the fitted values, is invariant with respect to the choice of constraint system. In particular, the invariant variances of the fitted values in the final year, 2015, are shown in the lower right panel. We denote these variances by $\boldsymbol{V}_A$ and we use $\boldsymbol{V}_A$ in the construction of the invariant standard errors of the forecasts of mortality.
To illustrate the forecasting of mortality we consider a ten year forecast. The upper left panel of Figure 2 is divided into three regions. Region A is

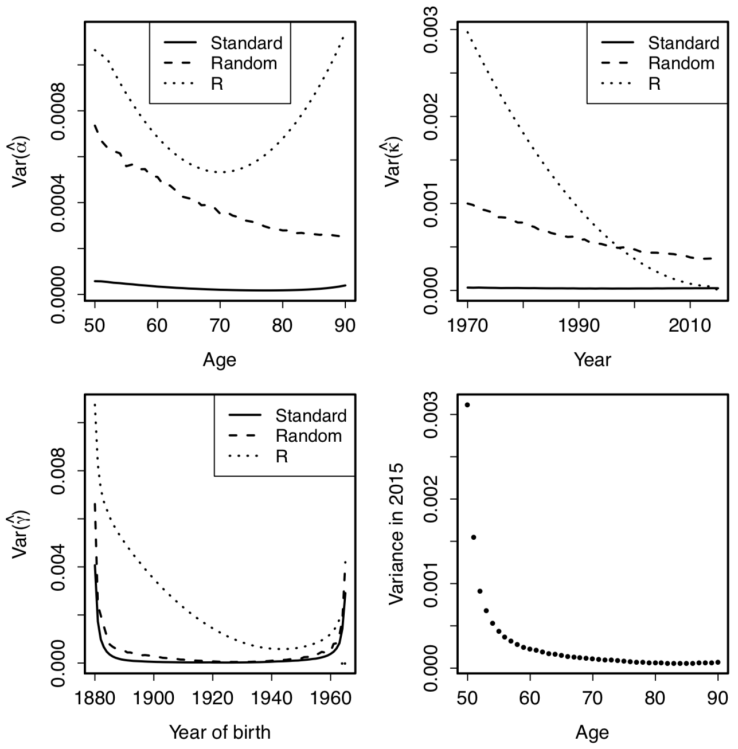


FIGURE 1. Variances of $\hat{\boldsymbol{\alpha}}$ (top left), $\hat{\boldsymbol{\kappa}}$ (top right) and $\hat{\boldsymbol{\gamma}}$ (bottom left) in the APC model under three sets of constraints; invariant variances $\boldsymbol{V}_A$ of fitted values in the final year, 2015, (bottom right).

the data region, in region B the forecast error depends on $\boldsymbol{V}_A$ and the forecast error for $\boldsymbol{\kappa}$, while in region C the forecast error depends additionally of the forecast error for $\boldsymbol{\gamma}$. The resulting invariant forecast standard errors are shown in Figure 2.

4 Conclusions

We showed (Currie, to appear) that two constraint systems lead to identical fitted and forecast values of mortality. In the present paper we show that this result extends to their standard errors.

Our results have important financial consequences. Forecasts of mortality for policyholders in their fifties are necessarily for thirty, forty or even fifty years ahead. Perversely, these ages are exactly those for which estimates and forecasts of mortality are their least reliable; lower right panel in Figure 2. Actuaries routinely drop the cohort parameters from a model where the

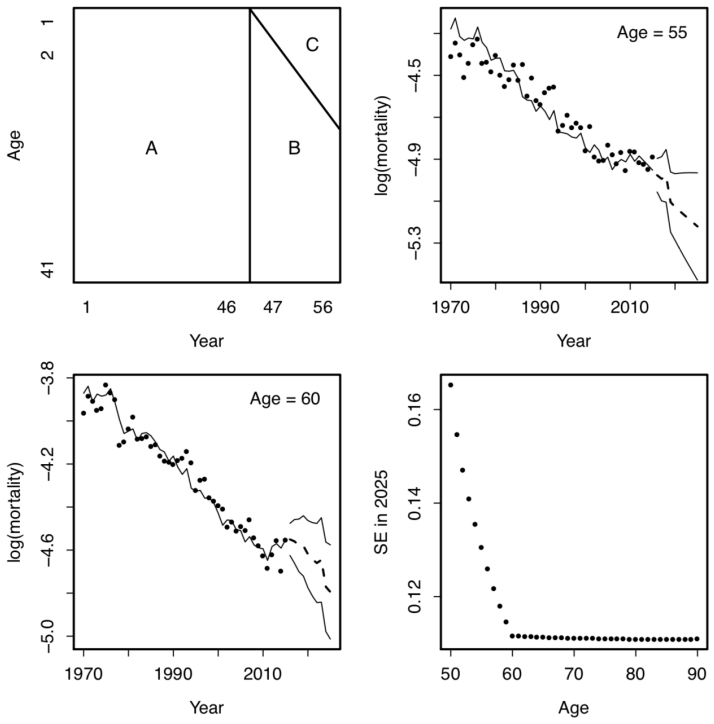


FIGURE 2. Top left: three regions; forecasts with 95% confidence intervals at age 55 (top right) and age 60 (bottom left); bottom right: invariant standard errors of forecast in final forecast year, 2025.

number of cells with such parameters is small, say fewer than four. The assumption here is that it is better to forecast such cohort parameters than to estimate them. We will use our results on standard errors to examine this assumption in future work.

References

Currie, I.D. (2013). Smoothing constrained generalized linear models with an application to the Lee-Carter model. *Statistical Modelling*, 13, 69 – 93.

Currie, I.D. (2019). Invariance and the forecasting of mortality. In: *Proceedings of the 34th International Workshop on Statistical Modelling*, Guimarães, Portugal, 95 – 100.

Currie, I.D. (to appear). Constraints, the identifiability problem and the forecasting of mortality. *Annals of Actuarial Science*.

Improved statistical emulation for a soft-tissue cardiac mechanical model

David Dalton[1], Dirk Husmeier[1]

[1] Department of Mathematics and Statistics, University of Glasgow, UK

E-mail for correspondence: d.dalton.1@research.gla.ac.uk

Abstract: This paper outlines a new approach to emulation based parameter inference in a cardiac mechanic model of the left ventricle (LV) of the heart that allows for prediction uncertainty to be accounted for. The emulation is performed using Gaussian processes, which are designed to build on the results of previous research in this area. This approach yields more accurate parameter estimates than previously reported in the literature.

Keywords: Cardiac Modelling; Holzapfel-Ogden Law; Statistical Emulation

1 Introduction

The Holzapfel-Ogden (HO) model (Holzapfel and Ogden, 2009) is a system of coupled partial differential equations that define the stretch-strain dependence of the inner tissue of the LV, known as the endocardium. The model depends on various material parameters, for example those which are related to the stiffness of the cardiac fibres. Interest is in determining these parameters for their potential to aid in the diagnosis of cardiac defects, however they can only be directly measured *in-vivo* by invasive procedures. An alternative, non-invasive approach to inferring the parameters which has potential for use in clinical decision support is to use magnetic resonance imaging (MRI). This is done by taking MRI scans of a subject's LV at end diastole, to determine the myocardium responses that are modelled by the HO law. The material properties can then be estimated as those values which minimise the discrepancy between the observed myocardium response, and the response predicted by the model. The diagnostic value of this approach has been shown in previous work (Gao *et al.* 2017). The problem, however, is that the HO cardio-mechanical equations describing

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

the kinematics of the heart have no closed form solutions. Instead, numerical procedures based on finite element discretisation are required, which typically take on the order of 15 minutes per evaluation on a high performance computer. Since solving for the material properties by iterative optimisation methods may require hundreds or thousands of such evaluations, the approach is rendered unsuitable as a real-time clinical decision support tool.
A number of methods exists which can be used to overcome this problem, one of which is *statistical emulation.*

2 Statistical Emulation

Statistical emulation involves approximating a computationally expensive model, referred to as a simulator, with a much cheaper surrogate model, known as an emulator. This is done by first choosing a set of points to cover the input region of interest, and then running the simulator $\mathbf{f}$ from each point. This creates a dataset of input-output pairs

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{f}(\mathbf{x}_i))_{i=1}^{N}\} \tag{1}$$

on which the surrogate model $\hat{\mathbf{f}}$ is trained. While the creation of a dataset in this manner for the HO law is extremely computationally expensive, all simulations can be done in advance of clinical deployment. In clinic, $\hat{\mathbf{f}}$ can be used in place of $\mathbf{f}$ in the parameter optimisation routine, allowing estimates to be obtained in real time. The inputs required for the HO model are a LV geometry, $\mathcal{H}$, and a four dimensional parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^{\mathsf{T}}$, which we are interested in inferring. In this study, we considered a fixed LV geometry, and then used a Sobol sequence (Fang *et al.*, 2006) to generate $10,100$ parameter configurations within the physiologically realistic boundaries $(0.1, 5)^4$. The simulator was then run from each point, and 25 outputs were extracted: circumferential strains at $K = 24$ regions along the endocardial surface, and the LV volume, all measured at end-diastole.
Having created the dataset, the approach for constructing the emulator $\hat{\mathbf{f}}$ must be considered. Gaussian process regression is a Bayesian non-parametric approach that is commonly used for emulation (Kennedy and O'Hagan, 2001). A Gaussian process (GP) is a stochastic process where any finite collection of random variables from the process are Gaussian distributed. GPs can be used for regression to define a prior directly over a space of functions

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{2}$$

where the GP is completely specified by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. Given a finite set of known training points and unknown test points, the GP marginalises to a multivariate Gaussian distribution, with mean and covariance found by evaluation of m and k at

the given points. Standard rules for conditional Gaussian distributions can then be used to find the posterior distribution of the test points, given the training points. For further details, the reader is directed to (Rasmussen and Williams, 2006).

One drawback of the GP modelling framework is that training and prediction times grow with the size of the dataset under consideration. Local Gaussian process regression (Gramacy and Apley, 2015) is an approach which can alleviate this complexity. With local GPs, a prediction is made at a given point using a GP trained on only the k nearest neighbours of the point in the training data. Initial work on the HO simulation data we analyse in this paper demonstrated the effectiveness of an emulator comprised of 25 independent local GPs, one for each dimension of the simulator output (Davies *et al.*, 2019). The problem with the local GP approach in our application context is that, as we adjust the input parameter values during the optimisation routine, the local neighbourhood will also change. This in turn means that the emulator will need to be refit at each iteration. Further work however has shown that a multivariate-output local GP emulator trained on the nearest neighbours of a test point in *output space* can accurately model the HO law (Noe *et al.*, 2019). The advantage of considering neighbours in output space is that this neighbourhood does not change during the parameter optimisation, meaning that the emulator only has to be fit once for each test point.

Given the above results, in this paper we consider an emulator made up of 25 independent local GPs trained on $k = 200$ neighbours in output space. The GPs were fit with linear mean functions and with squared exponential kernel function, where the length scales for each input dimension were allowed to vary. This is known as an ARD (Automatic Relevance Determination) prior in the Machine Learning community (Rasmussen and Williams, 2006). Although the data under consideration is deterministic, a small nugget term (10^{-6}) was added for reasons of numerical stability.

3 Maximum Likelihood Parameter Inference

Our objective is to find the optimal parameter estimates which minimise the loss between measured data, and the values predicted by the emulator. In what follows, we denote the measured quantities, after non-dimensionalisation, by

$$\mathbf{y} = (y_0, y_1, \ldots, y_K)^\mathsf{T} \tag{3}$$

where y_0 is the non-dimensionalised LV volume, and $y_1, \ldots, y_{24}$ are the non-dimensionalised circumferential strains.

The corresponding outputs from the GP emulator, which depend on the cardio-mechanic parameters $\boldsymbol{\theta}$ and the LV geometry, $\mathcal{H}$, are denoted:

$$\hat{\mathbf{f}}(\boldsymbol{\theta}, \mathcal{H}) = \left(\hat{\mathrm{f}}_0(\boldsymbol{\theta}, \mathcal{H}), \hat{\mathrm{f}}_1(\boldsymbol{\theta}, \mathcal{H}), \ldots, \hat{\mathrm{f}}_K(\boldsymbol{\theta}, \mathcal{H})\right)^\mathsf{T} \tag{4}$$

In previous work, described in Section 2, the cardio-mechanic parameters $\boldsymbol{\theta}$ for a given LV geometry $\mathcal{H}$ were estimated by minimizing the L2 norm of the difference between $\mathbf{y}$ and $\hat{\mathbf{f}}$:

$$E(\boldsymbol{\theta}, \mathcal{H}) \;=\; ||\mathbf{y} - \hat{\mathbf{f}}(\boldsymbol{\theta}, \mathcal{H})||^2 \;=\; \sum_{i=0}^{K} \Big(y_i - \hat{\mathrm{f}}_i(\boldsymbol{\theta}, \mathcal{H}) \Big)^2 \tag{5}$$

where each output $\hat{\mathrm{f}}_i(\boldsymbol{\theta}, \mathcal{H})$ was set to the corresponding posterior GP mean $\mu_i(\boldsymbol{\theta}, \mathcal{H})$. Under the assumption that the measurement noise is iid additive Gaussian with variance σ_m^2:

$$\mathbf{y} \;=\; \hat{\mathbf{f}}(\boldsymbol{\theta}, \mathcal{H}) + \boldsymbol{\varepsilon}; \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_m^2 \mathbf{I}) \tag{6}$$

we obtain the log likelihood, conditional on the emulator output $\hat{\mathbf{f}}(\boldsymbol{\theta}, \mathcal{H})$:

$$\log p\Big(\mathbf{y} | \hat{\mathbf{f}}(\boldsymbol{\theta}, \mathcal{H})\Big) \;=\; \frac{-1}{2\sigma_m^2} \sum_{i=0}^{K} \Big(y_i - \hat{\mathrm{f}}_i(\boldsymbol{\theta}, \mathcal{H}) \Big)^2 - \frac{K+1}{2} \log(2\pi\sigma_m^2) \tag{7}$$

Maximizing this conditional likelihood with respect to $\boldsymbol{\theta}$, for a given LV geometry $\mathcal{H}$, is equivalent to minimizing the original objective function (5), where again the emulator outputs are set to the posterior GP mean values. However, a disadvantage of this approach is that the uncertainty of the emulator, naturally predicted by the GP variance, is not taken into consideration. To rectify this, we can compute the marginal likelihood by integrating over the emulator outputs

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{H}, \sigma_m^2) = \int p(\mathbf{y}|\hat{\mathbf{f}}, \sigma_m^2) p(\hat{\mathbf{f}}|\boldsymbol{\theta}, \mathcal{H}) d\hat{\mathbf{f}} = \prod_{i=0}^{K} \int p(y_i|\hat{\mathrm{f}}_i, \sigma_m^2) p(\hat{\mathrm{f}}_i|\boldsymbol{\theta}, \mathcal{H}) d\hat{\mathrm{f}}_i \tag{8}$$

where conditional independence between the outputs has been assumed. The two probability distributions under the integral are given by

$$\begin{aligned} p(y_i|\hat{\mathrm{f}}_i, \sigma_m^2) &= \mathcal{N}(y_i|\hat{\mathrm{f}}_i, \sigma_m^2) \\ p(\hat{\mathrm{f}}_i|\boldsymbol{\theta}, \mathcal{H}) &= \mathcal{N}\Big(\hat{\mathrm{f}}_i|\mu_i(\boldsymbol{\theta}, \mathcal{H}), \sigma_i^2(\boldsymbol{\theta}, \mathcal{H})\Big) \end{aligned} \tag{9}$$

where $\mu_i(\boldsymbol{\theta}, \mathcal{H})$ is the mean of the i^{th} GP emulator, and $\sigma_i^2(\boldsymbol{\theta}, \mathcal{H})$ is its variance. The integral in (8) is a therefore a standard Gaussian integral with closed-form solution

$$p\left(\mathbf{y}|\boldsymbol{\theta}, \mathcal{H}, \sigma_m^2\right) = \prod_{i=0}^{K} \mathcal{N}\left(y_i|\mu_i(\boldsymbol{\theta}, \mathcal{H}), \sigma_m^2 + \sigma_i^2(\boldsymbol{\theta}, \mathcal{H})\right) \tag{10}$$

which gives

$$\log p(\mathbf{y}|\boldsymbol{\theta}, \mathcal{H}, \sigma_m^2) = -\frac{1}{2} \sum_{i=0}^{K} \left\{ \frac{\Big(y_i - \mu_i(\boldsymbol{\theta}, \mathcal{H})\Big)^2}{[\sigma_m^2 + \sigma_i^2(\boldsymbol{\theta}, \mathcal{H})]} + \log\Big(2\pi[\sigma_m^2 + \sigma_i^2(\boldsymbol{\theta}, \mathcal{H})]\Big) \right\} \tag{11}$$

as a better objective function to optimise.

4 Results and Discussion

In the absence of large quantities of real patient data, we instead reserved the final 100 points of our simulated dataset as an independent test set on which to quantify the difference in parameter estimation accuracy when emulation uncertainty is accounted for. Using local GP emulators trained on the remaining simulation data, we used iterative optimisation methods to estimate each of the independent test parameter values with loss functions (5) and (11) respectively. By then evaluating the mean squared error (MSE) between the known true values and the estimated values, we obtain a list of 100 errors for each loss function. The median of these lists is displayed in Table 1, alongside the first and third quartiles.

TABLE 1. Test Set MSE (Parameter Space)

Emulator	Loss Function	1^{st} Quartile	Median	3^{rd} Quartile
Local GP	Equation (5)	9.2×10^{-8}	3.1×10^{-7}	3.1×10^{-6}
Local GP	Equation (11)	7.1×10^{-8}	3.0×10^{-7}	2.0×10^{-6}

The results in Table 1 quantify the improvement in parameter estimation accuracy that can be achieved by accounting for emulation uncertainty. The gain in performance is slight, which may be due to the prediction variance for each output dimension being quite similar. Of note is that our parameter estimation accuracy has improved by more than one order of magnitude over the best results from the literature, particularly as a consequence of a decreased nugget term. This accuracy is visualised in Figure 1, which plots the 100 out of sample test parameter values, broken down into each dimension respectively, versus the corresponding values predicted when using loss function (11). We see extremely good agreement between the true and predicted values across the entire parameter space.

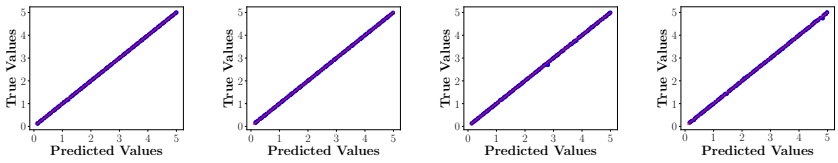


FIGURE 1. Plot of the true test set parameter values versus those predicted by the local GP emulator using loss function (11), when broken down into each dimension respectively. Points lying on the red lines of unit slope indicate perfect prediction accuracy.

The limitation of the analysis presented here is that we have performed emulation for a fixed, known LV geometry $\mathcal{H}$. To be of clinical use however,

the emulator must be able to account for the unique LV geometry of a given patient. Therefore, the construction of emulators that can account for LV geometry variations will be the remit of further work.

Acknowledgments: This work was carried out as part of the SoftMech^{MP} project, funded by EPSRC, grant reference number EP/S030875/1. Dirk Husmeier was funded by the Royal Society of Edinburgh, grant reference number 62335.

References

Davies, V., Noe, U., Lazarus, A., Gao, H., Macdonald, B., Berry, C., Luo, X. and Husmeier, D. (2019). Fast parameter inference in a biomechanical model of the left ventricle by using statistical emulation *Journ. of the R. Stat. Soc., Series C (Appl. Statistics)*, **68**, Part 5, 15551576

Fang K., Li R., Sudjianto A. (2006). *Design and modeling for computer experiments*. London, UK: Chapman & Hall/CRC

Gao, H., Li, W., Cai, L., Berry, C. and Luo, X. (2015). Parameter estimation in a HolzapfelOgden law for healthy myocardium. *J. Engng Math.*, **95**, 231248.3)

Gramacy R.B., Apley D.W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Stat.*, **24**, 561 578.

Holzapfel, G. A. and Ogden, R. W. (2009). Constitutive modelling of passive myocardium: a structurally based framework for material characterization *Phil. Trans. R. Soc. A,*, **367**, 34453475.

Kennedy M.C., OHagan A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. B (Statistical Methodology)*, **63**, 425 464. (doi:10.1111/rssb. 2001.63.issue-3)

Noe, U., Lazarus, A., Gao, H., Davies, V., Macdonald, B., Mangion, K., Berry, C., Luo, X. and Husmeier, D. (2019). Gaussian process emulation to accelerate parameter estimation in a mechanical model of the left ventricle: a critical step towards clinical end-user relevance. *J. R. Soc. Interface*, **16**, 20190114. http://dx.doi.org/10.1098/rsif.2019.0114

Rasmussen C.E., Williams C.K.I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Log-ratio diagrams for compositions with zero counts

Paul H. C. Eilers[1]

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

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

Abstract: The ternary diagram is a popular tool for displaying compositions with three components. When one or two components are close to zero, it is hard to judge the display. The TrioScale diagram is an effective alternative, using log-ratios. Unfortunately, it cannot handle zeros, which is a serious drawback when studying count data. A variant of PRIDE is proposed to adjust the counts, guaranteeing positive numbers.

Keywords: Compositional data, PRIDE, shrinkage, penalties.

1 Introduction

The ternary diagram is a popular and effective way to display triples of fractions. Figure 1 shows an example, for concentrations of three metals (Manganese, Rubidium and Strontium) in moss (data set *moss* in the *R* package *StatDA*). The underlying principle is that for any point in an isosceles triangle the sum of the lengths of the perpendiculars on the sides is constant.

If a part of the fractions is close to 0 or 1, the dots gets pushed to the sides, or into a corner, making the ternary plot hard to interpret, as can be seen in Figure 1. An alternative diagram, called TrioScale (de Rooij and Eilers, 2013), does not have this problem, as is shown in the right panel. The axes now have a different meaning: they represent logs of ratios, like log(Sr/Rb) for the horizontal axis. The coordinates in the new diagram are easy to compute: $x = \log(p_2/p_1)$ and $y = [2\log(p_3/p_1) - x]/\sqrt{3}$. Notice that they are based on log-ratios and that the sum of the proportions can be arbitrary.

Unlike the ternary plot, the TrioScale diagram has no bounds. The axes can be moved to parallel positions to create room for the data points. The log-

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

ratios for any point in the diagram can be read by perpendicular projection on the axes.

Logs of ratios are preferable for statistical modelling of compositional data (van den Boogaart and Tolosana-Delgado, 2013). de Rooij and Eilers (2013) showed that, on the transformed scales, linear relationships between log-ratios form straight lines. In contrast, in a ternary diagram they show strong curvatures.

The point cloud in the TrioScale plot in Figure 2 suggests a bivariate normal distribution as a decent approximation. This is certainly not the case in the ternary diagram. Also the domain is bounded in the latter, which is not the case in the TrioScale plot.

However, count data can create problems. With zeros in one or more fractions it is not possible to compute log-ratios. This is a nuisance when analyzing small counts. A simple, but inelegant, solution is to add a small number (like 0.5) to each count. Instead, I propose to use the PRIDE model (Perperoglou and Eilers, 2010) to replace zeros by positive numbers, as will be explained in the next section.

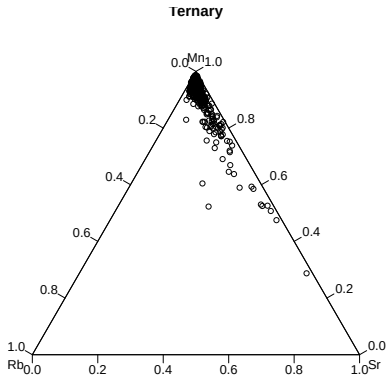


FIGURE 1. Concentrations of three metals (Manganese, Rubidium and Strontium) in moss, displayed in a ternary diagram. The axes indicate fractions.

2 PRIDE

PRIDE stands for Penalized Random Individual Deviance Effects. Let y be an $n-$vector of observed counts and let X be an n by m design matrix. In a log-linear model for for the expected values, μ, a random effect is introduced for each individual observation:

$$\log(\mu_i) = \sum_j x_{ij}\beta_j + \gamma_i \quad \text{or} \quad \log \mu = [X\, I_n] \begin{bmatrix} \beta \\ \gamma \end{bmatrix} = [X\, I_n]\theta = B\theta,$$

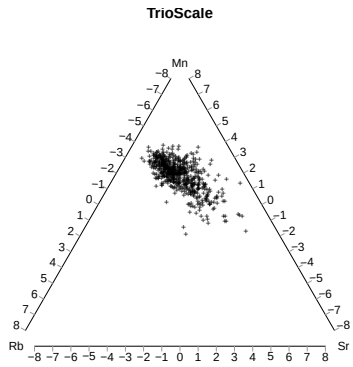


FIGURE 2. Concentrations of three metals (Manganese, Rubidium and Strontium) in moss, displayed in a TrioScale diagram. The axes indicate log-ratios of fractions.

where β contains the regression coefficients and γ the individual random effects. To estimate β and γ, the penalized deviance

$$D = 2\sum_i y_i \log(y_i/\mu_i) + \lambda \sum_i \gamma_i^2 + \kappa \sum_j \beta_j^2,$$

is minimized. The second penalty is introduced to avoid problems when X is collinear; κ is a small number (like 10^{-6}). The penalized likelihood equations are $B'(y-\mu) = P\theta$, where P is a diagonal matrix with blocks κI_m and λI_n on the diagonal. The resulting linearized equations, that have to be solved iteratively, are

$$(B'\tilde{M}B + P)\theta = B'(y - \tilde{\mu} + \tilde{M}B\tilde{\theta}),$$

where a tilde indicates the current approximation and $M = \mathrm{diag}(\mu)$. To determine the value of the penalty parameter λ, I use AIC. It combines the (unpenalized) deviance and the effective model dimension, which is computed (at convergence) as trace(G), with

$$G = (B'\hat{M}B + P)^{-1}B'\hat{M}B$$

PRIDE was designed to improve estimation of the regression parameters (and especially their standard errors) when over-dispersion makes a model with only β and the Poisson assumption unrealistic. My goal here is different: after estimating the model I replace the observed y, containing zeros, with $\hat{\mu}$. All elements of $\hat{\mu}$ are positive and thus suitable for presentation with TrioScale.

The recipe is as follows. Let the data be given in a matrix $Y = [y_{ij}]$ with m columns and n rows. Fit the model $\log \mu_{ij} = \eta_{ij} = \alpha_i + \beta_j + \gamma_{ij}$, so that

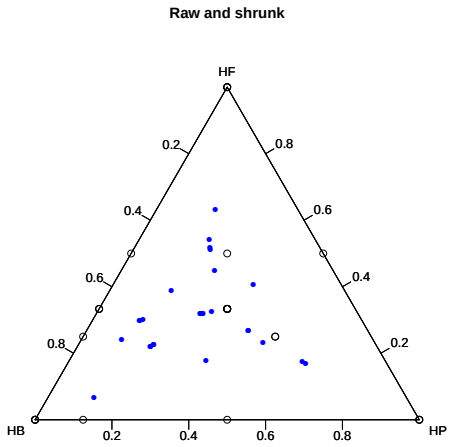


FIGURE 3. Ternary diagram of the proportions of time spent by pigs in three locations in a stable (HB: half in straw bed; HF: half in feeder; HP: half in dunging passage). The circles represent the raw proportions, the dots their adjusted values.

$\theta' = [\alpha' \beta' \gamma']'$. The minimum of AIC is searched for on a grid for $\log \lambda$. To fit the model, Y is converted to the vector y, and the design matrix $X = [X_1|X_2|I_p]$ is constructed, where $X_1 = e_m \otimes I_n$, $X_2 = I_m \otimes e_n$, e_q is a vector of ones of length q, I_q is and identity matrix of size q, and $p = mn$. With many observations, the number of equations, $n + m + mn$, gets large. The system is very sparse, so it can be solved quickly with sparse matrix software. That does not help much with the computation of (the trace of) G. However, Perperoglou and Eilers (2010) presented a very efficient algorithm that exploits structure of the equations in (2). They consists of blocks, with the largest block diagonal with dimensions mn by mn. There is no need to store it explicitly as a matrix: a vector of the diagonal elements is enough. By rearranging the equations, computation time becomes essentially proportional to mn. Hence there is no practical limit to the size of the data set that can be handled.

3 An application

The R package *zCompositions* contains the data set *Pigs*. At 97 instants, 5 minutes apart, the locations of 29 sows in a stable were observed. Six locations were considered: straw bed (BED), half in the straw bed (HB), dunging passage (PASSAGE), half in the dunging passage (HP), feeder (FEEDER) and half in the feeder (HF). The triples (HB, HP, HF), are interesting, because they show low counts and many zeros. I first apply PRIDE to the complete matrix with all six locations and then select the

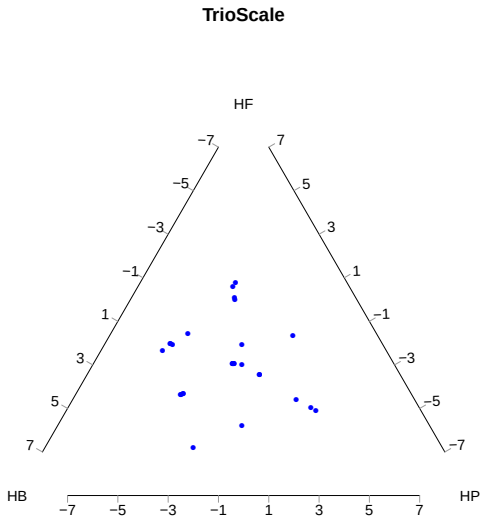


FIGURE 4. TrioScale diagram of the adjusted proportions of time spent by pigs in three locations in a stable (HB: half in straw bed; HF: half in feeder; HP: half in dunging passage).

triple (HB, HP, HF) for display. The AIC profile is shown in Figure 5. It shows a clear minimum around $\lambda = 0.8$.

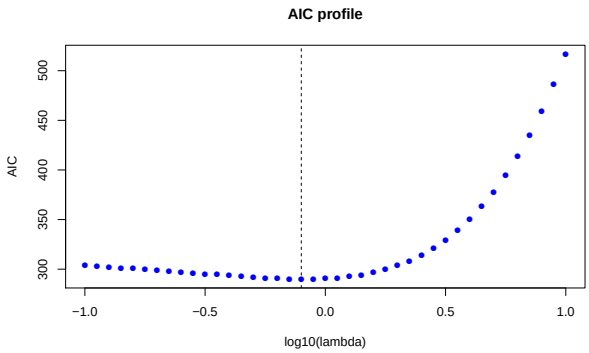


FIGURE 5. AIC profile for the Pigs data.

4 Discussion

The proposed model will always generate positive replacements for zeros, removing the main obstacle for using TrioScale with counts. It also changes the values of the non-zero observations. This is not common in recipes for compositional data, which usually correct only the zeros. The Bayesian algorithm in the package *zCompositions* (Martin-Fernadez et al., 2015) is an example. I think that it is reasonable to adjust all observations. Why modify only the zeros and keep the other numbers untouched?
An interpretation of the model is that we shrink towards the independence model $\log \mu_{ij} = \alpha_i + \beta_j$. In principle shrinking towards more complicated models is possible, if we have enough prior information. Covariates can also be included, when relevant.
Adjusting counts with PRIDE can be used in more places. It will work on any table with zeroes, as long as all row sums and all column sums are not zero. There is a rich literature on correction of zeros in outcomes of clinical trials with small observations. I proposed a bivariate display of log-odds (Eilers, 2007) that cannot handle exact zeros. It will be interesting to investigate the prior use of PRIDE there.

References

de Rooij, M. and Eilers, P.H.C. (2013). TrioScale: A new diagram for compositional data. in *Proceedings of CoDaWork 2013.*

Eilers, P.H.C. (2007). Data exploration in meta-analysis with smooth latent distributions. *Statistics in Medicine*, **26**, 3358 – 3368.

Martin-Fernandez J.A. et al. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, **15**, 134 – 158.

Perperoglou, A. and Eilers, P.H.C. (2010) Penalized regression with individual deviance effects. *Computational Statistics*, **25**, 341 – 361

van den Boogaart, K.G. and Tolosana-Delgado, R. (2013) *Analyzing Compositional Data with R.* Springer.

Modelling proposals in competing risk studies: empirical likelihood approaches to compare different risks

Hammou El Barmi[1], Vicente Núñez-Antón[2]

[1] The City University of New York, USA
[2] University of the Basque Country UPV/EHU, Spain

E-mail for correspondence: Hammou.Elbarmi@baruch.cuny.edu

Abstract: In standard competing risks studies, every unit or subject is exposed to different risks at the same time, but its actual failure or death is attributed to exactly one of them which is then called the cause of failure. In general, the goal of these studies is to distinguish between the following three alternatives: (1) the risks are equal, (2) the risks are not equal, and (3) the risks are linearly ordered. We concentrate on modelling proposals in competing risk studies and develop empirical likelihood (EL) based tests for testing the hypothesis that the cumulative incidence functions (CIF) corresponding to k-competing risks are equal against the alternative that they are not equal or that they are linearly ordered. The proposed test statistics are functionals of localized empirical likelihood statistics. Their asymptotic null distributions are distribution-free and have a simple representation in terms of a standard Brownian motion or a standard Brownian bridge. The tests we propose here are extended to the case of right-censored survival data via multiple imputation. In order to assess the usefulness of the proposed tests, and to illustrate the theoretical results for their asymptotic distributions, we include a simulation study and also discuss an example involving survival times of mice exposed to radiation.

Keywords: Competing risks; Cumulative incidence functions; Empirical likelihood, Hypotheses testing.

1 Introduction

In standard competing risks studies, every unit or subject is exposed to different risks at the same time, but its actual failure or death is attributed to exactly one of them which is then called the cause of failure. In general,

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

the goal of these studies is to distinguish between the following three alternatives: (1) the risks are equal, (2) the risks are not equal, and (3) the risks are linearly ordered. This is established on the basis of the observed data which is a random sample from (T, δ), where T is the lifetime and δ is the cause of death. As an example, when comparing brands of a component from different suppliers, the components may be tested in series. In such a setting, the components are functioning in the same environment and their times to failure are generally dependent. When this is the case, the system will fail as soon as one of the components fails. Consequently, we only observe the lifetime of the system and its cause of failure. The procedures we develop here, within the context of modelling proposals in competing risk studies will allow us to test whether these components are of the same quality against the alternatives that they are either: (a) of different quality, or (b) one is superior to the others. A key role in such comparisons is played by the cumulative incidence function (CIF). We assume that we have k risks in which case the possible values of δ are $1, \ldots, k$ The CIF corresponding to the j-th risk is a subdistribution function whose value at time t is the probability of failure or death by time t from risk j:

$$F_j(t) = P[T \leq t,\ \delta = j], \quad j = 1, \ldots, k, \tag{1}$$

with $F(t) = \sum_j F_j(t)$ being the distribution function (DF) of T, which we assume throughout to be continuous with survival function S. The cause specific hazard rate due to cause j is defined by

$$\lambda_j = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t \leq T \leq t + \Delta t, \delta = j | T \geq t], \quad j = 1, \ldots, k,$$

with the overall hazard rate $\lambda = \sum_j \lambda_j$. Comparison of competing risks based on their CIFs has been considered before in the literature. For the case of $k = 2$, several tests are available in the literature. In the continuous case, Aly et al. (1994) and El Barmi et al. (2004) used very closely related Kolmogorov-Smirnov type statistics to test $H_0 : F_1 = F_2$ against $H_1 : F_1 \leq F_2$, whereas El Barmi and Kochar (2002) developed a likelihood ratio test for the same problem for the discrete or grouped data situation. Extensions of these tests to the k-sample case have been considered in El Barmi and Mukerjee (2006) and El Barmi et al. (2006), for the general and the discrete/grouped data case, respectively. In this paper we provide alternative tests based on the empirical likelihood approach. The test developed in El Barmi and Mukerjee (2006) is the only test designed for the k-sample case in the general case. Recently El Barmi and El Bermi (2015) developed an EL approach to test H_0 against H_1. Clearly when H_0 is true, the time and the cause of failure are independent. On the other hand, the hypothesis of ordered CIFs, H_1, is equivalent to

$$H_1 : P[\delta = 1 | T \leq t] \leq P[\delta = 2 | T \leq t], \quad \forall t \geq 0$$

We extend the results in El Barmi and El Bermi (2015) to the k-risks case by developing an EL based tests in the uncensored case for testing H_0 against $H_2 - H_0$ and H_0 versus $H_1 - H_0$, where

$$H_0 : F_1 = F_2 = \cdots = F_k, \tag{2}$$

$$H_1 : F_1 \leq F_2 \leq \cdots \leq F_k, \tag{3}$$

and H_2 imposes no constraints on $F_j, j = 1, \ldots, k$. As already stated, when H_0 is true, the time and the cause of failure are independent. Moreover, the hypothesis os ordered CIFs, H_1 is equivalent to

$$H_1 : P[\delta = j | T \leq t] \leq P[\delta = j + 1 | T \leq t], \quad \forall t \geq 0, \, j = 1, 2, \ldots, k - 1.$$

In this form, H_1 states that, given that a unit has failed by time t, the conditional probability of its failing from cause $j + 1$ is uniformly greater than that from cause j.

Our objective is to develop a novel EL approach to the important problem of nonparametrically testing H_0 against H_2 and H_0 against H_1 based on a competing risks data modelling approach among the k CIFs. The proposed tests is computationally efficient to implement and could be used with massive data sets because they do not rely on the bootstrap or any other simulation technique, and they reduce to a local test for an ordering of binomial probabilities, which only requires a single sweep through the pooled data. The proposed test statistics are functionals of localized empirical likelihood statistics and their asymptotic null distributions are distribution-free and have a simple representation. In order to implement the test, we need to obtain he critical values of the corresponding test statistic, where its finite sample as well as its asymptotic null distributions are not tractable but the latter is distribution free. The approximate critical values can be obtained by simulating 10000 data sets. The i-th dataset, $\{(T_{ij}, \delta_{ij}), j = 1, 2, \ldots, 100\}$ is a sample of size 100, where $T_{ij} = X_{1ij} \wedge X_{2ij} \wedge \ldots \wedge X_{kij}$, and $\delta_{ij} = \ell$, if $T_{ij} = X_{\ell ij}$. Here $X_{1ij}, X_{2ij}, \ldots, X_{kij}$ are independent exponential random variables with mean one. The R program used to compute these approximate critical values is available from the authors upon request.

The proposed tests are also extended to the case of right-censored survival data via multiple imputation. Consider first the situation of Type I censoring in which we assume that all the units enter at baseline and are followed for a set time-period, say $[0, \tau]$. At the end of the follow-up period, the remaining subjects at risk are right-censored. The right-censored subjects can be viewed as failing in some (unknown) random order after the end of follow-up period. In addition, only the order in which they fail and the cause of failure affect the complete-data test statistics $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$, which would be available if all times and the causes of failure were observed. Our proposal is to simply average $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$ over all possible permutations of these unobserved failure times. An average based on Monte Carlo sampling

could be used to reduce the computational cost when the censoring rate is large. The null distribution is unchanged. A similar idea can be used to handle the case of random right-censoring as follows. First, let us note that the Kaplan-Meier estimator, $\hat{S}$ (Kaplan and Meier, 1958), can be plugged-in to provide an estimate of the residual survival function $e(s) = S(s)/S(t)$, for $s > t$, provided that $S(t) > 0$. If we specify $\tau > 0$ such that $\hat{S}(\tau) > 0$, we have that, for any right-censored observation at $t < \tau$, the estimated residual survival function is well-defined. Simulating from this estimated residual survival distribution produces a new "uncensored" observation if it falls in $[t; \tau)$, and, otherwise, a Type I censored observation. Either way, its probability of failing from risk j is $1/k$. Any observation (censored or non-censored) at $s > \tau$ becomes right-censored at τ. In this way, we reduce the problem to the Type I censored case discussed above. Clearly, it would be important to set the value of τ as large as possible to be able to minimize the amount of extraneous right-censoring at τ. In practice, this could be achieved by setting it slightly to the left of the largest uncensored observation. In the sequel, when using this proposed procedure, we average the complete-data test statistics $\mathcal{S}_{01}$ and $\mathcal{S}_{02}$ over 1000 "simulated complete" data samples. The theoretical justification for motivating the proposed imputation procedure can be derived by using a result of Akritas (1986, Theorem 2.2) on bootstrapping the Kaplan-Meier estimator.

2 Application and discussion

To illustrate the theoretical results, we discuss an example involving survival times of mice exposed to radiation. We analyze a set of mortality data kindly provided by Dr. H.E. Walburg, Jr. of the Oak Ridge National Laboratory and reported by Hoel (1972). The data were obtained from a laboratory experiment and consisted of the survival times of 82 male mice who were exposed to radiation at an age between 5 and 6 weeks, and that were kept in a germ-free environment. After autopsy, the cause of death was attributed to one of three causes: reticulum cell sarcoma (blue: 3), other causes (red: 2) and thymic lymphoma (black: 1) (see Figure 1). Our proposed test results in a value providing an estimated p-value between 0.01 and 0.02, leading to the rejection of the hypothesis that the CIFs are equal, which implies that risks are not the same (see Figure 1). Ordered alternative of interest based on medical recommendations and specific hypothesis of interest were also considered and tested for. Our proposed test is an asymptotically distribution-free empirical likelihood ratio type tests for testing the null hypotheses that k cumulative incidence functions corresponding to competing risks are equal against the alternative that they are not equal, and against the alternative that they are linearly ordered. We also provide approximate critical values for these tests and studied its behavior in both a simulation study and a real dataset application. In ad-

dition, we also discuss a new approach that can be used to extend our test to the right censored data situations.

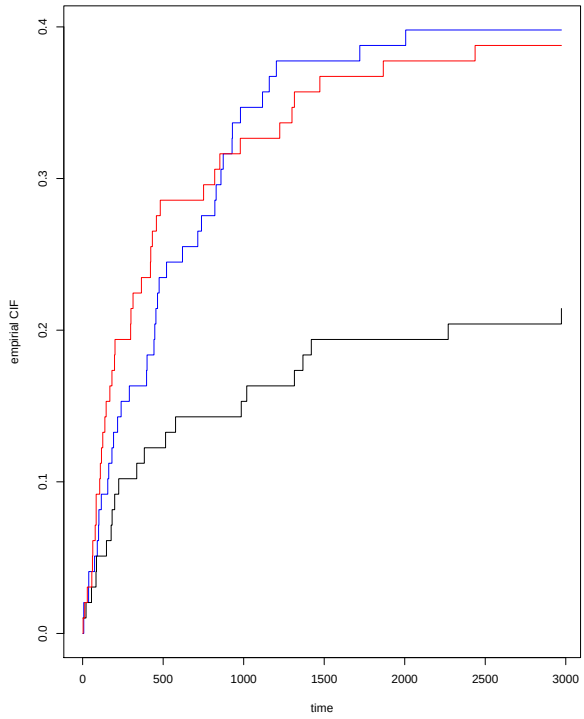


FIGURE 1. CIFs for radiation mice data (Hoel, 1972).

3 Concluding remarks

In this paper, we have developed an asymptotically distribution-free empirical likelihood ratio type tests for testing the null hypotheses that k cumulative incidence functions corresponding to competing risks are equal against the alternative that they are not equal and against the alternative that they are linearly ordered. We have also provided approximate critical values for these tests and, in order to illustrate our results, we have analyzed a dataset that has been previously analyzed within these settings. In addition, we have also discussed a new approach that can use to extend our test to the right censored data situations.

Acknowledgments: This work was supported by Ministerio de Economía y Competitividad, Agencia Estatal de Investigación (AEI), Fondo Europeo de Desarrollo Regional (FEDER), the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group), and Universidad del País Vasco UPV/EHU under research grants MTM2016-74931-P (AEI/FEDER, UE), IT1359-19 and UFI11/03.

References

El Barmi, H., and El Bermi, L. (2015). On comparing cumulative incidence functions using an empirical likelihood ratio type test. *Communications in Statistics: Theory and Methods*,**44**, 4940 – 4952.

El Barmi, H., and Kockar, S. (2002). Inference for survival functions under order restrictions. *Journal of the Indian Statistical Association*, **40(2)**, 85 – 102.

El Barmi, H., and McKeague, I.W. (2016). Testing for uniform stochastic ordering via empirical likelihood. *Annals of the Institute of Statistical Mathematics*, **68**, 955 – 976.

El Barmi, H., Kockar, S., and Tsimikas, J. (2006). Likelihood ratio test for and against ordering of cumulative incidence functions in multiple competing risks and discrete mark variable models. *Journal of Statistical Planning and Inference*, **136**, 1588 – 1607.

El Barmi, H., Nuñez-Antón, V., and Zimmerman, D.L. (2009). Testing for and against a set of inequality constraints: the k-sample case. *Journal of Statistical Planning and Inference*, **139(3)**, 1012 – 1022.

El Barmi, H., Kochar, A., Mukerjee, H., and Samaniego, F. (2004). Inference for subsurvival functions under an Order Restriction. *Journal of Statistical Planning and Inference*, **118**, 145 – 165.

Hoel, D.G. (1972). A representation of mortality data by competing risks. *Biometrics*, **28**, 475 – 488.

Robertson, T., Wright, F.T., and Dykstra, R.L. (1988). *Order Restricted Inference*. New York: Wiley.

A new model for multivariate functional data classification with application to the prediction of difficulty in web surveys using mouse movement trajectories

Amanda Fernández-Fontelo[1], Felix Henninger[2], Pascal J. Kieslich[2], Frauke Kreuter[2,3,4], Sonja Greven[1]

[1] Chair of Statistics, School of Business and Economics, Humboldt-Universitt zu Berlin, Berlin, Germany
[2] Mannheim Centre for European Social Research, University of Mannheim, Mannheim, Germany
[3] University of Maryland, College Park, Maryland, USA
[4] Institute for Employment Research, Mannheim, Germany

E-mail for correspondence: fernanda@hu-berlin.de

Abstract: A semi-metric-based model for multivariate functional data classification is presented and used to improve difficulty prediction in web surveys with mouse movement trajectories.

Keywords: ensemble; nearest neighbors; non-parametric functional kernel; online surveys; semi-metrics

1 Introduction

Functional data is a relatively new branch of statistics devoted to the study of curves, surfaces, images, etc., which has experienced rapid development in recent years. Improvements are continually emerging to answer real-world functional data questions from many scientific disciplines.
This paper aims to improve difficulty prediction through mouse movement trajectories gathered from respondents in a web survey, where several questions were manipulated to create different scenarios of difficulty. In survey research, difficulty in understanding and responding to survey questions in the way researchers intended is one of the most frequent sources of error that impedes the collection of robust and reliable data (Kreuter 2013).

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

Therefore, to efficiently identify difficulty in surveys, robust and precise methods are vital to minimize measurement errors from respondents' data. To do so, this paper introduces a new model for multivariate functional data classification that uses novel semi-metrics to assess dissimilarities between trajectories.
The remainder of this paper is as follows. Section 2 describes the model, and Section 3 gives some preliminary results of the application.

2 The model

Consider the learning sample $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$ where $y_1, \ldots, y_n$ are values of a categorical random variable Y with classes $L = \{1, \ldots, l\}$. Consider also that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are realizations over $t \in \mathbb{T} \subset \mathbb{R}$ of independent and identically distributed copies of a multivariate functional random variable $\mathbf{X} \in \mathcal{F}$, where $\mathcal{F}$ is an appropriate space of d-dimensional functions. Hence, $\mathbf{x}(t) \in \mathbb{R}^d$ for each t with $d \geq 2$. Assume that $\mathbf{x}^{(a)}$ is the a-th derivative of $\mathbf{x}$ that exists and is square-integrable in $\mathcal{F}$. However, in real-world data, functions are evaluated over a finite grid where time is discretely observed, and that grid may differ between observed functions.
Suppose the following classification problem: a new observation $\mathbf{x}_*$ with unknown class membership y_* is given, and thus we want to infer it from the learning sample where predictors are multivariate functions.

2.1 Semi-metrics in the multivariate functional framework

Semi-metrics were proposed to measure distances and capture specific characteristics of functions. Formally, let $D(\mathbf{x}, \mathbf{x}_*)$ be the semi-metric between the functions $\mathbf{x}$ and $\mathbf{x}_*$ fulfilling:

$$
\begin{aligned}
&D(\mathbf{x}, \mathbf{x}_*) \geq 0 \\
&D(\mathbf{x}, \mathbf{x}) = 0 \\
&D(\mathbf{x}, \mathbf{x}_*) \leq D(\mathbf{x}, \tilde{\mathbf{x}}) + D(\tilde{\mathbf{x}}, \mathbf{x}_*),
\end{aligned}
$$

$\forall \mathbf{x}, \mathbf{x}_*, \tilde{\mathbf{x}} \in \mathcal{F}$. However, they are different from metrics in that $D(\mathbf{x}, \mathbf{x}_*) = 0$ does not always imply that $\mathbf{x} = \mathbf{x}_*$. In addition, they can also be computed on the functions' derivatives $\mathbf{x}^{(a)}$.
In the literature, several such semi-metrics have been proposed, notably by Ferraty and Vieu (2006), Fuchs et al. (2015) and Fuchs et al. (2017). However, these have been limited to the univariate functional case. Therefore, this work extends these semi-metrics to the multivariate case, and additionally, considers other distances such as the Frchet, Hausdorff, and Needleman-Wunsch distances, and also application-specific semi-metrics. Table 1 serves as an example of some of the semi-metrics studied in this work.

TABLE 1. Example of semi-metrics extended to functions from $\mathbb{T} \in (0,1)$ to $\mathbb{R}^d$.

	$D(\mathbf{x}, \mathbf{x}_*)$
Manhattan	$\sum_{k=1}^{d} \int_{\mathbb{T}} \mid x_{(k)}(t) - x_{*(k)}(t) \mid \mathrm{dt}$
Euclidean	$\left(\sum_{k=1}^{d} \int_{\mathbb{T}} \left(x_{(k)}(t) - x_{*(k)}(t) \right)^2 \mathrm{dt} \right)^{\frac{1}{2}}$
mean	$\left(\sum_{k=1}^{d} \left(\frac{1}{\lvert\mathbb{T}\rvert} \int_{\mathbb{T}} x_{(k)}(t)\mathrm{dt} - \frac{1}{\lvert\mathbb{T}\rvert} \int_{\mathbb{T}} x_{*(k)}(t)\mathrm{dt} \right)^2 \right)^{\frac{1}{2}}$
globMax	$\left(\sum_{k=1}^{d} \left(\max_{t\in\mathbb{T}}\left(x_{(k)}(t)\right) - \max_{t\in\mathbb{T}}\left(x_{*(k)}(t)\right) \right)^2 \right)^{\frac{1}{2}}$
globMin	$\left(\sum_{k=1}^{d} \left(\min_{t\in\mathbb{T}}\left(x_{(k)}(t)\right) - \min_{t\in\mathbb{T}}\left(x_{*(k)}(t)\right) \right)^2 \right)^{\frac{1}{2}}$
dynamic time warp (dtw)	$\left(\sum_{k=1}^{d} \int_{\mathbb{T}} \left(x_{(k)}(t) \circ \gamma^* - x_{*(k)}(t) \right)^2 \mathrm{dt} \right)^{\frac{1}{2}}$ $\gamma^* = \underset{\gamma\in\Gamma}{\operatorname{argmin}} = \left(\sum_{k=1}^{d} \int_{\mathbb{T}} \left(x_{(k)}(t) \circ \gamma - x_{*(k)}(t) \right)^2 \mathrm{dt} \right)^{\frac{1}{2}}$ $\circ$ denotes the composition operator and Γ denotes the set of all warping functions
Hausdorff	$\max \left\{ \sup_{t\in\mathbb{T}} \inf_{t'\in\mathbb{T}} \left(\left(\sum_{k=1}^{d} \int_{\mathbb{T}} \left(x_{(k)}(t) - x_{*(k)}(t') \right)^2 \mathrm{dt} \right)^{\frac{1}{2}} \right), \sup_{t'\in\mathbb{T}} \inf_{t\in\mathbb{T}} \left(\left(\sum_{k=1}^{d} \int_{\mathbb{T}} \left(x_{(k)}(t) - x_{*(k)}(t') \right)^2 \mathrm{dt} \right)^{\frac{1}{2}} \right) \right\}$
Frchet	$\inf_{\alpha,\beta} \max_{t\in\mathbb{T}} \left(\left(\sum_{k=1}^{d} \int_{\mathbb{T}} \left(x_{(k)}(\alpha(t)) - x_{*(k)}(\beta(t)) \right)^2 \mathrm{dt} \right)^{\frac{1}{2}} \right)$ where α, β are continuous non-decreasing functions

2.2 Ensemble

Following Ferraty and Vieu (2006), Fuchs et al. (2015) and Fuchs et al. (2017), classification models such as the functional k-nearest neighbour (FkNN) and non-parametric functional kernel estimator (NPFKE) can be used to predict the class membership of $\mathbf{x}_*$ according to the proximity to $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The idea of FkNN is to order the semi-metric values, and then predict the class membership of $\mathbf{x}_*$ with the most frequent category of the k closest functions. However, NPFKE weights the probabilities of class memberships of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to predict the probabilities of the class membership of $\mathbf{x}_*$. The weights are given by $\frac{K(D(\mathbf{x}_i,\mathbf{x}_*)/h)}{\sum_{i=1}^{n} K(D(\mathbf{x}_i,\mathbf{x}_*)/h)}$,where $K(\cdot)$ and h are the kernel function (e.g., gaussian or uniform kernels) and bandwidth parameter, respectively. Notice that these weights are assigned in accordance with the similarity between $\mathbf{x}_*$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
In this paper, the linear combination and the stacked-based ensembles of FkNN and NPFKE are extended to the multivariate functional case through new and appropriate semi-metrics. In the case of the stacked-based ensemble, random forest, boosting, and neural network with one hidden layer are considered super-learners candidates.

3 Application

The data analyzed in this work are based on a web survey where several questions were manipulated to create two scenarios of difficulty. One of these questions was related to the type of employment, for which easy and difficult versions with respectively concise and complex language were created and randomly assigned between respondents ($n = 551$). As they responded to the question, participants' mouse movements were collected in pairs of x- and y-coordinates which were time-normalized and considered as bivariate functional predictors.
Table 2 gives some preliminary accuracies of FkNN and NPFKE with semi-metrics in Table 1, application-specific semi-metrics and some preliminary ensembles; for example, the Euclidean distance between flips, hovers, and response times (RT). In survey research, x-flips, y-flips, and hovers are respectively, the number of directional changes in the horizontal direction, vertical direction, and periods without movement. In addition, a personalization method was considered to incorporate the respondents' baseline behavior into the model.To do so, mouse movements from five non-manipulated questions were added to the semi-metrics with smaller weights to incorporate the baseline behavior of the participants in these measures of similarity.
Sub-sampling cross-validation was used with 100 repetitions and weights of 70% and 30% in the training and testing sets, respectively. For the NPFKE, the normal kernel function was considered, and the weights for personalization were 0.5 and 0.1 for the target and baseline variables, respectively.

Since different semi-metrics capture different features of the trajectories, an ensemble will be considered next by combining several non-personalized and personalized semi-metrics to improve the accuracies in Table 2.

TABLE 2. Preliminary accuracies with FkNN and NPFKE models and semi–metrics either in Table 1 or application-specific semi-metrics.

model	semi-metric	unpersonalized			personalized		
		a=0	$a = 1$	a=2	a=0	$a = 1$	a=2
FkNN	Manhattan	0.567	0.536	0.554	0.564	0.539	0.546
	Euclidean	0.556	0.531	0.550	0.556	0.522	0.537
	mean	0.561	0.536	0.523	0.549	0.514	0.523
	globMax	0.593	0.543	0.530	0.601	0.545	0.521
	globMin	0.556	0.560	0.531	0.542	0.578	0.545
	dtw	0.546	0.526	0.528	-	-	-
	Hausdorff	0.581	0.559	0.512	-	-	-
	Frchet	0.562	-	-	-	-	-
	flips	0.530	-	-	-	-	-
	hovers	0.525	-	-	-	-	-
	RT	0.514	-	-	-	-	-
NPFKE	Manhattan	0.581	0.530	0.530	0.573	0.529	0.530
	Euclidean	0.569	0.528	0.534	0.562	0.528	0.530
	mean	0.563	0.531	0.527	0.559	0.531	0.529
	globMax	0.603	0.527	0.526	0.594	0.530	0.526
	globMin	0.548	0.551	0.531	0.527	0.567	0.540
	flips	0.526	-	-	-	-	-
	hovers	0.528	-	-	-	-	-
	RT	0.528	-	-	-	-	-

Acknowledgements

The authors acknowledge financial support from the German Research Foundation (DFG) through the grant "Statistical modeling using mouse movements to model measurement error and improve data quality in web surveys" (GR 3793/2-1 and KR 2211/5-1).

References

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

Fuchs, K., Gertheiss, J. and Tutz, G. (2015). Nearest neighbor ensembles for functional data with interpretable feature selection. *Chemometrics and Intelligent Laboratory Systems*, 146, 186 – 197.

Fuchs, K., Pößneckerb, W. and Tutz, G. (2017). *Classification of Functional Data with k-Nearest-Neighbor Ensembles by Fitting Constrained Multinomial Logit Models.* arXiv:1612.04710

Horwitz, R., Brockhaus, S., Henninger, F., Kieslich, P. J., Schierholz, M. et al. (2019). Learning from Mouse Movements: Improving Questionnaires and Respondents User Experience Through Passive Data Collection. In: *Advances in Questionnaire Design, Development, Evaluation and Testing*, Wiley & Sons, Ltd, 403 – 425.

Kreuter, F. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information.* John Wiley & Sons.

A computationally efficient estimator for large clustered non-Gaussian data

Alvaro J. Flórez [1], Geert Molenberghs[12], Geert Verbeke[2], Pavlos Mamouris[2], Bert Vaes[3]

[1] I-Biostat, Universiteit Hasselt, Belgium
[2] I-BioStat, KU Leuven, Belgium
[3] Academisch Centrum voor Huisartsgeneeskunde, Belgium

E-mail for correspondence: alvaro.florez@uhasselt.be

Abstract: The generalized linear mixed model (GLMM) is one of the most frequently used techniques to analyze clustered non-Gaussian data. Commonly, the GLMM is fitted by maximizing the marginal (log-)likelihood, i.e., integrating out the random effects. However, this whole maximisation may require a considerable amount of computing resources. Although computationally manageable with medium to large data, it can be too time-consuming or computationally intractable with very large clusters and/or with a large number of clusters. To overcome this, a fast two-stage estimator for correlated non-Gaussian data is presented. It is rooted in the pseudo-likelihood split-sample methodology. Based on simulations, it shows good statistical properties, and it is computationally much faster than full maximum likelihood. The approach is illustrated using a large dataset belonging to a network of Belgian general practices.

Keywords: Generalized linear mixed model; Hierarchical data; Random effects; Split-sample

1 Introduction

The analysis of clustered non-Gaussian data is commonly done within the generalized linear mixed model (GLMM) framework. In the GLMM methodology, we assume that, conditionally on normally distributed random effects, the outcomes are independent and their distribution belongs to the exponential family, encompassing models for a wide range of outcomes types, such as binary, count, and time-to-event. The main idea of including these random effects is to address correlation and some variability due to clustering.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

Let Y_{ij} be the jth outcome measured for cluster i, with $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. The GLMM assumes that, conditionally on a q-dimensional vector of random effects $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$, the outcomes Y_{ij} are independent with a density that belongs to the exponential family, that is:

$$f(y_{ij}|\boldsymbol{b}_i) = \exp\left\{\phi^{-1}\left[y_{ij}\theta_{ij} - \psi\left(\theta_{ij}\right)\right] + c\left(y_{ij}, \psi\right)\right\}, \tag{1}$$

where θ_{ij} and ϕ are called natural and scale parameter, respectively; $\psi(\cdot)$ and $c(\cdot, \cdot)$ are known functions. Here, the conditional mean μ_{ij} is modeled by a known link function, $\mu_{ij} = h\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i\right)$, where $X_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})$ and $Z_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i})$ are $(n_i \times p)$ and $(n_i \times q)$ matrices of covariates; and $\boldsymbol{\beta}$ is an unknown p-dimensional vector of fixed-effects coefficients.

Even though (1) is expressed in this hierarchical form, it is customarily fitted by maximizing the marginal (log-)likelihood, i.e.,

$$L\left(\boldsymbol{\beta}, D, \phi\right) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\boldsymbol{b}_i\right) f(\boldsymbol{b}_i|D) d\boldsymbol{b}_i. \tag{2}$$

As it can be seen from (2), maximization of the likelihood involves N integrals over the q-dimensional random effects $\boldsymbol{b}_i$. Except from the Gaussian case, the derivation of the marginal joint distribution can be complicated, or even not possible in analytical form. Therefore, the marginalization is done numerically, at the cost of requiring more computing resource.

To facilitate the estimation procedure with large datasets, Molenberghs et al (2011) proposed the split-sample methodology. Here, the sample is partitioned into K sub-samples, which are analyzed separately and afterwards the results are combined to obtain overall inferences. For clustered data, the most efficient partitioning consists of sub-samples with equally distributed clusters, i.e., clusters with the same design matrices (Molenberghs et al, 2018). However, this constraint is very restrictive in many cases, and we may end up in the most extreme case, all sub-samples with a single cluster, leading to the so-called cluster-by-cluster (CbC) estimator.

The paper is organized as follows. In Section 2, we propose the CbC estimator for a GLMM. The main findings of extensive simulations and a real data analysis are briefly showed in Section 3 and 4, respectively. Finally, Section 5 is reserved for concluding remarks. More details of the CbC estimator can be found in Flórez et al (2019a, 2019b, 2020).

2 Cluster-by-cluster estimator

The cluster-by-cluster (CbC) estimator follows the same two steps of the split-sample methodology. For simplicity, we will assume that $X_i = Z_i$. Nevertheless, the general expression requires some further but straightforward algebra.

Given the conditional independence assumption, in the first stage, we can fit a generalized linear model (GLM) within each cluster. Evidently, it requires that Z_i is full column rank within each cluster, allowing estimation of $(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_N)$.
In the second stage, a global estimator of $\boldsymbol{\beta}$ is obtained by weighted averaging the sets of estimates of each cluster. Then, the estimator and its variance are:

$$\widetilde{\boldsymbol{\beta}} = \sum_{i=1}^{N} A_i \widehat{\boldsymbol{\beta}}, \text{ and } V\left(\widetilde{\boldsymbol{\beta}}\right) = \sum_{i=1}^{N} A_i V\left(\widehat{\boldsymbol{\beta}}\right) A_i^T,$$

respectively. For the weighting matrices (A_i), we opt for an approximation of the so-called optimal weights (Molenberghs et al, 2018).
The variance matrix of the random effects (D) measures the variability between clusters, and consequently, it cannot be estimated using a single cluster. Hence, a method-of-moments approach is proposed. It is based on the sum of the cross-product of the difference between the cluster-specific estimates $(\widehat{\boldsymbol{\beta}}_i)$ and the global estimate $(\widetilde{\boldsymbol{\beta}})$, i.e., $S_{\boldsymbol{b}} = \sum_{i=1}^{N} \left(\widehat{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}}_i - \widetilde{\boldsymbol{\beta}}\right)^T$.
Then, the estimator is found by equating $S_{\boldsymbol{b}}$ to its expected value and solving for D. Since $E\left(\widetilde{\boldsymbol{b}}_i\right) \approx \mathbf{0}$,

$$E\left(\widetilde{\boldsymbol{b}}_i\right) \approx \sum_{i=1}^{N} (I - A_i) V\left(\widehat{\boldsymbol{\beta}}_i\right) (I - A_i)^T + \sum_{k \neq i} A_j V\left(\widehat{\boldsymbol{\beta}}_j\right) A_j^T, \qquad (3)$$

where $V\left(\widehat{\boldsymbol{\beta}}_i\right) \approx D + V\left(\widehat{\boldsymbol{\beta}}_i | \boldsymbol{b}_i\right)$. Depending on the type of outcome, $V\left(\widehat{\boldsymbol{\beta}}_i | \boldsymbol{b}_i\right)$ can be found analytically or approximated using Taylor series expansions.
Given that (3) is non-linear, an iterative procedure, e.g., Newton-Raphson, is needed to find the solution of D. Furthermore, an expression for the variance of $\widetilde{D}$ can be found using the delta method.

3 Simulation

For the data-generating model we consider the following model:

$$Y_{ij} | \boldsymbol{b}_i \sim \text{Bern.}\left[\pi_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}\right], \text{ and } Y_{ij} | \boldsymbol{b}_i \sim \text{Pois.}\left[\lambda_{ij} = \exp\left(\eta_{ij}\right)\right],$$

where $\eta_{ij} = \beta_0 + b_{0i} + z_i\beta_1 + x_{ij}(\beta_2 + b_{1i}) + z_i x_{ij}\beta_3$, x_{ij} is continuous covariate ranging in $[0, 1]$, z_i is a binary covariate, and $(b_{0i}, b_{1i})' \sim N(\mathbf{0}, \boldsymbol{D})$.

We set $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)' = (1, 0.1, -2, -1)'$ and $\boldsymbol{D} = (d_{11}, d_{12}, d_{22}) = (2, -0.5, 0.5)$, for the logistic model. On the other hand, for Poisson model, $\boldsymbol{\beta} = (1.5, -0.1, -0.5, -0.2)'$ and $\boldsymbol{D} = (0.4, -0.2, 0.6)$.
For the simulations, we fixed the number of clusters (N) and the cluster size (n_i) is generated by $n_i \sim N\left[\mu_n, (0.25\mu_n)^2\right]$ (rounded to the nearest integer). For the logistic model, $N = 100$ and $\mu_n = \{100, 200, 500\}$. For the Poisson model, $N = 50$ $\mu_n = \{50, 100, 200\}$.
For each scenario, 1000 datasets were generated and analyzed using the CbC estimator and the MLE based on adaptive quadrature. The comparison between both estimators is done by the relative bias (RB) and efficiency (RE), separately for each parameter. The later is defined as the ratio of the mean square error ratio of the CbC estimator over the MLE.
To evaluate the computational performance for large data, we set $N = 500$ and vary $\mu_n = \{500, 1000, 2000\}$. Here, we generated 25 datasets.

TABLE 1. Relative bias (in percentage) and efficiency of the CbC estimator for each parameter of the logistic and Poisson model with random slope.

	Logistic model						Poisson model					
	Relative bias(%)			Relative bias(%)			Relative bias(%)			Relative bias(%)		
	μ_n			μ_n			μ_n			μ_n		
Parm	100	200	500	100	200	500	50	100	200	50	100	200
β_0	3.8	1.7	0.8	1.12	1.04	1.02	0.1	-0.2	0.2	1.01	1.01	1.00
β_1	30.8	17.5	1.0	1.09	1.04	1.02	2.7	12.9	5.0	1.00	1.02	1.00
β_2	4.0	2.0	0.9	1.53	1.19	1.08	1.5	-0.2	2.0	1.01	1.00	1.00
β_3	5.5	3.4	0.7	1.34	1.14	1.06	1.8	-1.2	-6.4	1.01	1.00	1.01
d_{11}	6.6	3.5	0.9	1.57	1.34	1.17	3.4	0.7	0.7	1.16	1.07	1.05
d_{12}	20.3	9.3	2.0	1.95	1.59	1.22	5.1	2.8	1.9	1.13	1.11	1.09
d_{22}	42.2	16.3	5.9	3.50	1.93	1.28	6.6	3.6	0.4	1.19	1.12	1.03

Table 1 exhibits the relative bias and efficiency of the CbC estimator for each parameter of the logistic and Poisson model with random effects in all scenarios. For the Poisson model, the estimator of the fixed effects is practically unbiased and as efficient as the MLE. For the variance components, it is asymptotically efficient. For the logistic model, it provides somewhat biased estimates for all parameters, especially for the variance components. However, as in the Poisson case, the bias and the efficiency loss decrease with μ_n but at a slower rate.
Table 2 displays the median computation time, in seconds, of the CbC estimator and full MLE for the logistic and Poisson models with random effects. As expected, the CbC estimator is faster than the MLE for the three types of outcomes, with a larger computing time for the binary case.

4 Data analysis

The CbC estimator is implemented using different datasets from the Intego-project, a large database of continuous recording of patient information in a

TABLE 2. Median computation time (in seconds) for the cluster-by-cluster (CbC) estimator and the maximum likelihood estimator (MLE), in parenthesis the median ratio, for the logistic and Poisson models with random slopes.

	Logistic model			Poisson model		
Estimator	$\mu_n = 500$	$\mu_n = 1000$	$\mu_n = 2000$	$\mu_n = 500$	$\mu_n = 1000$	$\mu_n = 2000$
CbC	22.16(11.5)	35.13(14.5)	82.49(21)	3.05(27.8)	8.05(44.3)	8.78(102.3)
MLE	254.39(1)	508.72(1)	1735.22(1)	84.67(1)	356.23(1)	898(1)

network of Belgian general practices. The samples contain information from around 49 practices spread throughout Flanders, Belgium. The number of patients per practice ranges between 389 and 9676. Furthermore, there are missing data in the data (particularly in the variables used as covariates). Therefore, before fitting the GLMM, multiple imputation procedure was performed (drawing 20 multiply imputed datasets).

Here, we model the outcome hypertension in the Intego database from 2015. Let Y_{ij} be the outcome (absent/present) for patient j in practice i. The model is:

$$\begin{aligned} Y_{ij}|b_i \sim & \text{Bernoulli}\left(\pi_{ij}\right), \\ \text{logit}\left(\pi_{ij}\right) = & \beta_0 + b_i + \text{age}_{ij}\beta_1 + \text{gender}_{ij}\beta_2 + \text{BMI}_{ij}\beta_3 + \\ & \text{diabetes}_{ij}\beta_4 + \text{cholesterol}_{ij}\beta_4, \end{aligned} \tag{4}$$

where gender_{ij} and diabetes_{ij} are indicator variables. $\text{Gender}_{ij} = 1$ if the jth patient at practice i is male, $\text{gender}_{ij} = 0$ otherwise; $\text{diabetes}_{ij} = 1$ if the jth patient at practice i has diabetes, $\text{diabetes}_{ij} = 0$ otherwise. Furthermore, we assume $b_i \sim N(0, d)$.

Table 3 displays the estimates and standard errors of the parameters of the model (4) by the CbC estimator and MLE. Both estimators provide somewhat similar estimates and standard errors for the fixed effects. Regarding the variance of the random intercept, the CbC estimate, and its standard error, is slightly larger than the ones observed with the MLE. Based on the Wald test, all covariates have a significant effect. Regarding gender, the probability of suffering hypertension is 11% higher in men than in women. Age, BMI, diabetes, and systolic and diastolic blood pressure are also considered risk factors. Furthermore, fitting the CbC estimator for each multiply imputed dataset needed around 3 seconds. On the other hand, the MLE took more than 50 times as long.

5 Concluding remarks

Given the statistical and computational properties, we suggest that the CbC estimator is an attractive alternative to fit a GLMM with several large-size clusters. Although large clusters is not common in a longitudinal study, this can be encountered in another hierarchical settings, such as meta-analyses.

TABLE 3. Intego data - binary case. Parameter estimates for the logistic model after with random intercept multiple imputation by the cluster-by-cluster (CbC) estimator and maximum likelihood estimator (MLE).

		CbC		MLE	
Effect	Parm	Est	S.E.	Est	S.E.
Intercept	β_0	-2.4	0.1	-2.304	0.0790
Age	β_1	0.06	0.0007	0.057	0.0007
Gender	β_2	-0.112	0.0225	-0.109	0.0220
BMI	β_3	0.066	0.0039	0.066	0.0037
Diabetes	β_4	0.854	0.0296	0.837	0.0285
Cholesterol	β_5	-0.003	0.0004	-0.003	0.0004
Systolic	β_6	0.022	0.0012	0.022	0.0012
Diastolic	β_7	0.018	0.0024	0.018	0.0023
Var rand. eff	d	0.308	0.0602	0.286	0.0134

Although the estimator still has attractive properties with medium cluster-sizes, its implementation can be problematic, especially during the first stage. With few observations or several zeros in a cluster (in the binary or Poisson cases), the GLM estimator may diverge or converge to a spurious solution, leading to unstable overall estimates. Therefore, we suggest performing a sensitivity analysis by excluding any problematic clusters and evaluating the overall estimates. Furthermore, the addition of weights in the estimator of D reduces the influence of small and unstable clusters.

References

Flórez, A. J., G., Molenberghs, G., Verbeke, G. and Alonso Abad, A. (2019a). A closed-form estimator for meta-analysis and surrogate markers evaluation. *Journal of Biopharmaceutical Statistics*, **29(2)**, 318 – 332.

Flórez, A. J., G., Molenberghs, G., Verbeke, G., Kenward, M., Mamouris, P. and Vaes, B. (2019b). Fast two-stage estimator for clustered count data with overdispersion. *Journal of Statistical Computation and Simulation*, **89(14)**, 2678 – 2693.

Flórez, A. J., G., Molenberghs, G., Verbeke, G., Mamouris, P. and Vaes, B. (2020). A computationally efficient estimator for large clustered non-Gaussian data. *Submitted for publication*, 1 – 30.

Molenberghs, G., Verbeke, G. and Iddi, S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics & Probability Letters*, **81(7)**, 892 – 901.

Molenberghs, G., Hermans, L., Nassiri, V., Kenward, M., Vand der Elst, W., Aerts, M. and Verbeke, G. (2018). Clusters with random size: maximum likelihood versus weighted estimation. *Statistica Sinica*, **28(3)**, 1107 – 1132.

Capture–recapture in case of one-inflation

Herwig Friedl[1], Dankmar Böhning[2]

[1] Institute of Statistics, Graz University of Technology, Austria
[2] Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

E-mail for correspondence: hfriedl@tugraz.at

Abstract: Estimating the size of a hard-to-count population is a challenging matter. In particular, when only few observations of the population to be estimated are available. The matter gets even more complex when one-inflation occurs. This situation is illustrated at hand of a real problem: the size of a dice snake population in Graz (Austria). The paper discusses how one-inflation can be easily handled in likelihood approaches and also discusses how variances and confidence intervals can be obtained.

Keywords: Population Size; Zero-Truncation; One-Inflation.

1 Introduction and motivation

Tranninger (2018) tried to estimate the size of a dice snake population along the river Mur in Graz (Austria). The work was motivated by a resettlement project of the population due to the development of a water power plant in the vicinity of the living ground of the dice snakes. The major question was: how many dice snakes are there? In the year 2014 in which there were 31 capture occasions between April and September, snakes were found under artificial hiding places and photos of their undersides were taken. These photos then allowed to uniquely identify each animal repeatedly. Hence, the count X informs about the number of identifications of each animal. However, there is the well-known complication (Böhning et al., 2018; McCrea and Morgan, 2015) that any population unit with $X_i = 0$ would not be observed leading to a reduced observed sample. The empirical distribution of X is provided in Table 1. The objective now is to estimate the size N of such an elusive target population. However, we have that $f_0 = N - n$ is unknown. The frequency f_0 is also labelled as the *dark* or *hidden* figure and its estimate is the prime interest here.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

TABLE 1. Frequency distribution of count X of repeated snake identifications.

x	0	1	2	3	4	5	$\cdots$	observed size
f_x	–	f_1	f_2	f_3	f_4	f_5	$\cdots$	$n = f_1 + f_2 + \cdots$
snake sightings	–	59	8	1	1	1		$n = 70$

2 Modelling

For predicting f_0 some sort of modelling is unavoidable as the nonparametric estimates f_x $(0 < x)$ carry no information for f_0. Hence, we need a model for $P_\theta(X = x) = p_x(\theta)$ so that an estimate $\hat{\theta}$ can be found. This leads to fitted probabilities $p_x(\hat{\theta})$ for $0 \leq x$. In particular, we can use the Horvitz-Thompson-type estimator for estimating f_0, i.e.

$$\hat{f}_0 = n\frac{p_0(\hat{\theta})}{1 - p_0(\hat{\theta})}\,, \tag{1}$$

from which, ultimately, the population size estimator $\hat{N} = n + \hat{f}_0$ follows. For valid inference, the valid specification of the model $p_x(\theta)$ is crucial. Since we see a large number of counts of ones, the *singletons*, we are concerned about *one-inflation*, a situation where more counts of ones occur than compatible with the baseline model $p_1(\theta)$ as this can lead to a highly inflated estimate of f_0. To accommodate one-inflation we need to include it into the model as

$$p'_x(\theta) = \begin{cases} (1-\alpha) + \alpha p_x^+(\theta), & x = 1 \\ \alpha p_x^+(\theta), & x \neq 1\,, \end{cases} \tag{2}$$

where $p_x^+(\theta) = p_x(\theta)/(1 - p_0(\theta))$ is a zero-truncated base distribution. The modelling is greatly simplified using the general result in Böhning and van der Heijden (2019). Consider an *arbitrary* inflation point x_I and an *arbitrary* count pmf $p_x(\theta)$ with associated x_I-inflation as

$$p'_x(\theta) = \begin{cases} (1-\alpha) + \alpha p_x(\theta), & x = x_I \\ \alpha p_x(\theta), & x \neq x_I\,, \end{cases}$$

where $\alpha \in [0, 1]$. The associated log-likelihood

$$\log L(\theta, \alpha | x) = f_{x_I} \log[1 - \alpha + \alpha p_{x_I}(\theta)] + \sum_{x \neq x_{x_I}} f_x \log p_x(\theta) + (n - f_{x_I}) \log \alpha\,,$$

is maximized in

$$\hat{\alpha} = \frac{1 - f_{x_I}/n}{1 - p_{x_I}(\theta)} \tag{3}$$

for fixed θ: Thus, the x_{x_I}-inflated profile log-likelihood function

$$\begin{aligned}\log L(\theta, \hat{\alpha}|x) &= f_{x_I} \log(f_{x_I}/n) + (n - f_{x_I}) \log(1 - f_{x_I}/n) \\ &\quad + \sum_{x \neq x_{x_I}} f_x \log\left(\frac{p_x(\theta)}{1 - p_{x_I}(\theta)}\right)\end{aligned}$$

equals the x_{x_I}-truncated log-likelihood

$$\sum_{x \neq x_{x_I}} f_x \log\left(\frac{p_x(\theta)}{1 - p_{x_I}(\theta)}\right)$$

plus a term that is independent of θ. This implies that x_{x_I}-inflation models can be simply fitted by x_{x_I}-truncated models.
Accounting for one-inflation ($x_I = 1$) and utilizing the above result we restrict inference on the *zero–one*-truncated pmf

$$p_x^{++}(\theta) = \frac{p_x(\theta)}{1 - p_0(\theta) - p_1(\theta)}, \qquad x = 2, 3, \ldots, \tag{4}$$

which then provides the one–inflated, zero–truncated density.

3 Horvitz-Thompson estimation

The Horvitz-Thompson estimator (1) has the property $E(\hat{f}_0) = Np_0(\theta)$, if there is no inflation. A modification is needed here as n contains the one-inflated part. This leads to

$$\hat{f}_0 = (n - f_1)\frac{p_0(\hat{\theta})}{1 - p_0(\hat{\theta}) - p_1(\hat{\theta})}, \tag{5}$$

which is again unbiased for $Np_0(\theta)$ and, ultimately, we can define the *modified Horvitz-Thompson estimator* $\hat{N} = n + \hat{f}_0$, which is unbiased for N, if the base distribution is correctly specified.
Table 2 contains the estimated population size under a geometric base model. This model choice does much better than the Poisson and also better, in some sense, than the negative binomial base model.The conventional estimator (cHTE) uses the zero-truncated geometric distribution whereas the modified estimator (mHTE) uses the zero-one-truncated geometric as described above.

4 Marginal (unconditional) likelihood

So far we maximized the conditional (zero-truncated) likelihood of the observed counts. Now we discuss the general sampling mechanism that generated the data. Let m be the largest number of sightings, then the joint

TABLE 2. Population size estimates under a zero-one-truncated model (mHTE) and a zero-truncated geometric model (cHTE).

	$\hat{N}$	
n	mHTE	cHTE
70	127	358

marginal pmf of the sample is a multinomial model defined on the counts $0, 1, \ldots, m$ from a population of size N. Since we only observe counts of $1, \ldots, m$, the conditional model is the zero-truncated multinomial for the n observed counts. This conditioning process is described by a binomial variable that splits the population into an observed (of size n) and unobserved part (of size $N - n = f_0$). Together we have

$$\begin{aligned}\text{multinom}(p_0(\theta), \ldots, p_m(\theta)|N) &= \text{multinom}\Big(\frac{p_1(\theta)}{1-p_0(\theta)}, \ldots, \frac{p_m(\theta)}{1-p_0(\theta)}\Big|n\Big) \\ &\quad \times \text{binom}(1-p_0(\theta)|N)\,,\end{aligned}$$

or equivalently

$$\begin{aligned}\frac{N!}{f_0!f_1!\cdots f_m!}\prod_{x=0}^{m} p_x(\theta)^{f_x} &= \frac{n!}{f_1!\cdots f_m!}\prod_{x=1}^{m}\Big(\frac{p_x(\theta)}{1-p_0(\theta)}\Big)^{f_x} \\ &\quad \times \frac{N!}{f_0!n!}p_0(\theta)^{f_0}(1-p_0(\theta))^n\,,\end{aligned}$$

which proofs the validity of the factorization.
Since $f_1, \ldots, f_m$ are fixed given the observed counts, the relevant part of the marginal likelihood is

$$L(f_0, \theta|f_1, \ldots, f_m) = \frac{N!}{f_0!}\prod_{x=0}^{m} p_x(\theta)^{f_x}\,.$$

Thus, we maximize the marginal log-likelihood function

$$\ell(f_0, \theta|f_1, \ldots, f_m) = \sum_{x=0}^{m} f_x \log p_x(\theta) + \log(N!/f_0!)\,.$$

For a given value of f_0, the θ-score function is

$$\frac{\partial}{\partial\theta}\ell(f_0, \theta|\cdot) = \sum_{x=0}^{m} f_x \frac{dp_x(\theta)/d\theta}{p_x(\theta)}\,.$$

If we specify the base model to be geometric, i.e. $p_x(\theta) = \theta(1-\theta)^x$, then

$$\frac{dp_x(\theta)/d\theta}{p_x(\theta)} = \frac{(1-\theta) - x\theta}{\theta(1-\theta)}$$

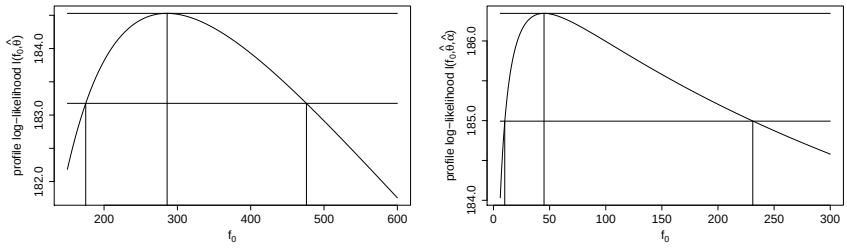


FIGURE 1. Marginal profile log-likelihood functions under a geometric model (left) and under a one-inflated geometric model (right).

and the marginal maximum likelihood estimator becomes

$$\hat{\theta} = \left(1 + \frac{1}{N}\sum_{x=1}^{m} x f_x\right)^{-1} .$$

This estimator depends on the value of N and thus on the unknown f_0. We propose to evaluate the marginal profile log-likelihood $\ell(f_0, \hat{\theta}|\cdot)$ for a grid of f_0 values to find the maximizer $\hat{f}_0$. This is shown in Figure 1.
Since $\hat{f}_0 = 286$ with 90% profile confidence interval $(175, 476)$ for f_0, the total size of the population is estimated to be 356 snakes, which seems to be a plausible number. This marginal estimate can now be compared to the conditional estimate $\hat{N}_c = 358$ given in Table 2.
Under an arbitrary **one-inflated** count model the conditional log-likelihood function is

$$f_1 \log(1 - \alpha + \alpha p_1(\theta)) + \sum_{x \neq 1} f_x \log \frac{p_x(\theta)}{1 - p_1(\theta)} , \qquad x = 0, 2, \ldots, m .$$

Adding the respective binomial part finally gives its marginal version as

$$\ell(f_0, \theta, \alpha | f_1, \ldots, f_m) = f_1 \log(1-\alpha+\alpha p_1(\theta)) + \sum_{x \neq 1} f_x \log \frac{p_x(\theta)}{1 - p_1(\theta)} + \log(N!/f_0!) .$$

Since $\hat{\alpha}$ defined in (3) for $x_I = 1$ maximizes the conditional as also the marginal likelihood, we define the marginal profile log-likelihood as

$$\begin{aligned} \ell(f_0, \theta, \hat{\alpha} | f_1, \ldots, f_m) &= f_1 \log(f_1/N) + (N - f_1) \log(1 - f_1/N) \\ &\quad + \sum_{x \neq 1} f_x \log \frac{p_x(\theta)}{1 - p_1(\theta)} + \log(N!/f_0!) . \end{aligned}$$

Under the geometric one-inflated situation the relevant term depending on θ becomes

$$\sum_{x \neq 1} f_x \log \frac{p_x(\theta)}{1 - p_1(\theta)} = \log \frac{\theta}{1 - \theta(1 - \theta)} \sum_{x \neq 1} f_x + \log(1 - \theta) \sum_{x \neq 1} f_x x$$

where $x = 0, 2, \ldots, m$. With

$$N_{(-1)} = \sum_{x \neq 1} f_x \quad \text{and} \quad S_{(-1)} = \sum_{x \neq 1} f_x x = \sum_{x=2}^{m} f_x x$$

the above marginal profile log-likelihood simplifies to

$$\begin{aligned} \ell(f_0, \theta, \hat{\alpha} | f_1, \ldots, f_m) &= N_{(-1)} \left(\log(1 - f_1/N) + \log \frac{\theta}{1 - \theta(1-\theta)} \right) \\ &\quad + f_1 \log(f_1/N) + S_{(-1)} \log(1-\theta) + \log(N!/f_0!) \end{aligned}$$

with corresponding θ-score function

$$\frac{\partial}{\partial \theta} \ell(f_0, \theta, \hat{\alpha} | f_1, \ldots, f_m) = N_{(-1)} \Big(\frac{1}{\theta} + \frac{1 - 2\theta}{1 - \theta(1-\theta)} \Big) - S_{(-1)} \frac{1}{1-\theta} \, .$$

Since $N_{(-1)}$ is a sum over all frequencies except f_1, this score function actually depends on both, θ and the unobserved f_0. Thus, it is natural to find the maximizer of the marginal profile log-likelihood using again a grid of f_0 values and maximize the corresponding likelihood function in θ conditional on each f_0 value which is shown in Figure 1.
The estimate $\hat{f}_0 = 45$ maximizes the profile likelihood. Therefore, the respective population size estimate $\hat{N} = 115$ is rather small but compares well with the conditional estimate $\hat{N} = 127$ in Table 2. A perhaps disadvantageous result is the fairly wide 90% profile confidence interval $(10, 231)$, reflecting the enormous variance of the estimator in this application.

References

Böhning, D., van der Heijden, P.G.M., and Bunge, J. (2018). *Capture-Recapture Methods for the Social and Medical Sciences*. Boca Raton: Chapman & Hall/CRC.

Böhning, D. and van der Heijden, P.G.M. (2019). The identity of the zero-truncated, one-inflated likelihood and the zero-one-truncated likelihood for general count densities with an application to drink-driving in Britain. *Annals of Applied Statistics*, **13** 1198 – 1211.

McCrea, R.S. and Morgan, B.J.T. (2015). *Analysis of Capture-Recapture Data*. Boca Raton: Chapman & Hall/CRC.

Tranninger, J. (2018). *The size of the dice snake population at the river Mur in Graz (Austria)*. Master Thesis, Institute of Statistics, Graz University of Technology, Austria.

Inference for the overlap coefficient based on P-splines and Dirichlet process mixtures

Javier E. Garrido Guillén[1], Vanda Inácio[1], María Xosé Rodríguez–Álvarez [2]

[1] School of Mathematics, University of Edinburgh, UK
[2] BCAM - Basque Center for Applied Mathematics & IKERBASQUE, Spain

E-mail for correspondence: javier.garridog@ed.ac.uk

Abstract: Accurate diagnosis of disease is of great importance in clinical practice and medical research. Before a diagnostic test is routinely used in practice its ability to discriminate between diseased and nondiseased states must be rigorously assessed. Further, its performance may depend on covariates (e.g., age and/or gender). This motivates us to propose the covariate-specific overlap coefficient, which will help to determine the optimal populations where to perform the tests on. We assume a location-scale regression model for the test outcomes in each group, relying on an additive formulation based on Penalised splines, while the regression error follows a Dirichlet process mixture of normal distributions. Our approach is illustrated through an application concerning diagnosis of diabetes.

Keywords: Diagnostic test; Dirichlet process mixtures; Overlap coefficient; Penalised splines.

1 Introduction

Disease diagnosis is a fundamental task in clinical practice and medical research. The ability of a diagnostic test to distinguish diseased from nondiseased individuals must be thoroughly evaluated before the test can be widely used in practice. Furthermore, in many situations the behaviour of the test may be influenced by external covariates.
The overlap coefficient (OVL), defined as the proportion of overlap area between two density functions, has been proposed as a summary measure of diagnostic accuracy. An OVL value of zero means that the distributions do not overlap at all (perfect diagnostic accuracy), whereas a value of one means that the distributions are identical and thus, the test is useless from

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

a diagnostic viewpoint. Let $Y_{\bar{D}}$ and Y_D be two independent continuous random variables representing the test outcomes from the nondiseased and diseased group, with covariate vectors $\boldsymbol{X}_{\bar{D}}$ and $\boldsymbol{X}_D$, and conditional density functions given by $f_{\bar{D}}(\cdot \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x}_{\bar{D}})$ and $f_D(\cdot \mid \boldsymbol{X}_D = \boldsymbol{x}_D)$, respectively. Given a covariates value $\boldsymbol{x}$, the covariate-specific overlap coefficient (cOVL) is defined as

$$\mathrm{OVL}(\boldsymbol{x}) = \int_{-\infty}^{\infty} \min\left\{f_{\bar{D}}(y \mid \boldsymbol{X}_{\bar{D}} = \boldsymbol{x}), f_D(y \mid \boldsymbol{X}_D = \boldsymbol{x})\right\} dy. \qquad (1)$$

The goal of this work is to propose a flexible Bayesian method to estimate the cOVL, so that it can be used for many populations and large number of diseases. Further, by working under a Bayesian context, point and interval estimates are obtained into a single integrated framework.

2 Methods

Let $\{(y_{\bar{D}i}, \boldsymbol{x}_{\bar{D}i}^T)\}_{i=1}^{n_{\bar{D}}}$ and $\{(y_{Dj}, \boldsymbol{x}_{Dj}^T)\}_{j=1}^{n_D}$ be independent random samples of size $n_{\bar{D}}$ and n_D from the nondiseased and diseased population, respectively. Further, let $\boldsymbol{x}_{\bar{D}i} = (x_{\bar{D}i,1}, \ldots, x_{\bar{D}i,p})^T$ and $\boldsymbol{x}_{Dj} = (x_{Dj,1}, \ldots, x_{Dj,p})^T$ be two p–dimensional covariate vectors, for $i = 1, \ldots, n_{\bar{D}}$ and $j = 1, \ldots, n_D$. For the sake of simplicity, we will assume that all the covariates are continuous and affect both the location and scale of each group. However, our modelling approach can easily incorporate categorical covariates as well as interactions between continuous and categorical covariates for both components. In what follows, we will describe our modelling procedure only with the diseased population as the same one is applicable to the nondiseased group. We will assume a location-scale regression model for the test outcomes where the error follows a Dirichlet process mixture of normal distributions. Such setting induces the following conditional density for the test outcomes in the diseased group

$$f(y_{Di} \mid \boldsymbol{X}_D = \boldsymbol{x}_{Di}) = \int \phi(y_{Di} \mid \eta_D(\boldsymbol{x}_{Di}) + \mu, s_D(\boldsymbol{x}_{Di})\sigma^2)\mathrm{d}G_D(\mu, \sigma^2),$$

where $\phi(\cdot \mid \mu, \sigma^2)$ is the density function of the normal distribution with mean μ and variance σ^2, and G_D follows a Dirichlet process with precision parameter $\alpha_D > 0$ and baseline distribution $G_D^*(\mu, \sigma^2)$. For conjugacy reasons and to ensure identifiability, we set $G_D^*(\mu, \sigma^2) = \mathrm{N}(\mu \mid 0, b_\mu^2)\mathrm{IG}(\sigma^2 \mid a_{\sigma^2}, b_{\sigma^2})$. Moreover, to allow us to easily simulate from the posterior distribution, we will employ a truncated stick-breaking construction for G_D, therefore

$$f(y_{Di} \mid \boldsymbol{X}_D = \boldsymbol{x}_{Di}) = \sum_{l=1}^{L_D} \omega_{Dl}\phi(y_{Di} \mid \eta_D(\boldsymbol{x}_{Di}) + \mu_{Dl}, s_D(\boldsymbol{x}_{Di})\sigma_{Dl}^2),$$

where $(\mu_{Dl}, \sigma^2_{Dl}) \overset{\text{iid}}{\sim} G^*_D$, and the weights are such that $\omega_{D1} = v_{D1}$, $\omega_{Dl} = v_{Dl}\prod_{t<l}(1-v_{Dt})$, for $l = 2, \ldots, L_D$; the inputs v's are distributed according to a beta distribution, i.e., $v_{D1}, \ldots, v_{DL_D-1} \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_D)$ and $v_{DL_D} = 1$. Regarding the specification of the predictors η_D and s_D, we model the mean and variance as an additive combination of smooth functions (Kobayashi and Ogasawara, 2016), that is,

$$\eta_D(\boldsymbol{x}_{Di}) = h_{D1}(x_{Di,1}) + \cdots + h_{Dp}(x_{Di,p}),$$
$$s_D(\boldsymbol{x}_{Di}) = \exp\left\{g_{D1}(x_{Di,1}) + \cdots + g_{Dp}(x_{Di,p})\right\},$$

where each smooth function is approximated by a cubic B-splines basis. For example, for h_{Dj} and $j = 1, \ldots, p$, let $x_{Dj,\min} = \xi_{Dj,0} < \cdots < \xi_{Dj,m_j} = x_{Dj,\max}$ be equally spaced knots. Thus, we can write h_{Dj} as a linear combination of $R_j = m_j + 3$ B-splines basis functions $B^h_{Dj,r}$, that is,

$$h_{Dj}(x_{Di,j}) = \sum_{r=1}^{R_j} B^h_{Dj,r}(x_{Di,j})\beta_{Dj,r} = \boldsymbol{B}^h_{Dj}(x_{Di,j})\boldsymbol{\beta}_{Dj},$$

where $\boldsymbol{\beta}_{Dj} = (\beta_{Dj,1}, \ldots, \beta_{Dj,R_j})^T$ is the corresponding vector of coefficients. Similarly, g_{Dj} can be approximated using $\boldsymbol{B}^g_{Dj}(x_{Di,j})\boldsymbol{\delta}_{Dj}$. It is well-known that the position and number of knots can have a large influence on the fitted functions. To overcome this problem we will use penalised splines (P-splines), where the penalty is based on differences of adjacent B-splines coefficients as described in (Eilers and Marx, 1996). We will follow the Bayesian P-splines approach proposed by Lang and Brezger (2004), where second-order random walk priors are assumed for all coefficients. More precisely, for $\boldsymbol{\beta}_{Dj}$

$$\beta_{Dj,r} = 2\beta_{Dj,r-1} - \beta_{Dj,r-2} + u_{Dj,r}, \quad r = 3, \ldots, R_j, \quad j = 1, \ldots, p,$$

where $u_{Dj,r} \overset{\text{iid}}{\sim} \text{N}(0, \tau^2_{Dj})$. The random walk variance τ^2_{Dj} controls the smoothness of the fitted functions. For $\boldsymbol{\delta}_{Dj}$, the prior variance is denoted by ψ^2_{Dj}. Note that to ensure identifiability, all functions h_{Dj} and g_{Dj}, are centred around zero. To complete our model specification, let

$$\alpha_D \sim \Gamma(a_\alpha, b_\alpha), \quad \tau^{-2}_{Dj} \sim \Gamma(a_{\tau^2}, b_{\tau^2}), \quad \text{and} \quad \psi^{-2}_{Dj} \sim \Gamma(a_{\psi^2}, b_{\psi^2}),$$

where $\Gamma(a, b)$ stands for a gamma distribution with shape a and rate b. Because explicit full conditionals are available, we implemented a Gibbs sampler to simulate from the posterior distribution. Finally, to get an estimate of the cOVL, the integral involved in (1) is approximated numerically (trapezoidal rule).

3 Application

We applied our method to data from a population based survey of diabetes in Cairo, Egypt. The data comprises measurements on 88 subjects with

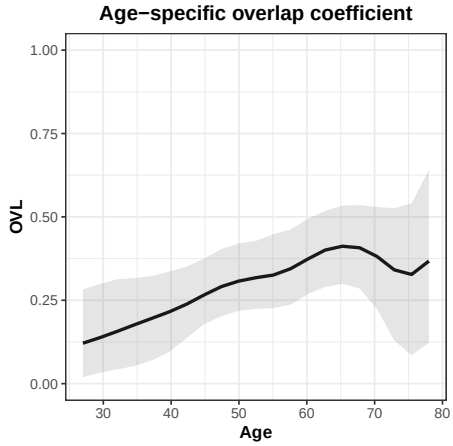


FIGURE 1. Posterior mean and 95% credible bands for the age-specific OVL.

diabetes and 198 non diabetic. Our primary goal is to evaluate the effect of age in the accuracy of glucose as a biomarker of diabetes.
In Figure 1, we depict the estimated age-specific coefficient of overlap, where we can see that it increases with age, thus meaning that the accuracy of the glucose levels as a marker of diabetes decreases with age. Figure 2 shows the estimated posterior (mean) mean functions for the non-diabetic (left) and diabetic (right) group. And finally, Figure 3 displays the conditional histograms and densities at ages of 41 and 60 (left and right panels, respectively) for the non diabetic (top row) and diabetic (bottom row) group. These values correspond to the first and third quartiles of the covariate age, respectively.

References

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with *B*-Splines and Penalties. *Statistical Science*, **11**(2), 89 – 102.

Kobayashi, G., and Ogasawara, K. (2016). Bayesian Nonparametric Instrumental Variable Regression Approach to Quantile Inference. *arXiv preprint arXiv:1608.07921.*

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**(1), 183 – 212.

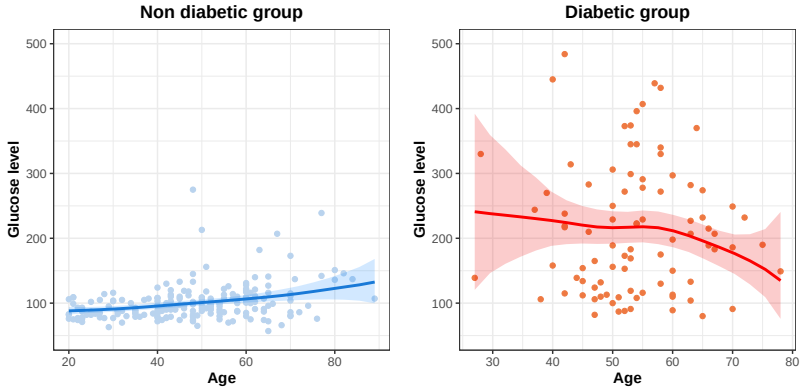


FIGURE 2. Mean function: posterior mean and 95% pointwise credible bands for the non diabetic (top left) and diabetic (top right) group.

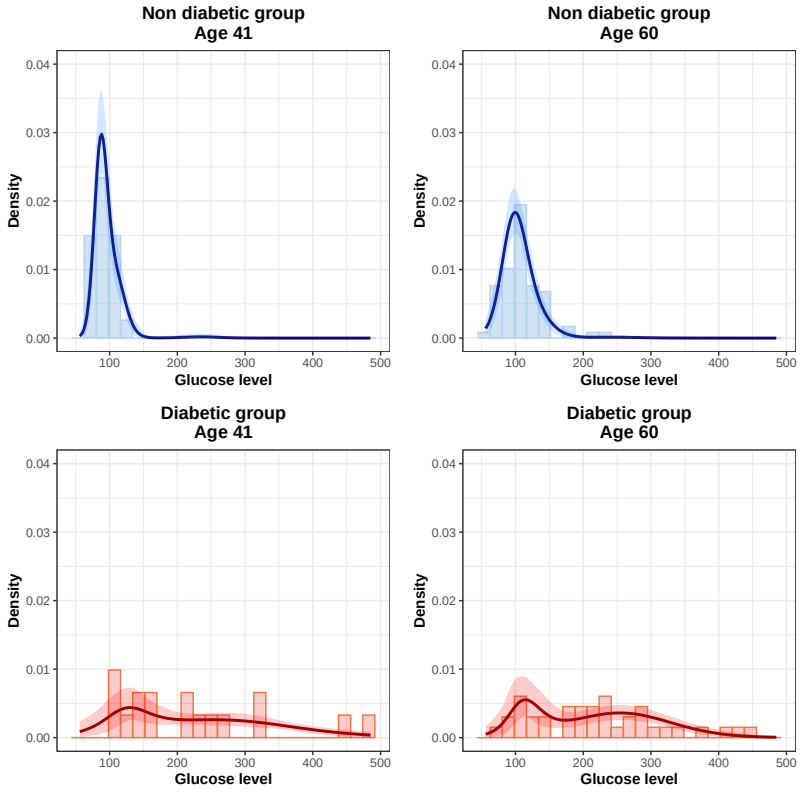


FIGURE 3. Conditional histograms and densities: posterior mean and 95% credible bands of the conditional densities at ages of 41 (left) and 60 (right) for the non diabetic (top row) and diabetic (bottom row) group.

Deducing neighborhoods of classes from a fitted classification model

Alexander Gerharz[1], Andreas Groll[1], Gunther Schauberger[2]

[1] Chair of Statistical Methods for Big Data, Department of Statistics, TU Dortmund University, Dortmund, Germany
[2] Chair of Epidemiology, Department of Sport and Health Sciences, Technical University of Munich, Munich, Germany

E-mail for correspondence: `gerharz@tu-dortmund.de`

Abstract: For complex statistical or machine learning models, *interpretable machine learning* methods can be used to make up for the lack of interpretability. The method proposed here helps to understand the partitioning of the feature space into predicted classes in a classification model. Basically, it observes the changes of the predictions after slight manipulations of specific metric features. The observed changes can then be interpreted as neighboring classes in the feature space. An example is shown with the `iris` classification task.

Keywords: Interpretable Machine Learning; Explainable Artificial Intelligence; Classification Task; Feature Space Partitioning; Chordgraphs.

1 Introduction

Currently, various interpretable machine learning (IML) methods exist, including traditional methods like the permutation feature importance as described by Breiman (2001), but also more recent methods like the usage of anchors (see, e.g., Ribeiro et al., 2018). The latter is used to find specific features and their respective feature values that determine the prediction of an observation, while the other features could be randomly altered without affecting the prediction too much.
Our proposed method works similar to the anchor method, but has a completely different purpose and interpretation. Its aim is to find neighboring classes in a fitted classification model by using small manipulations of metric features of interest and observing the changes of the predictions. Hence, instead of trying to find features and the respective feature values that de-

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

termine a prediction, we try to find those features changing the prediction when slightly manipulated, and then to interpret those changes.

2 Methodology

In the following, we will set the mathematical background for our new IML method. The aim is to slightly increase or decrease the value of metric features of interest and observe the changes in the predicted classes.
Suppose $\hat{f}(\mathbf{x})$ is a final model fitted for a classification task with K different classes, $K \geq 2$, and let L be the set of all the features used for this classification with a specific set-size $p = |L|$. Then, $\hat{\pi}_{\hat{f},k}(\mathbf{x}_i)$ denotes the estimated probability by the model $\hat{f}(\cdot)$ for observation i with feature vector $\mathbf{x}_i$ to belong to a specific class $k \in \{1, \ldots, K\}$. Next, we determine

$$k^*_{\hat{f}}(\mathbf{x}_i) = argmax_k\{\hat{\pi}_{\hat{f},k}(\mathbf{x}_i)\}\,,$$

i.e. $k^*_{\hat{f}}(\mathbf{x}_i)$ is the class with the highest probability as estimated by the model $\hat{f}(\cdot)$ for the observation $\mathbf{x}_i$ (from here on index $\hat{f}$ is dropped for better readability, as we will now always refer to the same fitted model).
Next, we choose a subset $M \subseteq L$ containing the features of interest. Mostly the size of set M is 1. Now, $\tilde{\mathbf{x}}_i$ represents the feature-vector for observation i, whose features from M each were manipulated componentwisely by a small amount. The manipulation is done by increasing or decreasing the quantile-function of the subset M containing the features of interest. For this purpose, a small value q_l is added componentwisely to $\hat{F}_l(\cdot)$, which denotes the empirical cumulative distribution function (ecdf) for all features $l = 1, ..., p$ (see Figure 1), with

$$q_l = \begin{cases} u_l, & for\ l \in M\ \ and\ \ u_l \in [-1, 1] \\ 0, & else\,. \end{cases}$$

To prevent extrapolation in the quantile function $\hat{F}_l^{-1}(\alpha)$, α is chosen from the interval $[0, 1]$. Now, we define for all $l = 1, ..., p$ componentwisely:

$$\begin{aligned} & \hat{F}_l(\tilde{\mathbf{x}}_{i,l}) = max\{min\{\hat{F}_l(\mathbf{x}_{i,l}) + q_l, 1\}, 0\} \\ \Longrightarrow \quad & \tilde{\mathbf{x}}_{i,l} \quad\ = \hat{F}_l^{-1}(max\{min\{\hat{F}_l(\mathbf{x}_{i,l}) + q_l, 1\}, 0\}). \end{aligned}$$

The modifying values q_l for each $l \in M$ are set by the user. Note that the inverse of the ecdf $\hat{F}_l^{-1}$ does not necessarily exist, as $\hat{F}_l$ is not necessarily continuous. Hence, we have to define

$$\hat{F}_l^{-1}(\alpha) = inf\{x : \hat{F}_l(x) \geq \alpha\}.$$

Now, $\tilde{\mathbf{x}}_i$ is the new manipulated observation, which has the same values as $\mathbf{x}_i$ for those covariates from $L \backslash M$, but different values for the features from

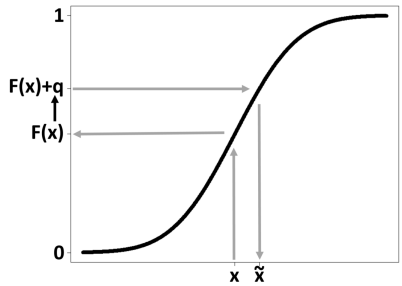


FIGURE 1. How to determine $\tilde{\mathbf{x}}$ for a specific feature.

M, which were increased or decreased componentwisely by the values that corresponded to manipulations by the amount q_l of the respective ecdf. Finally, for observation $i, i = 1, ..., n$ and $\mathbf{q} = (q_1, ..., q_p)^{\mathrm{T}}$, corresponding to the chosen $M \subseteq L$ and modifications $u_1, \ldots, u_p$, let $C_{\mathbf{q}}(\mathbf{x}_i)$ define the pair of the original and the (potentially) new class prediction resulting from this manipulation, i.e.

$$\begin{aligned} C_{\mathbf{q}}(\mathbf{x}_i) &= (k^*(\mathbf{x}_i), k^*(\tilde{\mathbf{x}}_i)) \\ &= (\hat{y}_{i,old}, \hat{y}_{i,new}). \end{aligned}$$

We obtain $\hat{y}_{i,old} = \hat{y}_{i,new}$, if the predicted class **has not changed** by manipulating $\mathbf{x}_{i,M}$, and $\hat{y}_{i,old} \neq \hat{y}_{i,new}$ otherwise.
The results could now be given in form of a migration matrix for all observations $i = 1, ..., n$, where the rows indicate the predicted classes of an observation before the manipulation of $\mathbf{x}_{i,M}$ and the columns indicate its predicted classes after the manipulation. The trace of this migration matrix counts the number of observations that have not changed classes despite the manipulation. The off-diagonal elements aggregate the number of observations that have changed their predicted class from the class indicated by the respective row to the predicted class indicated by the respective column. A general example of a migration matrix can be found in Table 1.

TABLE 1. Exemplary migration matrix for two classes.

	A_{after}	B_{after}
A_{before}	$n_{A\to A}$	$n_{A\to B}$
B_{before}	$n_{B\to A}$	$n_{B\to B}$

The off-diagonal elements of Table 1 can be interpreted as follows:

- if $n_{A\to B} > 0$, an area in the feature space is found, where class B is classified to be next to class A via the manipulation of the feature subset M

- if $n_{A \to B} = 0$, no area is found, where class B is classified to be next to class A via the manipulation of the feature subset M - but it could still exist! (the used manipulation might have been too weak or too strong)

Of course, $n_{B \to A}$ can be interpreted analogously. These migration matrices can then be visualized by using chordgraphs as shown in the application example below.
To guarantee that the method yields good interpretability, certain rules should be followed: (i) This method is built to find closely neighboring classes within the feature space. By using a too strong manipulation a class could be skipped and no direct neighborhood is found. To prove or disprove that a direct neighborhood between two classes exists, the complete feature space would have to be filled with infinitely many data points and manipulations with infinitely small steps would have to be performed. (ii) If we define the preference order

$$A \succ B := \begin{array}{l}\text{class } A \text{ is directly (or generally) next to class } B \text{ in} \\ \text{the direction of the manipulation,}\end{array}$$

then due to the possibility of a very complex partitioning of the feature space one could have

$$A \succ B \wedge B \succ C \not\Rightarrow A \succ C.$$

This expresses that the results of this method can not be interpreted transitively.

3 Application on the `Iris` Data

Next, we illustrate our method on the `iris` flowers dataset, an easy to understand standard example from classification (see Anderson, 1936, or Fisher, 1936). For the sake of simplicity, we fit a simple *classification tree* based on just two of the original four features, namely `Petal.Length` and `Petal.Width`. Figure 2 (left) displays the feature space partitioning by the fitted tree. It turns out that the class `virginica` is modeled "above" the class `versicolor` with respect to `Petal.Width`. When slightly increasing the `Petal.Width`, this neighborhood is found by the method as shown in the corresponding chordgraph in Figure 2 (right), where $n_{versicolor \to virginica} = 3$ is indicated by the lightgrey strang of chords starting at the class `versicolor` and ending in the class `virginica`.
As this example contains only two features, the classification model can be visualized in a 2-dimensional way as shown in Figure 2 (left). This means, it is easy to compare the neighborly indication by the cordgraph with the underlying model. If one would include more features for the classification task, graphical visualization would get rather complex and difficult. These

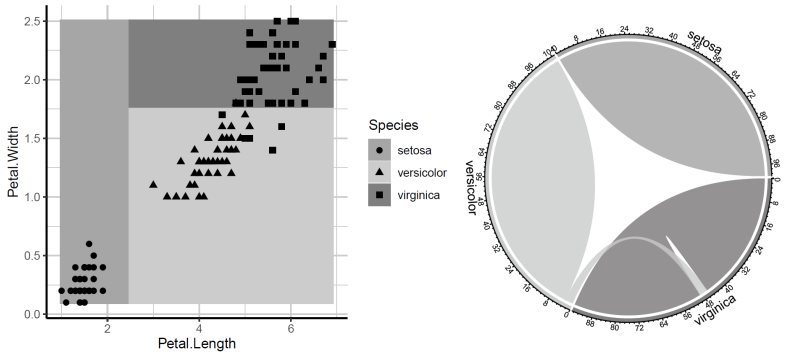


FIGURE 2. Feature space partitioning by a classification tree for the `iris` dataset (left); chordgraph for increasing `Petal.Width` with $q_{petal.width} = +0.02$ (right).

difficulties already arise when a single additional feature is added to the model and the corresponding partitioning of the feature space is visualized in a 3-dimensional graph. However, the proposed method overcomes these limitations. For a classification task an arbitrary amount of features can be used and the method would still indicate neighborly modeled classes as determined by the underlying model, even for high-dimensional feature spaces.

The original `iris` dataset contains two additional features, namely the `Sepal.Length` and `Sepal.Width`. When adding these features to the model, the complete partitioning of the feature space can not be displayed in a 2-dimensional way, but would require a 4-dimensional visualization. However, the proposed method could still show neighborly modeled classes in this 4-dimensional feature space in regard of specific manipulations.

Another benefit of the proposed method is that it is not restricted to a specific type of classification model. In the present example a simple classification tree was used, whose splits could be looked at directly and neighborly modeled classes could be determined by an experienced user. When using more complex models, e.g. a random forest, then it is hard to determine neighborhoods of the predicted classes just by directly looking at the multiple trees. The method proposed here, however, overcomes this limitation and is able to determine neighborly modeled classes even if the classification task in the present example would have been done with a highly complex model (or even with a black-box model).

4 Discussion

For classification models the method proposed here can help to determine neighborly modeled classes with regard to specific features of interest. When applying this method to a specific classification task, it is neither

limited to a specific amount of features, nor to a specific amount of classes of the target, and it is not bound to a specific set of classification models. As the chordgraph is a visualization tool to show relations between multiple classes and even can indicate the direction of the relation, it is a very good graphical tool to present the results of the proposed method. All in all, this method provides an alternative to improve interpretability of very complex models, which is the true benefit of this method. Thus, it qualifies as an interpretable machine learning method and adds some more variety to this field.

As of right now, the here proposed method can be directly used to determine neighborhoods between classes modeled by classification models with regard to specific manipulations of some features of interest. But there are a few edge cases, which demand further research, e.g. how the method handles multiple observations with equal values of a feature of interest. Multiple solutions are currently being worked on for a fair treatment of all observations in these cases. Another point is a comparable treatment of positive and negative choices of q. As of right now, even the tiniest positive manipulation would lead to a shift of all the values of a feature of interest, but a very tiny negative manipulation would not lead to a shift of even a single value of the same feature of interest. We currently work on a solution to this, which will be presented soon.

References

Anderson, E. (1936). The species problem in iris. *Annals of the Missouri Botanical Garden*, **23**, 457

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5 – 32.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179 – 188.

Ribeiro, M.T., Singh, S., Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *AAAI Conference on Artificial Intelligence*.

Median bias reduction in cumulative link models

Vincenzo Gioia[1], Euloge Clovis Kenne Pagui[2], Alessandra Salvan[2]

[1] Department of Economics and Statistics, University of Udine, Italy
[2] Department of Statistical Sciences, University of Padua, Italy

E-mail for correspondence: gioia.vincenzo@spes.uniud.it

Abstract: For cumulative link models, we propose a new estimation approach aiming at median bias reduction (Kenne Pagui et al., 2017). Such approach is based on an adjustment of the score function. The method does not require finiteness of the maximum likelihood estimate and is effective in preventing boundary estimates. The resulting estimator is componentwise third-order median unbiased in the continuous case and equivariant under componentwise monotone reparameterizations. Simulation studies and an application compare the proposed method with maximum likelihood and mean bias reduction.

Keywords: Adjusted score; Boundary estimates; Likelihood; Ordinal data.

1 Introduction

Ordinal responses are very common in many contexts, especially in the social sciences, in medical disciplines and in business analysis. Cumulative link models, proposed by McCullagh (1980), see also Agresti (2010), are the most popular tool to handle ordinal data. The reason relies on the use of a single regression parameter for all response levels, making the effects simple to interpret. For these models, maximum likelihood (ML) is the estimation method of choice. However, with small samples or sparse data, the asymptotic approximation for the distribution of the ML estimator may poorly reflect the exact sampling distribution that may be centered away from the true parameter value. Another problem with ML estimation lies in boundary estimates, which can arise with positive probability in models for ordinal data.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

Kosmidis (2014) developed mean bias reduction (meanBR) for cumulative link models following Firth (1993), whose adjusted score does not require finiteness of the ML estimate. An alternative modification of the score equation is proposed in Kenne Pagui et al. (2017), aiming at median bias reduction (medianBR). Like meanBR, medianBR estimation does not require finiteness of the ML estimate and is effective in preventing boundary estimates. The medianBR estimator is componentwise third-order median unbiased in the continuous case and equivariant under componentwise monotone reparameterizations. Here we develop medianBR for cumulative link models following the simplified algebraic form of the adjustment term in Kenne Pagui et al. (2019). We show, through simulation studies and an application, that the new method outperforms ML and is competitive with meanBR.

2 Cumulative link models

Let Y_i be the ordinal outcome, with c categories, for subject i, $i = 1, \ldots, n$. Let p_{ij} be the probability to observe category j, $j = 1, \ldots, c-1$, for subject i, and $\Pr(Y_i \leq j) = \sum_{k=1}^{j} p_{ik}$ the cumulative probability. With $\boldsymbol{x}_i$ a p-dimensional row vector of covariates, cumulative link models assume

$$g\{\Pr(Y_i \leq j|\boldsymbol{x_i})\} = \alpha_j + \boldsymbol{x_i}\beta, \qquad j = 1, \ldots, c-1,$$

where $g(\cdot)$ is a given link function and $\beta^{\mathrm{T}} = (\beta_1, \ldots, \beta_p)$ is the regression parameter vector. Therefore, the effects of $\boldsymbol{x}_i$, expressed through β, are the same for each $j = 1, \ldots, c-1$. The intercept parameters α_j satisfy $-\infty = \alpha_0 \leq \alpha_1 \leq \ldots \leq \alpha_{c-1} \leq \alpha_c = +\infty$, so that $\Pr(Y_i \leq j|\boldsymbol{x}_i)$ is increasing in j for each fixed $\boldsymbol{x}_i$. A problem with ML estimation lies in boundary estimates, that is estimates of β with infinite components, and/or consecutive intercept estimates having the same value.

3 Median bias reduction

For a general parametric model with p-dimensional parameter θ and log-likelihood $\ell(\theta)$, based on a sample of size n, let $U_r = U_r(\theta) = \partial\ell(\theta)/\partial\theta_r$ be the r-th component of the score function $U(\theta)$, $r = 1, \ldots, p$. Let $j(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}$ be the observed information and $i(\theta) = E_\theta\{j(\theta)\}$ the expected information, which we assume to be of order $O(n)$. The medianBR estimator, $\tilde{\theta}$, is obtained as solution of the estimating equation $\tilde{U}(\theta) = 0$, based on the adjusted score (Kenne Pagui et al., 2019)

$$\tilde{U}(\theta) = U(\theta) + \tilde{A}(\theta),$$

with $\tilde{A}(\theta) = A^*(\theta) - i(\theta)F(\theta)$. The vector $A^*(\theta)$ has components $A_r^* = \frac{1}{2}\mathrm{tr}\{i(\theta)^{-1}(P_r + Q_r)\}$, with $P_r = E_\theta\{U(\theta)U(\theta)^{\mathrm{T}}U_r\}$ and $Q_r =$

$E_\theta\{-j(\theta)U_r\}$, $r = 1, \ldots, p$. The vector $F(\theta)$ has components $F_r = [i(\theta)^{-1}]_r^{\mathrm{T}}\tilde{F}_r$, where $\tilde{F}_r$ has elements $\tilde{F}_{r,t} = \mathrm{tr}[h_r\{(1/3)P_t + (1/2)Q_t\}]$, $r, t = 1, \ldots, p$, with the matrix h_r obtained as $h_r = \{[i(\theta)^{-1}]_r[i(\theta)^{-1}]_r^{\mathrm{T}}\}/i^{rr}(\theta)$, $r = 1, \ldots, p$. Above, we denoted by $[i(\theta)^{-1}]_r$ the r-th column of $i(\theta)^{-1}$ and by $i^{rr}(\theta)$ the (r, r) element of $i(\theta)^{-1}$.
In general, the equation $\tilde{U}(\theta) = 0$ needs to be solved numerically using a Fisher scoring-type algorithm.
In the continuous case, each component of $\tilde{\theta}$, $\tilde{\theta}_r$, $r = 1, \ldots, p$, is median unbiased with error of order $O(n^{-3/2})$, i.e. $\Pr_\theta(\tilde{\theta}_r \leq \theta_r) = \frac{1}{2} + O(n^{-3/2})$, compared with $O(n^{-1/2})$ of ML estimator. Moreover, the asymptotic distribution of $\tilde{\theta}$ is the same as that of the ML and the meanBR estimators, that is $\mathcal{N}_p(\theta, i(\theta)^{-1})$.

4 Simulation results

Through simulation studies, we compared medianBR with ML and meanBR, in terms of empirical probability of underestimation (PU%), estimated relative bias (RB%) and empirical coverage of the 95% Wald-type confidence interval (WALD%). We consider different sample sizes, $n = 50, 100, 200$ and the logit link function. We generate the covariate x_1 from a standard Normal, x_2 and x_3 from Bernoulli distributions with probabilities 0.5 and 0.8 respectively, and x_4 from a Poisson with mean 2.5. Assuming that the response has three categories, we fit the model

$$\mathrm{logit}\{\Pr(Y_i \leq j|\boldsymbol{x_i})\} = \alpha_j + \sum_{k=1}^{4} x_{ik}\beta_k, \qquad j = 1, 2;\ i = 1, \ldots, n,$$

considering 10,000 replications, with covariates fixed at the observed value and true parameter $\theta_0 = (-1, 2, 1, -1, 1, -1)$. Figure 1 shows the numerical results for the regression parameters. We found 2.82% and 0.08% simulated samples with ML boundary estimates, for $n = 50$ and $n = 100$, respectively. Instead, meanBR and medianBR estimates are always finite. The new method proves to be remarkably accurate in achieving median centering and it shows a lower estimated mean bias than ML, as well as a good empirical coverage for confidence intervals. Unreported simulation results shown similar behaviors considering probit and complementary log-log link functions.

5 An application

We consider the wine dataset analyzed in Christensen (2019), based on Randall (1989), concerning a factorial experiment for investigating the effects of two factors on the bitterness of wine, evaluated according to five

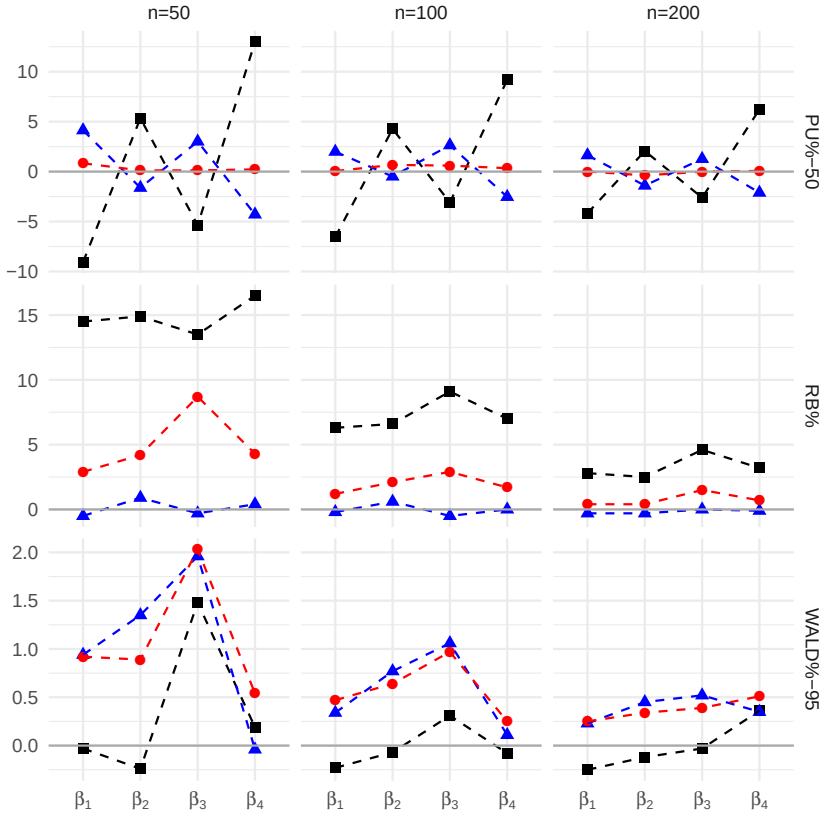


FIGURE 1. Simulation results of regression parameters, $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$, for ML (black square), meanBR (blue triangle) and medianBR (red circle) estimators. For ML, RB% and WALD% are conditional upon finiteness of the estimates.

ordered categories. The two factors are temperature at the time of crashing the grapes (x_1) and contact between juice and skin (x_2), each of them with two levels. For each of the four treatment conditions, two bottles were assessed by a panel of nine judges, giving $n = 72$ observations in all. We consider, as in Christensen (2019, Section 4.8), the outcomes obtained combining the three central categories and we fit the model

$$\text{logit}\{\Pr(Y_i \leq j|\boldsymbol{x_i})\} = \alpha_j + x_{i1}\beta_1 + x_{i2}\beta_2, \qquad j = 1, 2;\ i = 1, \ldots, 72.$$

Table 1 shows the ML, meanBR and medianBR estimates. Both meanBR and medianBR approaches are effective in preventing boundary estimates. Table 2 shows simulation results for the regression parameters with 10,000 replications, covariates fixed at the observed value and true parameter

$\theta_0 = (-1, 4, -2, -1)$. Whereas ML boundary estimates occurr in 9.79% of simulated samples, meanBR and medianBR estimates are always finite. It appears that the medianBR estimator is preferable in terms of PU%, outperforms ML in terms of RB% and shows a good empirical coverage for confidence intervals.

TABLE 1. ML, meanBR and medianBR estimates (s.e).

Method	α_1	α_2	β_1	β_2
ML	-1.32 (0.53)	$+\infty$ ($+\infty$)	$-\infty$ ($+\infty$)	-1.31 (0.71)
meanBR	-1.25 (0.51)	5.48 (1.48)	-3.43 (1.42)	-1.19 (0.67)
medianBR	-1.29 (0.52)	6.46 (2.32)	-4.48 (2.29)	-1.24 (0.68)

TABLE 2. Simulation results of regression coefficients $\beta = (\beta_1, \beta_2)$. For ML, RB% and WALD% are conditional upon finiteness of the estimates.

	Parameter β_1			Parameter β_2		
Method	PU%	RB%	WALD%	PU%	RB%	WALD%
ML	55.08	1.80	96.92	53.20	8.20	96.50
meanBR	43.91	-0.65	95.88	48.10	0.50	96.60
medianBR	49.71	8.95	96.48	50.35	4.90	96.28

References

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

Christensen, R. H. B. (2019). ordinal - regression models for ordinal data. *R package version 2019.12 – 10* http://CRAN.R-project.org/package=ordinal.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27 – 38.

Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika*, **104**, 923 – 938.

Kenne Pagui, E. C., Salvan, A. and Sartori, N. (2019). Efficient implementation of median bias reduction. *Submitted*, https://arxiv.org/abs/-2004.08630.

Kosmidis, I. (2014). Improved estimation in cumulative link models. *Journal of the Royal Statistical Society, Series B*, **76**, 169 – 196.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, **42**, 109 – 142.

Randall, J. H. (1989). The analysis of sensory data by generalized linear model. *Biometrical Journal*, **31**, 781 – 793.

Addressing cluster-constant covariates in mixed effects models via likelihood-based boosting techniques

Colin Griesbach[1], Andreas Groll[2], Elisabeth Waldmann[1]

[1] Friedrich-Alexander University Erlangen-Nürnberg, Germany
[2] TU Dortmund University, Germany

E-mail for correspondence: colin.griesbach@fau.de

Abstract: Boosting techniques from the field of statistical learning have grown to be a popular tool for estimating and selecting predictor effects in various regression models and can roughly be separated in two general approaches, namely gradient boosting and likelihood-based boosting. An extensive framework has been proposed in order to fit generalised mixed models based on boosting, however for the case of cluster-constant covariates likelihood-based boosting approaches tend to mischoose variables in the selection step leading to wrong estimates. We propose an improved boosting algorithm for linear mixed models, where the random effects are properly weighted, disentangled from the fixed effects updating scheme and corrected for correlations with cluster-constant covariates in order to improve quality of estimates and in addition reduce the computational effort. The method outperforms current state-of-the-art approaches from boosting and maximum likelihood inference.

Keywords: Statistical learning; Variable selection; Likelihood boosting; Prediction analysis.

1 Introduction

An extensive framework has been proposed in [2, 3] in order to fit various mixed models with likelihood-based boosting techniques [1] and is included in the R package `GMMBoost`. However, algorithms like `bGLMM` from the `GMMBoost` package tend to struggle with cluster-constant covariates, e.g. baseline covariates like gender or treatment group in longitudinal studies. As shown in Figure 1, this malfunction already occurs in a very basic data example with the popular Orthodont dataset available in the `nlme` pack-

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

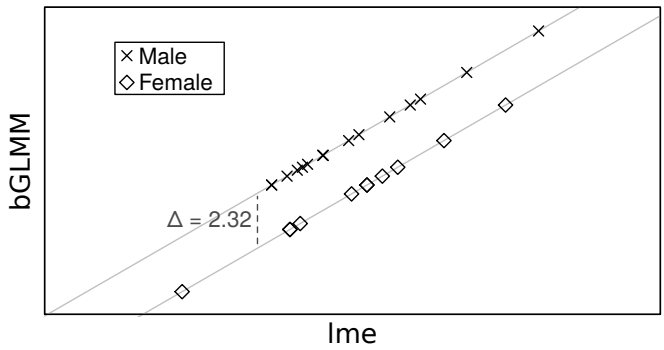


FIGURE 1. Comparison between random intercept estimates by `lme` and `bGLMM`.

age. A basic linear mixed model with random intercepts returns the two coefficient estimates $\hat{\beta}^{\mathrm{l}}_{\mathrm{sex}} = -2.32$ by `lme` and $\hat{\beta}^{\mathrm{b}}_{\mathrm{sex}} = 0.00$ by `bGLMM` for the effect of the cluster-constant covariate gender. The reason for this difference becomes clear when looking at the random intercepts, where `bGLMM` tends to compensate the missing effect for gender by assigning every female subject a random intercept lowered by 2.32.
We propose an updated algorithm with various changes in order to solve this identifiability issue and hence avoid the phenomenon of random intercepts growing too quickly. The major improvements include undocking the random effects update from the fixed effects boosting scheme and introducing a correction step for the random effects estimation to avoid possible correlations with observed covariates.

2 Methods

2.1 Model

We consider the linear mixed model

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

with design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ for fixed and random effects, variance-covariance-matrix $\boldsymbol{Q}$ for the random components and model error $\boldsymbol{\varepsilon}$ with variance σ^2. In order to perform likelihood inference, we formulate the penalized log-likelihood

$$\ell^{\mathrm{pen}}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{Q}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) - \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\gamma}_i^T \boldsymbol{Q}^{-1} \boldsymbol{\gamma}_i,$$

which is going to be maximized simultaneously for the effect estimates and random components by likelihood-based boosting.

2.2 Boosting Modifications

Amongst smaller modifications like weakening the random effects update and using smaller starting values for random components, we incorporate two major improvements into the algorithm.

Separate random effects update. A first update for the random effects $\boldsymbol{\gamma}$ is obtained by calculating

$$\boldsymbol{s}_{\text{ran}}(\boldsymbol{\gamma}) = \frac{\partial \ell^{\text{pen}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}, \quad \boldsymbol{F}_{\text{ran}}(\boldsymbol{\gamma}) = -\mathbb{E}\left[\frac{\partial^2 \ell^{\text{pen}}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\right]$$

and weakly updating

$$\tilde{\boldsymbol{\gamma}}^{[m]} = \hat{\boldsymbol{\gamma}}^{[m-1]} + \nu \boldsymbol{F}_{\text{ran}}(\boldsymbol{\gamma})^{-1} \boldsymbol{s}_{\text{ran}}(\boldsymbol{\gamma}).$$

The disentanglement of the random effects update from the fixed effects updating scheme guarantees a fair comparison of the single fixed effects, where the random effects do not play a crucial role. In addition the Fisher matrix

$$\boldsymbol{F}_{\text{ran}}(\boldsymbol{\gamma}) = diag(\boldsymbol{F}_1, \ldots, \boldsymbol{F}_n), \quad \boldsymbol{F}_i = \sigma^{-2} \boldsymbol{Z}_i^T \boldsymbol{Z}_i + \boldsymbol{Q}^{-1}$$

now has block-diagonal form making the inversion much easier and thus strongly reducing the computational effort.

Random effects correction. An additional correction is needed in order to solve the identifiability problem. Hence, instead of using the unaltered random intercept estimate $\tilde{\boldsymbol{\gamma}}_{\bullet 1}^{[m]}$, we proceed with the orthogonalised estimates

$$\hat{\boldsymbol{\gamma}}_{\bullet 1}^{[m]} = \tilde{\boldsymbol{\gamma}}_{\bullet 1}^{[m]} - (\tilde{\boldsymbol{X}}_{\text{c}}^T \tilde{\boldsymbol{X}}_{\text{c}})^{-1} \tilde{\boldsymbol{X}}_{\text{c}}^T \tilde{\boldsymbol{\gamma}}_{\bullet 1}^{[m]},$$

which result from counting out the orthogonal projections of $\tilde{\boldsymbol{\gamma}}_{\bullet 1}^{[m]}$ onto the subspace generated by the cluster-constant covariates $\tilde{\boldsymbol{X}}_{\text{c}}$. This ensures that the resulting estimates $\hat{\boldsymbol{\gamma}}_{\bullet 1}^{[m]}$ are uncorrelated with any cluster-constant covariates.

3 Data Examples

The new algorithm proved to perform well as shown by an extensive simulation study. In the following it is also evaluated based on two real world applications. The first one focuses solely on the novel estimation procedure regarding the random effects, the second showcases variable selection and shrinkage properties of the algorithm.

3.1 Orthodont Data

Applied to the Orthodont data mentioned in the beginning, the modified algorithm (`boostLMM`) yields the coefficient estimates for the covariates age and gender as well as the random intercepts variance $\boldsymbol{Q}$ depicted in Table 1. It is evident that `boostLMM` solves the random effects issues occurring with

TABLE 1. Estimates for the Orthodont dataset.

	$\hat{\beta}_0$	$\hat{\beta}_{\text{sex}}$	$\hat{\beta}_{\text{age}}$	$\hat{\boldsymbol{Q}}$
`lme`	17.71	-2.32	0.66	3.27
`boostLMM`	17.71	-2.32	0.66	3.11
`bGLMM`	16.82	0.00	0.65	5.41

`bGLMM`. Both the maximum likelihood approach in `lme` as well as `boostLMM` return matching estimates for fixed and random effects without any shift, which can be seen in Figure 2.

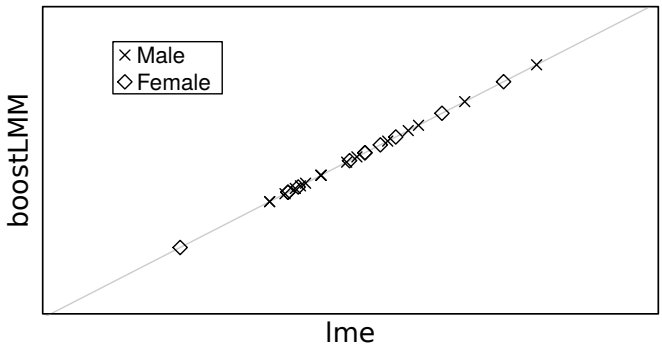


FIGURE 2. Comparison between random intercept estimates by `lme` and `boostLMM`.

3.2 Primary Biliary Cirrhosis

The popular primary biliary cirrhosis (PBC) dataset from 1994 tracks the change of the serum bilirubin level for a total of 312 PBC patients randomized into a treatment and a placebo group and additionally contains baseline covariates as well as follow-up measurements of several biomarkers. The serum bilirubin level, here modelled as the response variable, is considered a strong indicator for disease progression, hence an appropriate quantification of the impact of the given covariates on the serum bilirubin

level will lead to an adequate prediction model for the health status of PBC patients. Using boosting to carry out this quantification will optimize the prediction properties. Based on 10-fold cross validation, `boostLMM` determined $m_* = 93$ as the best performing number of iterations yielding the corresponding coefficient paths displayed in Figure 3. The algorithm

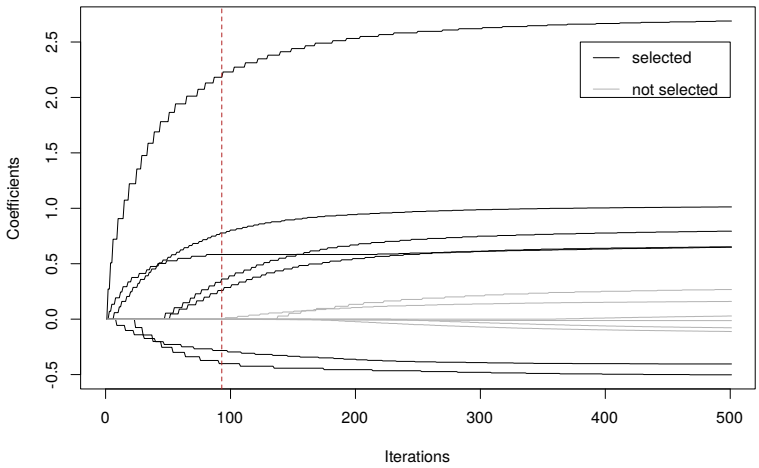


FIGURE 3. Coefficient paths for the PBC dataset generated by `boostLMM`.

stopped before six of the 13 covariates got selected into the model and thus their coefficient estimates are set to zero. The remaining effect estimates experienced various amounts of shrinkage in comparison to the maximum likelihood solution which prevents overfitting and hence offers an improved quality of prediction.

4 Conclusion

The updated algorithm is due to its minor and major tweaks capable of dealing with cluster-constant covariates in linear mixed models by preventing the random effects from taking up too much space. In addition, it preserves the well-known advantages of boosting techniques in general by offering variable selection and a good functionality even in high dimensional setups. As a very important side effect the computational effort receives a tremendous decrease making the algorithm more applicable to real world scenarios.

Acknowledgments: The work on this article was supported by the Interdisciplinary Center for Clinical Research (IZKF) of the Friedrich-Alexander-University Erlangen-Nürnberg (Project J61) and the Volkswagen Foundation.

References

Tutz, G. and Binder, H. (2006). Generalized additive models with implicit variable selection by likelihood-based boosting. *Biometrics*, **62** (4), 961 – 971

Tutz, G. and Groll, A. (2010). Generalized linear mixed models based on boosting. *Kneib, Thomas and Tutz, Gerhard (Eds.): Statistical Modelling and Regression Structures - Festschrift in the Honour of Ludwig Fahrmeir*, 197 – 216

Tutz, G. and Groll, A. (2013). Likelihood-based boosting in binary and ordinal random effects models. *Journal of Computational and Graphical Statistics*, **22** (2), 356 – 378

Beyond unidimensional poverty analysis using distributional copula models for mixed ordered-continuous outcomes

Maike Hohberg[1], Francesco Donat[2], Giampiero Marra[3], Thomas Kneib[1]

[1] Chair of Statistics, University of Goettingen, Germany
[2] Single Resolution Board, Brussels, Belgium
[3] Department of Statistical Science, University College London, UK

E-mail for correspondence: mhohber@uni-goettingen.de

Abstract: Poverty is a multidimensional concept often comprising a monetary outcome and other welfare dimensions such as education, subjective well-being or health, that are measured on an ordinal scale. In applied research, multidimensional poverty is ubiquitously assessed by studying each poverty dimension independently in univariate regression models or by combining several poverty dimensions into a scalar index. This inhibits a thorough analysis of the potentially varying interdependence between the poverty dimensions. We propose a multivariate copula generalized additive model for location, scale and shape (copula GAMLSS or distributional copula model) to tackle this challenge and we demonstrate its power by studying two important poverty dimensions: income and education. Since the level of education is often measured on an ordinal scale and income is continuous, we extend the bivariate copula GAMLSS to the case of mixed ordered-continuous outcomes. The new model is integrated into the `GJRM` package in `R` and applied to data from Indonesia.

Keywords: GAMLSS; copula; poverty

1 Introduction

Although poverty is widely regarded a multidimensional phenomenon and poverty measures moving beyond a single monetary dimension – such as the Multidimensional Poverty Index (MPI) – have emerged, little progress

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

[2]This paper should not be reported as representing the views of the Single Resolution Board. The views expressed are those of the authors and do not necessarily reflect those of the Board.

has been made on *analysing* poverty as a multidimensional concept. To study poverty at the micro level, univariate linear regression models are the standard tool but require either studying each poverty dimension separately in different equations, or using as response variable an index that subsumes all dimensions in a single number. Both approaches neglect the interdependence between poverty dimensions and ignore that the dependence itself should be part of the analysis. To overcome such limitations multivariate regression can be used to tackle multidimensionality in poverty analyses. The relationship between two or more outcomes can be modeled using copulas which have been proven to be useful and flexible tools in this regard.

A second issue in poverty analysis concerns distributional aspects. Especially for inequality and vulnerability analyses, it is important that poverty studies move beyond the simple mean effects. Generalized additive models for location, scale, and shape (GAMLSS, Rigby and Stasinopoulos, 2005) are able to capture the effects of covariates on the whole conditional distribution of a single poverty dimension.

Both issues of multidimensionality and distributional aspects can be addressed with a combination of GAMLSS and multivariate copula models, also referred to as copula GAMLSS. The advantage of embedding copula regression into GAMLSS is that each parameter of the marginals and the copula association parameter can be modeled to depend flexibly on covariates. This allows us to measure the strength of the dependence, which has been the focus of previous literature on interrelated poverty dimensions, and to analyse which factors related to household location and composition drive this dependence. This latter aspect has not been previously considered in poverty studies.

2 Model definition

A bivariate cumulative distribution function can be written as

$$F_{1,2}(r, y_2) = \mathcal{C}(F_1(r), F_2(y_2)) \in [0, 1], \tag{1}$$

where in our case Y_1 is a categorical variable with categories r. The variable Y_2 is assumed to be continuous. In the case study of Section 3, response Y_2 will represent the income and Y_1 the highest level of education attained by each individual surveyed. The copula function is $\mathcal{C} : [0, 1]^2 \longrightarrow [0, 1]$, with $F_1(r) := \mathbb{P}(Y_1 \preceq r)$ and $F_2(y_2) := \mathbb{P}(Y_2 \leq y_2)$ being the marginal distributions. To ensure that the copula function is uniquely determined, we represent the ordinal variable Y_1 as a coarse version of a latent continuous variable and define a cumulative link model

$$\mathbb{P}(Y_1 \preceq r) = \mathbb{P}(Y_1^* \leq \theta_r) = \mathbb{P}(\epsilon_1 \leq \theta_r - \mathbf{x}_1'\boldsymbol{\beta}_1) := F_1^*(\underbrace{\theta_r - \mathbf{x}_1^\top \boldsymbol{\beta}_1}_{:=\eta_{1r}}), \tag{2}$$

where Y_1^* denotes an unobserved (or latent) continuous variable that drives the decision for the observed categories, θ_r is a cut point on the latent continuum related to the level r of Y_1. We observe category r if the latent variable is between the cutoffs θ_{r-1} and θ_r. The predictor η_{1r} is associated with the ordinal categorical response. Later the predictor η_{1r} will be replaced with a generalized additive form. Equation (1) can be written as

$$F_{12}(r, y_2) = F_{12}^*(\eta_{1r}, y_2) = \mathcal{C}(F_1(r), F_2(y_2)) = \mathcal{C}(F_1^*(\eta_{1r}), F_2(y_2)). \quad (3)$$

The bivariate copula model is embedded into the distributional regression framework (or GAMLSS) to model flexibly both the dependence parameter and the marginal distributions. To this end, the response vector $\boldsymbol{y}_i = (y_{1i}^*, y_{2i})'$, $i = 1, \ldots, n$, is assumed to follow a parametric distribution where potentially all parameters, except of the cut-points, are related to a regression predictor and consequently to covariates. We write the joint conditional density as $f_{12}^*(\vartheta_{1i}, \ldots, \vartheta_{Ki}|\boldsymbol{\nu}_i)$, where the vector $\boldsymbol{\nu}_i$ collects any covariates associated to the parameters $\vartheta_{ki}, k = 1, \ldots, K$ of density f_{12}^*. Accordingly, the distributional parameter vector $\boldsymbol{\vartheta}_i = (\theta_1^*, \ldots, \theta_R^*, \vartheta_{1i}, \ldots, \vartheta_{Ki})'$ includes transformed cut-points $\{\theta_r^*\}$, the location parameter of the first marginal distribution, all other distributional parameters related to the second marginal distribution, and the copula parameter γ_i. Subscript i attached to parameters is made explicit to stress their potential dependence on individual-level covariates. For the ordinal response, logit and probit link functions can be applied and the scale parameter for density f_1 is set to one. in order to achieve identification as for a probit/logit model.
In the spirit of the GAMLSS approach, each distributional element in the parameter vector is related to an additive predictor via

$$\vartheta_{ki} = h_k(\eta_i^{\vartheta_k}) \quad \text{and} \quad \eta_i^{\vartheta_k} = g_k(\vartheta_{ki}), \quad (4)$$

where $\eta_i^{\vartheta_k}$ is the predictor belonging to distributional parameter ϑ_{ki}, and $h_k = g_k^{-1}$ is a response function mapping the real line into the domain of ϑ_{ki}.
For the ordinal equation, η_{1ri} in equation (2) can now be represented as $\eta_{ri}^{\mu_1} = \theta_r - \eta_i^{\mu_1}$, where $\eta_i^{\mu_1}$ is a predictor as in (4). The predictor $\eta_i^{\vartheta_k}$ takes on the additive form

$$\eta_i^{\vartheta_k} = \sum_{j=1}^{J_k} s_j^{\vartheta_k}(\boldsymbol{\nu}_i),$$

where functions $s_j^{\vartheta_k}(\boldsymbol{\nu}_i), j = 1, \ldots, J_k$, can be chosen to model a range of different effects of (a subset) of explanatory variables $\boldsymbol{\nu}_i$, such as linear, spatial, random, or nonlinear effects. Estimation is performed using a trust region algorithm as in Marra and Radice (2017).

3 Multidimensional poverty in Indonesia

To analyse the poverty dimensions in a bivariate copula model with a focus on a household's location, we rely on the most recent wave (IFLS 5) of the Indonesian Family Life Survey (IFLS). We fit a lognormal distribution for the continuous marginal as indicated by QQ plots, a logit model for the first marginal and a Gaussian copula to connect both marginals as supported by AIC and BIC. The predictors of the models $\eta_{educ}^{\mu_1}, \eta_{inc}^{\mu_2}, \eta^{\sigma_2}, \eta^{\gamma}$ are related to several household covariates and a spatial effect on the province level.

Model evaluation: To check the final bivariate model, we use a multivariate generalization of the quantile residuals that was proposed by Kalliovirta (2008). Multivariate quantile residuals for two continuous responses are defined as

$$\hat{\boldsymbol{q}}_i = \begin{pmatrix} \hat{q}_{1i} \\ \hat{q}_{2i} \end{pmatrix} = \begin{pmatrix} \Phi^{-1}(\hat{F}_1(y_{1i})) \\ \Phi^{-1}(\hat{F}_{2|1}(y_{2i}|y_{1i})), \end{pmatrix}$$

where $\hat{F}_{2|1}$ is the (estimated) conditional CDF of Y_2 given Y_1. In our case, the first marginal is discrete such that we resort to randomized quantile residuals, where uniformly distributed random variables on the interval corresponding to cumulative probabilities are plugged into $\Phi^{-1}(\cdot)$. If the model is correctly specified, then $\hat{\boldsymbol{q}}$ approximately follows a bivariate standard normal distribution. The contour plot for the bivariate model in Figure 1 (left) shows the density of the quantile residuals $\hat{\boldsymbol{q}}$ by means of a multivariate kernel density estimator. This estimated density is compared to the density of the standard normal distribution. The contour lines of both densities are close to each other indicating a good fit of the bivariate copula model. In Figure 1 (right), the sum of the squared elements of the multivariate quantile residuals are considered. That is, $\hat{\boldsymbol{q}}_i'\hat{\boldsymbol{q}}_i = \hat{q}_{1i}^2 + \hat{q}_{2i}^2$, where $\hat{\boldsymbol{q}}_i$ is the multivariate quantile residual for the i-th individual and $\hat{\boldsymbol{q}}_i'\hat{\boldsymbol{q}}_i \overset{a}{\sim} \chi^2(2)$ which is assessed in the QQ-plot.

Dependence for urban and rural households: To compare the dependence structure across different locations (urban/rural), we create an example of typical individual whose characteristics, other than the one under consideration, are set to their mean value or to their most frequent observation. The only exception is the education of the household head which is set to the second most frequent observation. This is the covariates' combination that we call an "example individual" henceforward. Figure 2 shows that the dependence is stronger for individuals in urban households compared to rural households. One reason might be that average education levels are lower in rural areas (x-axis) while at the same time high paid job opportunities are restricted in a rural environment, resulting in more equal incomes compared to an urban environment.

Dependence structure across provinces: Figure 3 shows the average

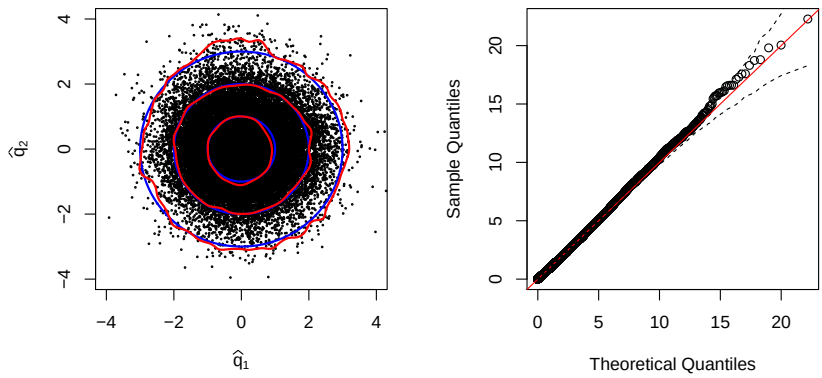


FIGURE 1. Left: Contour plot of multivariate quantile residuals. The red lines indicate the density of the quantile residuals estimated by a multivariate kernel density estimator. The blue circles are the contour lines of the density of the standard normal distribution with radius $1, 2$ and 3. Right: QQ-plot depicting the sum of the squared elements of the multivariate quantile residuals with 95% reference bands.

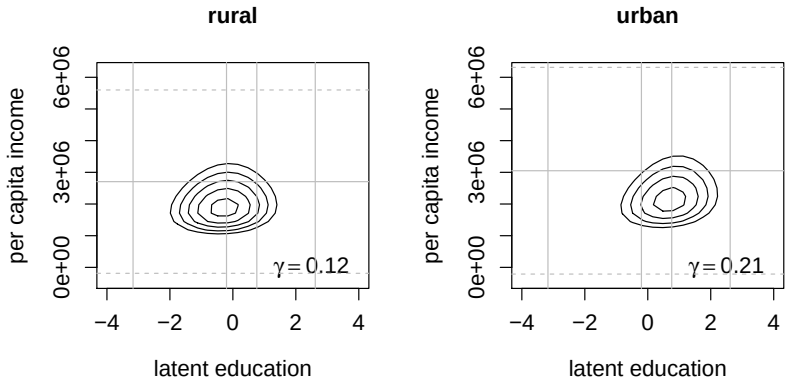


FIGURE 2. Contour plots for an example individual in an urban or rural household in the province of Jawa Timur. Contour plots for (education, income)' and a Gaussian copula. Contour lines of densities are at levels from 0.00000005 to 0.00000025 . The vertical straight lines represents the cut off values for the education categories, horizontal straight lines are the consumption average, and dashed horizontal line are at two standard deviations around this average.

of Kendall's τ over all individuals across in a particular province. The Kalimantan Selatan (South Borneo) is the province with the lowest average of Kendall's τ with a value of 0.045 and Kepulauan Riau (Riau Islands, northwest of Borneo) has a value of 0.150 which is the highest average value that also indicates spatial heterogeneity in the strength of the dependence. The provinces of Sumatra seem to have higher dependence between income and education than provinces in Borneo or Sulawesi. Interestingly, for Java and its neighbouring smaller islands on the east, the dependence seem to decrease from west to east.

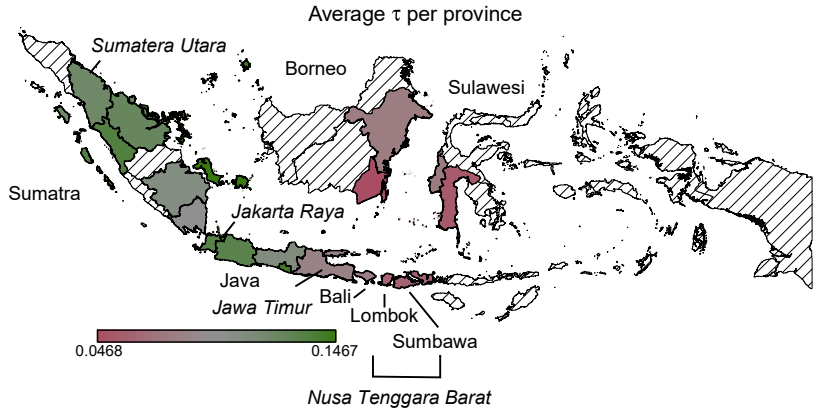


FIGURE 3. Kendall's τ for each individual averaged within provinces.

Joint probabilities: We calculate the probability for the example individual of being poor in both the education and income dimensions. We find that the probability for being poor in both dimensions is 2 times higher for the example individual in a rural household compared to the same individual in an urban household. Compared to Jawa Timur, the joint probabilities of Jakarta Raya, Nusa Tenggara Barat, and Sumatera Utara are about 8 times, 6 times, and 4 times higher, respectively.

References

Kalliovirta, L. (2008). Quantile residuals for multivariate models, *Technical Report 247*, Helsinki Center of Economic Research.

Marra, G. and Radice, R. (2017). Bivariate copula additive models for location, scale and shape *Computational Statistics & Data Analysis*, **112**, 99 – 113.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized Additive Models for Location, Scale and Shape *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507 – 554.

Closed-loop effects in coupling cardiac physiological models to clinical interventions

Dirk Husmeier[1] and L. Mihaela Paun[1]

[1] School of Mathematics & Statistics, University of Glasgow, Scotland, UK

E-mail for correspondence: dirk.husmeier@glasgow.ac.uk

Abstract: There have been impressive methodological advancements in the mathematical modelling of cardio-physiological processes. The majority of recent articles have focused on the *forward problem*: developing flexible mathematical models and robust numerical simulation procedures to match characteristics of physiological target data, and the *inverse problem*: inferring model parameters from cardiac physiological data with reliable uncertainty quantification. However, when connecting mathematical model predictions to the clinical decision process, new challenges arise. This paper briefly discusses the complications that potentially result from *closed-loop* effects, and the model extensions that are required to reduce the ensuing bias.

Keywords: Closed-loop effect, physiological model, pulmonary hypertension

1 Introduction and illustration

Consider a random variable $X \in \mathbb{R}$ that represents the value of a clinical disease indicator. Based on some adequate clinical data, which for the purpose of the following discussion do not need to be made specific, we monitor its posterior distribution $p(x)$ and the risk of the clinical indicator exceeding some tolerance threshold

$$P(X > \tau) \; = \; \int_{\tau}^{\infty} p(x)dx \tag{1}$$

If this risk exceeds some critical value α, $P(X > \tau) \; > \; \alpha$, medical treatment, for instance in the form of medication, is provided. While potentially only aiming at a symptomatic relief, this treatment is assumed to interfere with the patient's physiology or pathophysiology and affect the clinical disease indicator. Let $Y \in \mathbb{Y}$ denote a random variable that represents the

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

value of the disease indicator upon medical intervention, and let f describe the effect of the treatment: $y = f(x)$. This treatment effect implies a transformation of the probability distribution of the disease indicator:

$$p_y(y) = \int_{-\infty}^{\infty} \delta\Big(y - f(x)\Big) p_x(x) dx \tag{2}$$

where $\delta(.)$ is the Dirac delta function. The consequence is a potential prediction bias:

$$P(Y > \tau) \; = \; \int_{\tau}^{\infty} p(y) dy \; \neq \; P(X > \tau) \; = \; \alpha \tag{3}$$

which needs to be accounted for in the clinical decision process. As a simple illustration, assume the posterior distribution of the clinical indicator prior to the medical intervention is normal, $p_x(x) = N(x|\mu, \sigma^2)$, with mean $\mu = 0$ and variance $\sigma^2 = 1$, and that the effect of the medical treatment is a shift of the clinical indicator by $\psi \in \mathbb{R}^+$:

$$f(x) = \left[\begin{array}{ll} x & \text{if } x \leq \tau \\ x - \psi & \text{if } x > \tau \end{array} \right. \tag{4}$$

We obtain $p(y)$ by inserting (4) into (2) and making use of the following feature of the Dirac delta function:

$$\delta\Big(y - f(x)\Big) \; = \; \sum_i \frac{1}{|f'(x_i)|} \delta(x - x_i) \tag{5}$$

where $\{x_i\}$ are the roots of $y - f(x) = 0$. Inserting (4) and (5) into (2) gives:

$$p_y(y) \; = \; \left[\begin{array}{ll} p_x(y) & \text{if } y < \tau - \psi \\ p_x(y) + p_x(y + \psi) & \text{if } \tau - \psi \leq y \leq \tau \\ p_x(y + \psi) & \text{if } y > \tau \end{array} \right. \tag{6}$$

The apparent probability of the disease indicator to exceed the critical threshold τ will therefore be evaluated as

$$P(Y > \tau) = \int_{\tau}^{\infty} p_y(y) dy = \int_{\tau}^{\infty} p_x(y+\psi) dy = \int_{\tau+\psi}^{\infty} N(y|0,1) dy = \overline{G}(\tau+\psi)$$

where $\overline{G} = 1 - G$ and $G(.)$ is the normal cumulative distribution function, whereas the actual probability is

$$P(X > \tau) \; = \; \int_{\tau}^{\infty} p_x(x) dy \; = \; \int_{\tau}^{\infty} N(y|0,1) dy \; = \; \overline{G}(\tau) \tag{7}$$

Since $\overline{G}(.)$ is strictly monotonously decreasing, $\overline{G}(\tau + \psi) < \overline{G}(\tau)$, the apparent probability is biased and systematically underestimates the risk of exceeding the critical threshold τ: $P(Y > \tau) < P(X > \tau)$. Hence, by ignoring the effect of the treatment on the clinical indicator variable, any clinical decision support system based on this indicator variable will systematically underestimate the patient's state of risk.

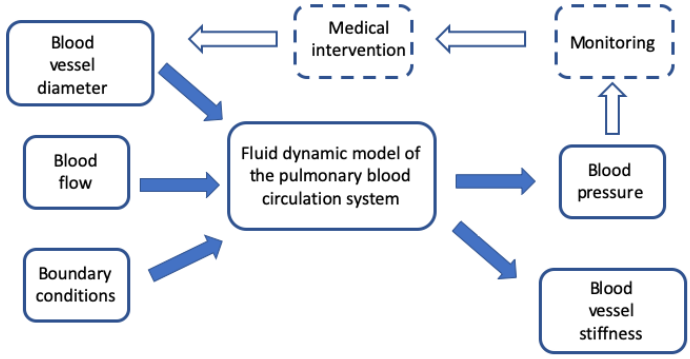


FIGURE 1. Schematic representation of our physiological model of pulmonary hypertension and how it is affected by closed-loop effects following a clinical intervention. See the main text for details.

2 Physiological application: pulmonary hypertension

Pulmonary hypertension, i.e. high blood pressure in the lungs, is a major risk factor for a variety of medical conditions, including inadequate coronary perfusion, stroke and heart failure. Pulmonary blood pressure can differ substantially from blood pressure in the rest of the body (the so-called systemic circuit) and, as opposed to the latter, can only be measured invasively. Standard techniques, which are based on right-heart catheterization, can have significant side effects, including internal bleeding and partial collapse of the lungs.

Recent advances in physiological modelling allow the pulmonary blood pressure to be predicted from the vasculature geometry and blood flow times series (Qureshi et al.), which can be measured non-invasively with computed tomography (CT) and ultrasound, respectively. The biophysical model depends on various boundary conditions and physiological parameters, most notably the blood vessel stiffness, which can be estimated with computational inference procedures (Paun et al.).

Figure 1 provides a schematic illustration. Given the geometry of the vasculature, most notably the blood vessel diameters (measured with CT), the blood flow (measured with ultrasound) and various boundary conditions (obtained from statistical inference, see Qureshi et al.), the model allows the prediction of the pulmonary blood pressure and the blood vessel stiffness (with the statistical inference techniques described in Paun et al.). In a clinical application, the prediction of high pulmonary blood pressure above a critical threshold will trigger the administration of vasodilators, whose effect is the increase of the vessel diameter. However, as illustrated

TABLE 1. Closed-loop effect and its correction in the biophysical modelling of pulmonary hypertension. Systolic blood flow can be measured with ultrasound; the initial geometry of the vasculature, including the diastolic diameter of the main pulmonary artery (MPA), is available from an initial CT scan. The biophysical model allows the prediction of the pulmonary systolic blood pressure (column 1) and the vessel stiffness (columns 2-3) with the statistical inference procedure described in (Paun et al.). The table shows the relative blood vessel stiffness estimation error (median and 95% posterior credible interval) without (column 2) and with a correction for the closed-loop effect that results from medical interventions triggered by model predictions (column 3). Since this is a simulation study (Qureshi et al.), the true vessel stiffness is known. Computational Bayesian inference was carried out with the MCMC scheme described in Paun et al.

Peak blood pressure exceeding threshold	Relative error without closed-loop correction	Relative error with closed-loop correction
25 %	1.51% (1.04%,1.97%)	-2.0e-03% (-0.37%,0.37%)
50 %	2.52% (2.15%,2.90%)	-1.4e-03% (-0.47%,0.46%)
75 %	61.5% (60.5%,62.4%)	-0.26% (-7.78%,6.49%)

in Figure 1, this causes a closed-loop effect, whereby the prediction from the model causes an action that alters the conditions under which the original prediction was obtained.

3 Simulation study

Our simulations are based on the pulmonary circulation model described in Qureshi et al. The blood vessel geometry of the larger blood vessels has been obtained from a CT scan in a healthy mouse, the effect of the small terminal blood vessels is approximated with electronic circuit (so-called Windkessel) elements consisting of two resistances and a capacitance. This gives three parameters that define downstream boundary conditions of the partial differential equations (PDEs) describing the blood flow through the pulmonary circuit. We also assume that the blood flow at the main pulmonary artery (MPA) is measured (noninvasively with ultrasound), which provides the upstream boundary condition for the PDEs. Following Qureshi et al. and Paun et al., we assume the same stiffness parameter in all blood vessels, which adds one further parameter to the physiological model. We further assume that the blood flows in the two daughter vessels of the MPA are measured (with ultrasound). Our data used for inference are the time courses of the blood flows through three blood vessels. The parameters to be inferred are the vessel stiffness and three Windkessel parameters. Once these parameters have been estimated, the blood pressure in the MPA can be predicted. A graphical illustration is provided in Figure 2 .

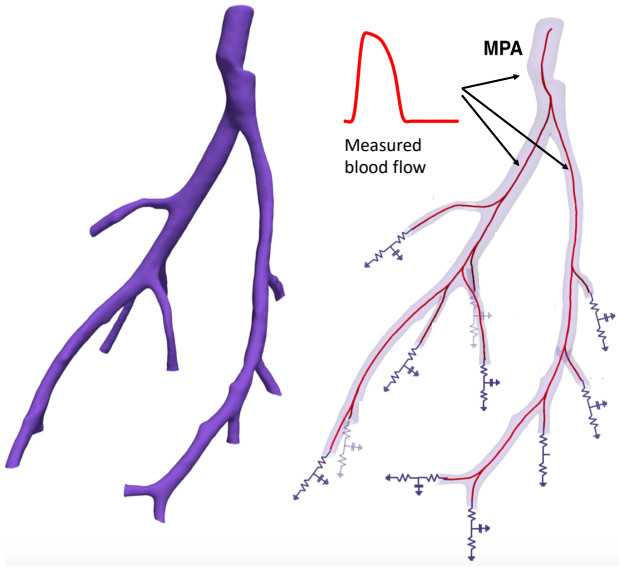


FIGURE 2. *Left panel:* 3D smoothed segmented network from a micro-CT image of a healthy mouse lung. *Right panel:* Directional graph of the same network. Blood flow waveforms are taken from ultrasound measurements at the main pulmonary artery (MPA) and two daughter vessels. At the outlet of each terminal vessel, three-element Windkessel elements with two resistors and a capacitor are attached as boundary conditions to mimic the effect of the microvasculature further downstream.

To simulate the effect of clinical interventions, we monitor the blood pressure in the MPA, and provide an in-silico vasodilator whenever the pressure exceeds a critical threshold. Since this is a proof-of-concept study, we use data from mice rather than humans, and set as an arbitrary threshold the peak pressure found in the hypoxic control mice used in the study of Qureshi et al. We simulate the effect of the vasodilator by increasing the diastolic trough diameter of all blood vessels by the same percentage amount, whose value is determined by the requirement that upon medical intervention, the peak blood pressure in the MPA must not exceed the critical value by more than 5%. This bandwidth defines the uncertainty that remains when explicitly including the closed-loop effect caused by the medical intervention in the model. We compare that with naive parameter inference that does not include any correction for the medical intervention, and assumes the diastolic blood vessel diameter to be fixed. We quantify the effect of ignoring the feedback loop with the percentage estimation error of the vessel stiffness.

4 Results

The results can be found in Table 1. They demonstrate that ignoring the closed-loop effect leads to a systematic bias in the estimation of the blood vessel stiffness, which is a critical risk indicator for vessel wall rupture, stroke and right-ventricle heart failure (Chen et al). Allowing for the medical intervention and including the ensuing feedback loop in the statistical inference corrects this bias and leads to a substantially improved estimation of the stiffness parameter.

5 Conclusions

Quantitative physiological models have great potential for improved and automated clinical decision support. However, it is important to correct for closed-loop effects in model calibration. Using a mathematical toy problem and a realistic fluid dynamics simulation of the pulmonary blood circulation system, we have shown that failing to allow for the effect of medical interventions – and not explicitly including them in the model – can lead to a systematic prediction bias. Our future work will focus on improved statistical inference when data on the effect of medical interventions are noisy and/or partially missing.

Acknowledgements

This work has been funded by EPSRC, grant reference numbers EP/R018634/1 (Centre for Closed Loop Data Science), EP/S030875/1 and EP/N014642/1 (Centre for Multiscale Soft Tissue Mechanics), and the Royal Society of Edinburgh, grant reference number 62335. We would like to thank Mette Olufsen, Mitchel Colebank and Nicholas Hill for helpful discussions about the clinical procedures and mathematical models for the pulmonary blood circulation system.

References

Chen et al. (2017). Arterial stiffness and stroke: de-stiffening strategy, a therapeutic target for stroke. *Stroke and Vascular Neurology*, **2** (2).

Qureshi et al. (2018). Hemodynamic assessment of pulmonary hypertension in mice: a model-based analysis of the disease mechanism. *Biomechanics and Modeling in Mechanobiology*, **18**, 219–243.

Paun et al. (2018). MCMC methods for inference in a mathematical model of pulmonary circulation. *Statistical Neerlandica*, **72**, 306 – 338.

Genome-Wide Association Studies: a Distance-Based approach

Itziar Irigoien[1], Bru Cormand[2], Maria Soler[3], Cristina Sánchez-Mora[3], Josep Antoni Ramos-Quiroga[3], Concepción Arenas[2]

[1] University of the Basque Country UPV/EHU, Donostia, Spain
[2] University of Barcelona, Barcelona, Spain
[3] Hospital Universitari Vall d'Hebron, Barcelona, Spain

E-mail for correspondence: itziar.irigoien@ehu.eus

Abstract: With the increase in the production of genome-wide association studies (GWAS), the analysis of such data sets with thousands of potential predictive single nucleotide-polymorphisms (SNPs) has become crucial in biomedical research. Here we propose a new method to identify SNPs related with a disease in case-control studies. The method provides two ordered lists of SNPs (with causal or protective alleles) that provide a useful tool to help the researcher to decide where to focus attention in a first stage.

Keywords: Attention-deficit/hyperactivity disorder; Distances; Genome-wide association studies; Nearest Neighbors.

1 Introduction

A typical GWAS data set may contain thousands of potential predictive single nucleotide-polymorphisms (SNPs) and the aim is to identify genes involved in human disease, by searching for SNP variants that occur more frequently in people with a particular phenotype than in people without this phenotype. GWAS analysis typically assume an underlying genetic model of association for each SNP (e.g., dominant, recessive, or additive), being the single additive model the one typically selected. In this case, each SNP is represented as the corresponding number of minor alleles (0, 1, or 2). Many methods for SNP identification use univariate tests which involve regressing each SNP separately on a given trait, adjusted for possible covariate variables and assessing the significance after correction for multiple

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

comparisons with a lost of sensitivity. As analyzing SNPs one at a time can neglect information about the joint distribution, multi-association analysis may be more suitable (Balding, 2006). Other possibilities involve grouping of SNPs over a moving window and look for associations of groups with the diseases, but the selection of the window is very subjective (Wu et al., 2010) or to consider stochastic search algorithms (Dobra and Massam, 2010). In the context of GWAS data the presence of population substructure can result in spurious associations. Usually, the first ten principal components (PC) are considered as covariate variables, assuming that these PCs capture information of the latent population substructure (Price et al., 2010).

2 Method

Let Y be a categorical variable indicating the presence (encoded as 1) or absence (encoded as 0) of the disease of interest. Let $\mathbf{X} = (x_{ij}^y)$ be an $n \times p$ data matrix containing the genotypes for the j^{th} SNP ($j = 1, \ldots, p$) on the i^{th} ($i = 1, \ldots, n$) individual, with $n = n_1 + n_2$ (n_1 is the number of cases and n_2 the number of controls), being y equal to 1 or 0 for cases and controls, respectively. We consider the single additive model as the underlying genetic model of association. In this case, each SNP tested in the case-control study and with alleles A and a generates three genotypes (AA, Aa, aa) and is represented as the corresponding number of minor alleles (0, 1, or 2). The model assumes that a SNP will be related to the disease if the number of values equal to 1 or 2 is greater in the case group than in the control group; that is, having one or two copies of the a allele will increase the probability of presenting the disease. Let $\mathbf{D} = (d_{ij})$ be the Manhattan distance matrix between all the individuals. For each individual $\mathbf{x}_i^y = (x_{i1}^y, \ldots, x_{ip}^y)^{\mathrm{T}}$ in the case or control group ($i = 1, \ldots, n$), we consider its 10-nearest neighbors among the n_1 cases ($NN_1(\mathbf{x}_i^y) = \{\mathbf{x}_{i_1}^1, \ldots, \mathbf{x}_{i_{10}}^1\}$) or among the n_2 controls ($NN_0(\mathbf{x}_i^y) = \{\mathbf{x}_{i_1}^0, \ldots, \mathbf{x}_{i_{10}}^0\}$), based on the $\mathbf{D}$ distance matrix. The method associates each SNP j with a value i_1^j obtained from variable I_1^j where

$$I_1^j = \frac{1}{10n_1} \sum_{i=1}^{n_1} \sum_{k=1}^{10} B(p_{ik}^j) - \frac{1}{10n_2} \sum_{i=1}^{n_2} \sum_{k=1}^{10} B(q_{ik}^j),$$

with $B(p_{ik}^j)$ a Bernoulli distribution taking value 1 with probability p_{ik}^j if the i case takes values 1 or 2 and its k neighbor control takes value 0 on the jth SNP; otherwise, it takes the value 0 with probability $1 - p_{ik}^j$. $B(q_{ik}^j)$ follows a Bernoulli distribution taking value 1 with probability q_{ik}^j if the i control takes values 1 or 2, and its k neighbor control takes value 0 on the jth SNP; otherwise, it takes the value 0 with probability $1 - q_{ik}^j$.

Proposition: Consider case i and its $NN_0(\mathbf{x}_i^1)$ neighbours. Let p_i be the probability of observing values 1 or 2 in SNP j for case i and let w_j be the probability that the jth SNP is related with the disease. Then,

$$p_{ik}^j = w_j p_i(1-p) + (1-w_j)Q^j \text{ and } q_{ik}^j = w_j p(1-p) + (1-w_j)Q^j,$$

with p the probability of observing values 1 or 2 by chance, and Q^j the probability of the event {case/control i takes values 1 or 2 and its k neighbor control takes value 0 | SNP j is not related with the disease}.

Proposition: SNPs that favor the presence of the disease have positive and large I_1^j values.

The decreasing ordered list with the i_1^j values provides a tool for a genetic study to focus the attention on SNPs with potentially casual alleles. As the distribution followed by I_1 with the i_1^j values is unknown in order to determine a threshold for the SNPs selection, it is necessary to obtain the associated bootstrap distribution or to adjust to a convenient Normal distribution if possible. Similarly, the method associates each SNP j with a value I_2^j with

$$I_2^j = \frac{1}{10n_2}\sum_{i=1}^{n_2}\sum_{k=1}^{10} B(p_{ik}^j) - \frac{1}{10n_1}\sum_{i=1}^{n_1}\sum_{k=1}^{10} B(q_{ik}^j),$$

where now $B(p_{ik}^j)$ follows a Bernoulli distribution taking value 1 with probability p_{ik}^j if the i control takes values 1 or 2 and its k neighbor case takes value 0 on the jth SNP; otherwise, it takes the value 0 with probability $1 - p_{ik}^j$. $B(q_{ik}^j)$ follows a Bernoulli distribution taking value 1 with probability q_{ik}^j if the i case takes values 1 or 2, and its k neighbor case takes value 0 on the jth SNP; otherwise, it takes the value 0 with probability $1 - q_{ik}^j$. In a similar way, SNPs with a corresponding i_2^j value that is big and positive are those potentially conferring protection against the disease.

3 Simulate data

The simulated case-control data set *simuCC* included in the genMOSS R package, contains 6000 SNPs, 1000 cases and 1000 controls. Two SNPs, rs4491689 and rs6869003, and a random environmental factor were associated with the disease. Our method identified these two SNPs as the first and second SNPs in the ranked list of SNPs favoring the disease, in agreement with the fact that they are the disease predisposing SNPs.

4 Application to a real data set

A real case-control data set, previously used in attention-deficit/hyperactivity disorder (ADHD; Snchez-Mora et al., 2015), including 418 cases, 428 controls and 155802 SNPs covering the whole genome was analyzed. We split the sample 20 times at random into train (90%) and test (10%) data. Taking SNPs that favor the presence of ADHD allows a highly reliable assignation of cases and controls, reaching correct classification percentages over 90% with only 200 SNPs (Table 1). The top finding is SNP rs739465 in the *VAV2* gene, encoding an angiogenic protein and previously associated with multiple sclerosis. Other findings point at the *NF1* gene, encoding neurofibromin 1 and causal for neurofibromatosis but also associated with risk-taking behavior, alcohol consumption or anxiety, and at *RBFOX1*, encoding a splicing factor and found associated with depression and also highlighted in a recent GWAS meta-analysis of 8 psychiatric disorders, including ADHD.

TABLE 1. AUC and percentage of correct classification of subjects into the case (ADHD) or control groups in the train-test situation, under different α values that correspond to different numbers of SNPs. Mean and standard deviation (in brackets) are indicated.

α	Number SNPs	AUC	Correct classification (%)
0.0001	22.00 (3.21)	0.78 (0.02)	71.87 (1.91)
0.0005	100.63 (7.91)	0.89 (0.01)	85.53 (1.47)
0.001	192.37 (11.16)	0.94 (0.01)	91.33 (0.91)
0.0025	457.65 (14.65)	0.98 (0.01)	97.10 (0.48)
0.005	894.26 (19.12)	0.98 (0.01)	99.02 (0.26)
0.01	1740.21 (21.15)	0.99 (0.003)	99.78 (0.15)
0.025	4220.26 (43.70)	0.99 (0.001)	99.92 (0.07)

References

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**, 781 – 791.

Dobra, A. and Massam, H. (2010). The Mode Oriented Stochastic Search (MOSS) Algorithm for Log-linear Models with Conjugate Priors. *Statistical Methodology*, **7**, 240 – 253.

Price, A.L., Zaitlen, N.A., Reich, D. and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, **11**, 459 – 463.

Snchez-Mora, C., Ramos-Quiroga, J.A., Bosch, R. et al., (2015). Case-control genome-wide association study of persistent attention-deficit hyperactivity disorder identifies FBXO33 as a novel susceptibility gene for the disorder. *Neuropsychopharmacology*,**40**, 915 – 26.

Wu, M.C., Kraft, P., Epstein, M.P. et al., (2010). Powerful SNP-set Analysis for Case-control Genome-wide Association Studies. *The American Journal of Human Genetics*, **86**, 929 – 942.

Multivariate Conditional Transformation Models

Thomas Kneib[1], Nadja Klein[2], Torsten Hothorn[3]

[1] Georg-August-Universität Göttingen, Germany
[2] Humboldt-Universität zu Berlin, Germany
[3] Universität Zürich, Switzerland

E-mail for correspondence: tkneib@uni-goettingen.de

Abstract: Regression models describing the joint distribution of multivariate response variables conditional on covariate information have become an important aspect of contemporary regression analysis. However, a limitation of such models is that they often rely on rather simplistic assumptions, e.g. a constant dependence structure that is not allowed to vary with the covariates or the restriction to linear dependence between the responses only. We propose a general framework for multivariate conditional transformation models that overcomes such limitations and describes the full joint distribution in a tractable and interpretable yet flexible way. Among the particular merits of the framework are that it can be embedded into likelihood-based inference (including results on asymptotic normality) and allows the dependence structure to vary with the covariates. In addition, the framework scales well beyond bivariate response situations.

Keywords: copulas; multivariate regression; most likely transformations; seemingly unrelated regression.

1 Basic Model Setup

We start by discussing transformation models developed for the analysis of the joint multivariate distribution of a J-dimensional, absolutely continuous random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_J)^\top \in \mathbb{R}^J$ with density $f_{\boldsymbol{Y}}(\boldsymbol{y})$ without conditioning on covariates. The key component of multivariate transformation models then is an unknown, bijective, strictly monotonically increasing transformation function $h : \mathbb{R}^J \to \mathbb{R}^J$. This function maps the vector $\boldsymbol{Y}$, whose distribution is unknown and shall be estimated from data, to a set of J independent and identically distributed, absolutely continuous random

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

variables $Z_j \sim \mathbb{P}_Z, j = 1, \ldots, J$ with an a priori defined distribution $\mathbb{P}_Z$ (in the following the standard normal distribution), such that

$$h(\boldsymbol{Y}) = (h_1(\boldsymbol{Y}), \ldots, h_J(\boldsymbol{Y}))^\top = (Z_1, \ldots, Z_J)^\top = \boldsymbol{Z} \in \mathbb{R}^J.$$

The density of $\boldsymbol{Y}$ implied by the transformation model is then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \left[\prod_{j=1}^{J} \phi_{0,1}(h_j(\boldsymbol{y}))\right] \cdot \left|\frac{\partial h(\boldsymbol{y})}{\partial \boldsymbol{y}}\right|$$

and, upon a suitable parameterisation of the transformation function, this enables maximum likelihood inference. However, in this generality, the model is cumbersome in terms of both interpretation and tractability. Thus, in the following, we introduce simplified parameterisations of h that lead to interpretable models.

2 Models with Recursive Structure

In a first step, we impose a triangular structure on the transformation function h by assuming

$$h_j(\boldsymbol{y}) = h_j(y_1, \ldots, y_J) = h_j(y_1, \ldots, y_j)$$

i.e. the jth component of the transformation function depends only on the first j elements of its argument $\boldsymbol{y}$. In a second step, we assume that the triangulary structured transformation functions are linear combinations of marginal transformation functions $\tilde{h}_j : \mathbb{R} \to \mathbb{R}$, i.e.

$$h_j(Y_1, \ldots, Y_j) = \lambda_{j1}\tilde{h}_1(Y_1) + \ldots + \lambda_{jj}\tilde{h}_j(Y_j)$$

where each $\tilde{h}_j$ increases strictly monotonically and $\lambda_{jj} > 0$ for all $j = 1, \ldots, J$ to ensure the bijectivity of h. Because the last coefficient, λ_{jj}, cannot be separated from the marginal transformation function $\tilde{h}_j(Y_j)$, we use the restriction $\lambda_{jj} \equiv 1$. Thus, the parameterisation of the transformation function h finally reads

$$h_j(Y_1, \ldots, Y_j) = \lambda_{j1}\tilde{h}_1(Y_1) + \ldots + \lambda_{j,j-1}\tilde{h}_{j-1}(Y_{j-1}) + \tilde{h}_j(Y_j)$$

and the model-based density function for $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{j=1}^{J} \phi_{0,1}\left(\lambda_{j1}\tilde{h}_1(Y_1) + \ldots + \lambda_{j,j-1}\tilde{h}_{j-1}(Y_{j-1}) + \tilde{h}_j(Y_j)\right) \frac{\partial \tilde{h}_j(Y_j)}{\partial Y_j}.$$

Summarising the model's specifications, our multivariate transformation model is characterised by a set of marginal transformations $\tilde{h}_j(Y_j)$, $j =$

$1, \ldots, J$, each applying to only a single component of the vector $\boldsymbol{Y}$, and by a lower triangular $(J \times J)$ matrix of transformation coefficients

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & & & & 0 \\ \lambda_{21} & 1 & & & \\ \lambda_{31} & \lambda_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ \lambda_{J1} & \lambda_{J2} & \ldots & \lambda_{J,J-1} & 1 \end{pmatrix}.$$

Under the standard normal reference distribution $\mathbb{P}_Z = \mathrm{N}(0, 1)$, the coefficients in $\boldsymbol{\Lambda}$ characterise the dependence structure via a Gaussian copula while the marginal transformation functions $\tilde{h}_j$ allow the generation of arbitrary marginal distributions for the components of $\boldsymbol{Y}$.

3 Conditional Transformation Models

By extending the unconditional transformation function, we define the J components of a multivariate conditional transformation function given covariates $\boldsymbol{X}$ as

$$h_j(\boldsymbol{y} \mid \boldsymbol{X}) = \sum_{j=1}^{j-1} \lambda_{jj}(\boldsymbol{X}) \tilde{h}_j(Y_j \mid \boldsymbol{X}) + \tilde{h}_j(Y_j \mid \boldsymbol{X})$$

where $\lambda_{jj}(\boldsymbol{X})$ and $\tilde{h}_j(Y_j \mid \boldsymbol{X})$ are expressed in terms of suitable basis function expansions, e.g. based on Bernstein polynomials that facilitate the consideration of the monotonicity constraints. For the marginal (with respect to the response Y_j) conditional (given covariates $\boldsymbol{X}$) transformation functions, this leads to a parameterisation

$$\tilde{h}_j(Y_j \mid \boldsymbol{X}) = \boldsymbol{c}_j(Y_j, \boldsymbol{x})^\top \boldsymbol{\vartheta}_j$$

where the basis functions $\boldsymbol{c}_j(Y_j, \boldsymbol{x})$, in general, depend on both element Y_j of the response and the covariates $\boldsymbol{x}$.

4 Simulation Study

In this section, we provide empirical evidence on the performance of our MCTMs via simulations. We simulated $R = 100$ data sets of size $n = 1,000$, following a method similar to that used in the parametric bootstrap procedure:

1. Covariate values x were simulated as i.i.d. variables, where $x \sim U[-0.9, 0.9]$.

2. The latent variables $\tilde{\boldsymbol{z}}_{ir} \in \mathbb{R}^2$ were generated as

$$\tilde{\boldsymbol{z}}_{ir} = \boldsymbol{\Lambda}_i^{-1} \boldsymbol{z}_{ir}, \quad i = 1, \ldots, n;\ r = 1, \ldots, R$$

with

$$\boldsymbol{z}_{ir} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_2) \text{ and } \boldsymbol{\Lambda}_i = \begin{pmatrix} 1 & 0 \\ x_{ir}^2 & 1 \end{pmatrix},$$

such that

$$\mathrm{Cov}(\tilde{z}_{i1}, \tilde{z}_{i2} | x_i) \equiv \boldsymbol{\Sigma}_i(x_i) = \begin{pmatrix} 1 & -x_i^2 \\ -x_i^2 & 1 + x_i^4 \end{pmatrix}.$$

3. From the latent variables, the observed responses were computed as

$$\boldsymbol{y}_{ir} = [F_1^{-1}\{\Phi_{0,1}(\tilde{z}_{ir,1})\}, F_2^{-1}\{\Phi_{0,\sigma_{i2}^2}(\tilde{z}_{ir,2})\}]^\top,$$

where $\sigma_{i2}^2 = 1 + x_i^4$ and F_1 and F_2 are the CDFs of two Dagum distributions with parameters $a_1 = \exp(2), b_1 = \exp(1), p_1 = \exp(1.3)$ and $a_2 = \exp(1.8), b_2 = \exp(0), p_2 = \exp(0.9)$, respectively. Note that the CDF of an unconditional Dagum distribution reads as

$$F(y) = \left(1 + \left(\frac{y}{b}\right)^{-a}\right)^{-p}, \quad \text{for } y > 0\ a > 0, b > 0, p > 0.$$

This model specification is equivalent to a Gaussian copula model with Dagum marginals, but note that, by its construction, the first margin is independent of the covariate x, while the scale of the second margin varies as a function of x. More precisely, the scale parameter b is affected by x while the shape parameters a, p remain constant.

As competitors for MCTMs, we considered Bayesian structured additive distributional regression models as implemented in the software package BayesX and vector generalised additive models as implemented in the corresponding R add-on package VGAM. For VGAM and BayesX, we employed the true specification, i.e. a Gaussian copula with correlation parameter

$$\rho(x_i) = \frac{-\lambda(x_i)}{\sqrt{1 + \lambda(x_i)^2}}$$

and Dagum marginals, in which the paramater b_2 of the second marginal depended on x but the first marginal as well as the parameters a_2 and p_2 did not. Both the predictor for b_2 and the correlation parameter ρ of the Gaussian copula were specified using cubic B-splines with 20 inner knots on an equidistant grid in the range of x with a second-order random walk prior; the other parameters of the margins were estimated as constants.

Because VGAM does not allow for simultaneous estimation of the margins and the dependence structure, we first estimated the Dagum marginals with

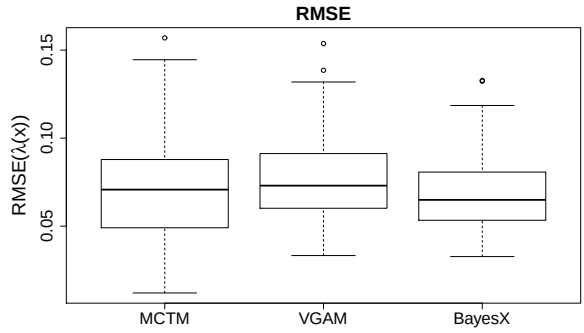


FIGURE 1. RMSE for $\hat{\lambda}(x)$ from MCTM (left), VGAM (middle) and BayesX (right).

constant parameters a_1, b_1, p_1, a_2, p_2 and covariate-dependent parameters b_2. The copula predictor was then estimated with plugged-in estimates of the margins, using cubic B-splines with 18 inner knots. For the multivariate transformation models (denoted as MCTM), we employed Bernstein polynomials of order eight (in y) and order three (in x), as in our application. Although BayesX and VGAM employed the correct model specification in terms of the parametric distribution assumption for the marginal distributions and the correlation parameter, the performance of MCTM was highly competitive in terms of the RMSE (Figure 1) without the requirement to either estimate the marginal distributions in a first step and plug the empirical copula data in to obtain the dependence structure (as for VGAM) or specifying parametric marginal distributions (as for BayesX). Both requirements are restrictive in practice since typically it is impossible to pick the 'correct' parametric distribution that exactly matches the marginal distributions of the underlying random variables.

5 Application: Trivariate Conditional Transformation Models for Undernutrition in India

To demonstrate practical aspects of multivariate conditional transformation models, we present a trivariate analysis of undernutrition in India. Overall, the available data set comprised 24,316 observations, after preprocessing of the data. We used three indicators, *stunting*, *wasting* and *underweight*, as the trivariate response vector, where *stunting* refers to stunted growth, measured as an insufficient height of a child with respect to his or her age, while *wasting* and *underweight* refer to insufficient weight for height and insufficient weight for age respectively. Hence *stunting* is an indicator of chronic undernutrition, *wasting* reflects acute undernutrition and *underweight* reflects both. Our aim is to model the joint distribution

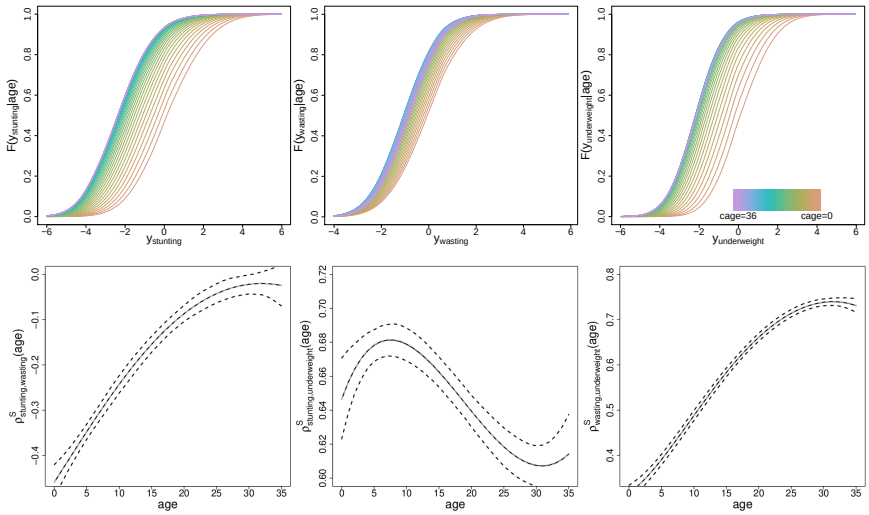


FIGURE 2. Estimated CDFs (top row) and rank correlation coefficients (bottom row) given age of the child. For the rank correlation coefficient, the maximum likelihood estimates are shown as solid black line and the 95% bootstrap confidence intervals as dashed black lines.

of *stunting*, *wasting* and *underweight* conditional on the age of the child. Figure 2 (top row) depicts the estimated marginal conditional CDFs $F_j(Y_j \mid \text{age})$, with the different colours indicating the ages of the children. Clearly, the shapes of the margins differ for the three indicators and change with the increasing age of the children. A shift to the left in the margins representing older ages indicates a higher risk of lower undernutrition scores. All of the distributions, but especially those of wasting, are asymmetric, as with increasing age the lower tails can be seen to vary less strongly than the upper tails.

Figure 2 (bottom row) depicts the conditional rank correlations ρ^S between stunting, wasting and underweight as functions of age along with the point and interval estimates obtained from $1,000$ parametrically drawn bootstrap samples. Interestingly, the correlation between stunting and wasting is initially negative for young children and then approaches zero with the increasing age of the children.

References

Klein, N., Hothorn, T., and Kneib, T. (2019). Multivariate Conditional Transformation Models. Working Paper, arXiv:1906.03151v2.

Percentiles curves based on multivariate conditional transformation models. Application to diabetes

Oscar Lado-Baleato[1], Carmen Cadarso-Suárez[1], Thomas Kneib[2] and Francisco Gude[3]

[1] Department of Statistics, Mathematical Analysis, and Optimization, Universidade de Santiago de Compostela, Galicia, Spain.
[2] Chair of Statistics, Georg-August-Universitt Gttingen, Gttingen, Germany.
[3] Clinical Epidemiology Unit, Complexo Hospitalario Universitario de Santiago de Compostela, Galicia, Spain.

E-mail for correspondence: oscarlado.baleato@usc.es

Abstract: Multivariate Conditional Transformation Models (MCTMs) were recently proposed as a new multivariate regression technique. MCTMs characterize jointly the covariates effects on the marginal distributions of the responses and their correlations. Flexibility, in both the responses and covariates effects are achieved using Bernstein polynomial basis. Based on MCTMs, in this paper percentile curves are constructed for each response. Simulation studies indicated the good performance of these estimated condtional percentiles. Finally, MCTMs percentile curves were obtained for three diabetes markers (fasting plasma glucose, glycated hemoglobin and fructosamine) condtionally on age.

Keywords: Diabetes; Multivariate regression ; Berstein basis; Multivariate transformation models.

1 Introduction

Diabetes diagnosis and control are mainly based on two tests: fasting plasma glucose (FPG) and glycated hemoglobin (HbA1c) concentrations. However, conditions that determine alterations in hemoglobin metabolism (anemia or kidney disease) can interfere with the reliability of HbA1c measurements. On the other hand, FPG is highly dependent on food ingestion and sample storage. Fructosamine (Fr), another glycated protein, is frequently used as an alternative glycemic marker. Nevertheless, its transla-

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

tion into average glucose levels is not as clear as for HbA1c and discordances are often encountered between Fr and HbA1c results. In addition, agreement among these glycemic markers may be affected by common factors like the age of individuals.
Studying simultaneously these three glycemic markers concentrations depending on age, may improve the diagnosis and treatment of diabetes. In a previous work (Espasandín-Domínguez et al (2019)) bivariate copula regression models were applied to identify factors that might affect the HbA1c and Fr distributions and explain discordant results between both tests. In this paper the FPG was included as an additional response variable thus extending this work to the trivariate case. To this aim, Multivariate Conditional Transformation Models (MCTMs; Klein et al, 2020) were considered. These models characterize the covariates effects on the Cumulative Distribution Functions (CDF) of each response and on their correlations. Also, MCTMs allow us to obtain percentile curves for each biomarker facilitating the interpretability to the practitioners.
The rest of the paper is structured as follows, the structure of MCTMs is briefly explained in Section 2 and the percentile curves are obtained. In Section 3 a simulation study is carried out to evaluate the percentile curves performance. The results of our clinical study are presented in Section 4 and finally the paper ends with some conclusions.

2 MCTM percentile curves

The Multivariate Conditional Transforamtion Models (MCTM) are based on the transformation of the original variable into a reference distribution (usually N(0,1)) applying an unknown, bijective and strictly monotonically increasing transformation function. In the trivariate case, given $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$, the transformation function $h : \mathbb{R}^3 \to \mathbb{R}^3$ maps the vector $\boldsymbol{Y}$ to a set of independent and identically distributed random variables. That is:

$$h(\boldsymbol{Y}) = (h_1(Y_1)), h_2(Y_2), h_3(Y_3))^T = (Z_1, Z_2, Z_3)^T = \boldsymbol{Z} \in R^3$$

Finally, the variable $\boldsymbol{Y}$ dependence structure is characterised by a lower triangular (3×3) matrix

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ \lambda_{21} & 1 & 0 \\ \lambda_{31} & \lambda_{32} & 1 \end{bmatrix}$$

defined by the coefficients $\lambda_{21}, \lambda_{31}$ and λ_{32} measuring the correlation between $(Y_1, Y_2), (Y_1, Y_3)$ and (Y_2, Y_3) respectively.
In the conditional case, given the covariates vector $\boldsymbol{x}$ the multivariate transformation function is given by $\tilde{h}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{j=1}^{j-1} \lambda_{jj}(\boldsymbol{x})\tilde{h}_j(y_j|\boldsymbol{x}) + \tilde{h}_j(y_j|\boldsymbol{x})$.

Where $\tilde{h}_j(y_j|\boldsymbol{x})$ and $\lambda_{ij}(\boldsymbol{x})$ are expressed in terms of basis functions expansions as:

$$\tilde{h}_j(y_j|\boldsymbol{x}) = c_j(y_j, \boldsymbol{x})^T \boldsymbol{\vartheta}_j = \boldsymbol{a}_j(y_j)^T \boldsymbol{\vartheta}_{j,1} - \boldsymbol{b}(\boldsymbol{x})^T \boldsymbol{\vartheta}_{j,2}$$
$$\lambda_{jj} = \alpha_{jj} + \boldsymbol{b}(\boldsymbol{x})^T \boldsymbol{\gamma}_{jj}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are a polynomial Bernstein basis and $(\alpha_{jj}, \boldsymbol{\vartheta}_{ji}, \boldsymbol{\gamma}_{jj})$ parametric coefficients whose estimation and inference is based on the model log-likelihood. The inference of some quantities, as the responses correlations, is achieved using parametric bootstrap (see Klein et al (2020) for details).
In MCTM the conditional CDF for each response $\boldsymbol{P}(\boldsymbol{Y} \leq \boldsymbol{y}|\boldsymbol{x})$ is given by $F(Y_j|\boldsymbol{X} = \boldsymbol{x}) = \Phi_{0,\sigma_j^2}(\tilde{h}_j(y_j|\boldsymbol{x})) = F_Z(\boldsymbol{a}_j(y_j)^T \boldsymbol{\vartheta}_{j,1} - \boldsymbol{b}(\boldsymbol{x})^T \boldsymbol{\vartheta}_{j,2})$. In order to make this model more interpretable the following percentile curves $Q_{\boldsymbol{Y}}(\tau) = F^{-1}(Y_j|\boldsymbol{X} = \boldsymbol{x})$ with $\tau \in (0, 1)$ may be obtained.

3 Simulation study

In this section we evaluate the percentile curves estimation. In the simulation set-up three continuous outcomes following a Dagum distribution were considered. The Dagum scale parameters were made dependent on one single covariate $x \in [-1, 1]$ as $b_1 = x^3$, $b_2 = x$ and $b_3 = x^2$. The trivariate response dependence structure was made dependent on x, using the correlation matrix given in Klein et al 2020 (section 5). The evaluation was done in 250 replicates considering three sample sizes $(500, 1000, 5000)$ and three τ values (0.05, 0.50 and 0.95).

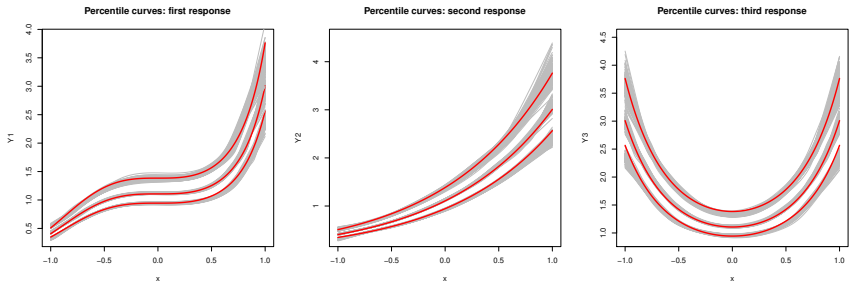


FIGURE 1. Estimation of the percentile curves for $\tau = 0.05, 0.50, 0.95$ and $n = 500$.

As can be seen in Figure 1, the estimated percentile curves (in grey) are very close to the theoretical ones (in red) for each marginal and τ. As shown in Table 1, based on the root mean square error, the percentile curves estimation error decreases with sample size for the three response variables being higher for $\tau = 0.05$ and $\tau = 0.95$.

TABLE 1. Root mean square error (RMSE) for the percentile curves estimation. The RMSE was obtained in 200 equally spaced points of the percentile curves as $\frac{1}{250}\sum_{i=1}^{250}\frac{1}{200}\sum_{k=1}^{200}\sqrt{(\hat{y}_{ik}^{\tau}-y_k^{\tau})^2}$.

	$Y_1 : x^3$			$Y_2 : x$			$Y_3 : x^2$		
	$\tau = 0.05$	$\tau = 0.50$	$\tau = 0.95$	$\tau = 0.05$	$\tau = 0.50$	$\tau = 0.95$	$\tau = 0.05$	$\tau = 0.50$	$\tau = 0.95$
500	0.0019	0.0013	0.0045	0.0028	0.0010	0.0037	0.0042	0.0014	0.0046
2000	0.0018	0.0007	0.0028	0.0024	0.0006	0.0015	0.0036	0.0007	0.0020
5000	0.0018	0.0007	0.0025	0.0022	0.0006	0.0011	0.0035	0.0005	0.0015

4 Multivariate regression modelling of glycemic markers

Using a sample of 1516 adults collected in the A-Estrada Glycation and Inflamation study (see Espasandn et al, 2019 for details) a MCTM trivariate regression model for the FPG, HbA1c and Fr concentrations depending on age was fitted. The marginal conditional transformation functions were parametrised as

$$\tilde{h}(y_j|age) = \boldsymbol{a}_j(y_j)^T\boldsymbol{\vartheta}_{j,1} - \boldsymbol{b}(age)\boldsymbol{\vartheta}_{j,2} \text{ for } j \in (FPG, HbA1c, Fr)$$

and the responses dependence structure as

$$\lambda_{jj}(age) = \boldsymbol{b}(age)^T\boldsymbol{\gamma}_{ij}$$

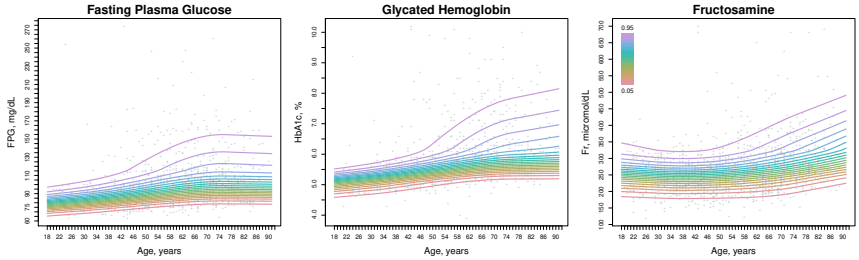


FIGURE 2. Percentile curves for FPG (fasting plasma glucose), glycated hemoglobin (HbA1c) and fructosamine (Fr) concentrations.

Figure 2 depicts the percentile curves for each glycemic markers with different colour indicating several τ value. The FPG, HbA1c and Fr concentrations increase with age and this increase is more pronounced for the upper percentile. Finally, as can be seen in Figure 3, the Fr and FPG association, as well as the Fr and HbA1c one, is higher in the older patients. While FPG and HbA1c show the higher association degree but it is not depend on age.

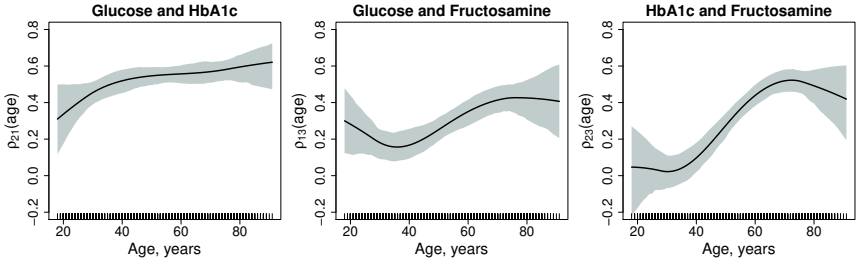


FIGURE 3. Age effect on the responses correlations along with the 95% pointwise confidence interval using 1000 bootstrap replicates.

5 Discussion and future work

In this work we demonstrate the usefulness of MCTMs in a real biomedical problem in diabetes research. These models allow for a joint estimation of the covariates effects on the distribution of the responses and their correlations. In this work, MCTMs percentile curves were proposed in order to get a more interpretable model output. MCTMs allowed us to model jointly, for the first time, the age effect on the concentrations of three glycemic markers, offering a better understanding these diabetes markers measurements.

References

Espasandín-Domínguez, J., Cadarso-Suárez, C., Kneib, T., Marra, G., Klein, N., Radice, R., Lado-Baleato, O., González-Quintela, A. and Gude, F. (2019). Assessing the relationship between markers of glycemic control through flexible copula regression models *Statistics in Medicine*, **38**, 5161 – 5181.

Klein N., Hothorn T. and Kneib T. (2020) Multivariate Conditional Transformation Models *Arxiv preprint*, doi:https://arxiv.org/abs/1906.03151.

Multivariate distributional regression forests for probabilistic nowcasting of wind profiles

Moritz N. Lang[1,2], Georg J. Mayr[2], Lisa Schlosser[1], Thorsten Simon[1,2], Reto Stauffer[1,3], Achim Zeileis[1]

[1] Department of Statistics, Universität Innsbruck, Innsbruck, Austria
[2] Department of Atmospheric and Cryospheric Science, Universität Innsbruck, Innsbruck, Austria
[3] Digital Science Center, Universität Innsbruck, Innsbruck, Austria

E-mail for correspondence: Moritz.Lang@uibk.ac.at

Abstract: This study presents statistical methods to probabilistically predict wind profiles along the approach path of an airport for one hour in advance. Accurate nowcasts of wind profiles increase safety and facilitate optimal air traffic management by timely re-routing of landing aircraft when wind direction shifts. Distributional regression trees and forests are enhanced to predict vertical wind profiles employing a multivariate normal distribution. To gain probabilistic forecasts for both wind speed and wind direction, the components of the two-dimensional Cartesian wind vector are modeled simultaneously for several height levels of a measurement tower. The resulting tree-based models can capture non-linear effects and interactions, and automatically select the relevant covariates that are associated with changes in any of the parameters of the (possibly) high-dimensional multivariate normal distribution employed. Extending the multivariate distributional regression trees to multivariate distributional regression forests can further improve the predictive performance by regularizing and smoothing the covariate effects.

Keywords: Distributional Trees; Random Forest; Multivariate Normal Distribution; Wind Profiles; Probabilistic Forecasting

1 Motivation

Statistical forecasting of numerical weather quantities has so far focused mainly on near-surface variables such as temperature, wind, and precipitation, presumably because most people are directly affected there. Accordingly, distributional regression trees and forests have already been success-

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

fully applied for probabilistic rain and wind direction forecasting by accounting for appropriate univariate response distributions (Schlosser et al., 2019; Lang et al., 2020). The nowcasting task of providing vertical wind profiles for aviation forecasters and air traffic control serves as a practical real-case application to extend univariate distributional regression trees and forest to multivariate response distributions.

2 Multivariate trees and forests

Distributional regression trees (Schlosser et al., 2019) fuse distributional modeling with regression trees based on the unbiased recursive partitioning algorithms MOB (Zeileis et al., 2008) or CTree (Hothorn et al., 2006). The basic idea is to recursively partition the covariate space into (approximately) homogeneous subgroups, so that a single distributional model is sufficient to be fitted to the response in each resulting subgroup. To capture the dependence on covariates, the association between the model's scores and each available covariate is assessed using either a parameter instability test (MOB) or a permutation test (CTree). After selecting the covariate with the highest significant association as split variable (i.e., lowest significant p-value, if any), the corresponding split point is chosen within the selected covariate either by optimizing the log-likelihood (MOB) or by using a two-sample test statistic (CTree) over all possible partitions. A natural extension of (distributional) regression trees is to build ensembles or forests of such trees which can further improve the predictive performance by regularizing and stabilizing the model (Breiman, 2001).

In comparison to preceding studies using distributional regression trees and forests, this study employs distributional trees and forests for probabilistic forecasting of a multivariate response distribution. Drawing on related work for tree models of psychometric networks (Jones et al., 2019), a p-dimensional multivariate normal distribution is employed in the leaves of the trees; however, the introduced methodology is conceptually transferable to any multivariate distribution. Based on the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, the density for a single p-dimensional observation vector $\mathbf{y}_i$ is given by

$$f_{\text{MVN}}(\mathbf{y}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right).$$

In the subsequent notation we collect all parameters in a single parameter vector $\boldsymbol{\theta}$ of length $k = p + p + p(p-1)/2$. Thus, this comprises the p means from $\boldsymbol{\mu}$ and the p variances and $p(p-1)/2$ correlations, respectively, from which the covariance $\boldsymbol{\Sigma}$ can be constructed.

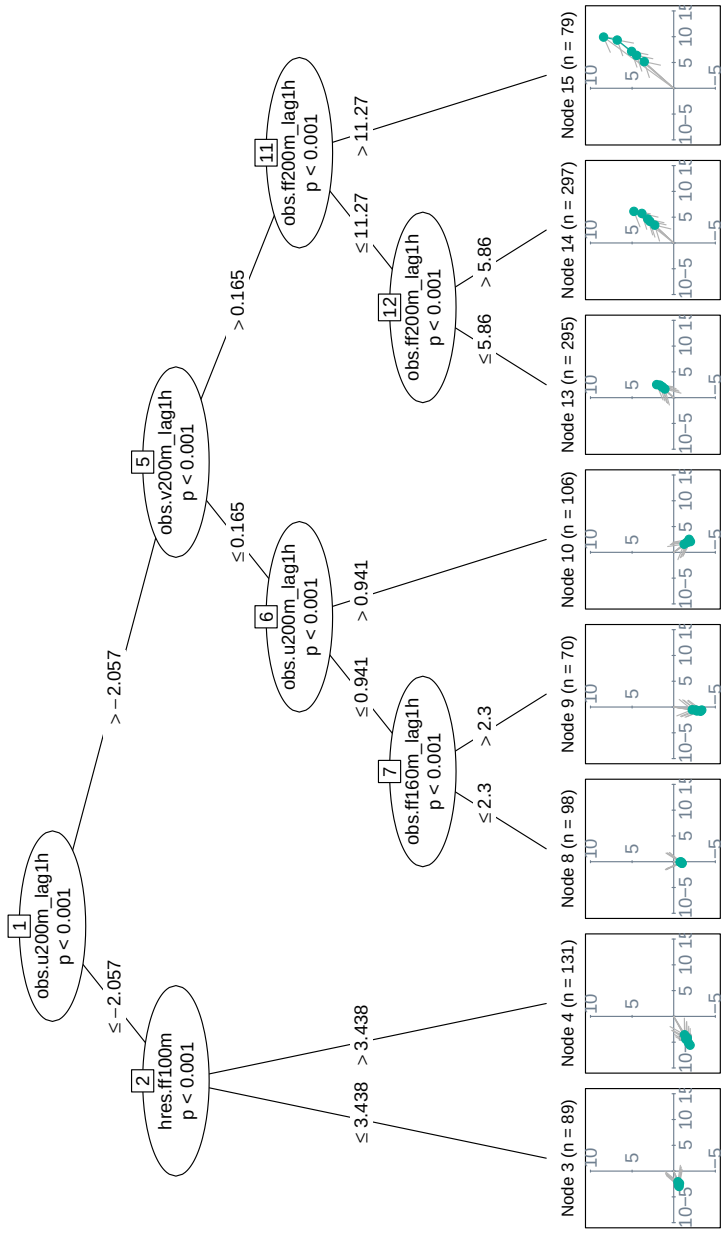


FIGURE 1. Fitted distributional regression tree based on the multivariate normal distribution for the u and v wind components at five different height levels for a measurement tower at Karlsruhe. In each terminal node, the location parameters of the wind vectors at all height levels are shown as colored points and gray arrows. The unit of the Cartesian coordinate system is in meter per second. The covariates employed are numerical high-resolution forecasts (hres), as well as 1-hourly lagged observations (obs) for wind speed (ff) and both wind vector components (u or v), all reported at different height levels (160 m, 100 m, 200 m).

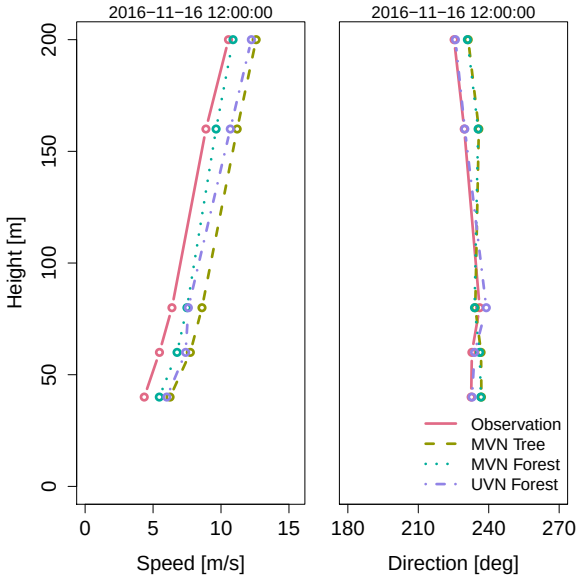


FIGURE 2. Derived wind direction and wind speed prediction at different height levels for a multivariate distributional tree and forest, as well as for an univariate distributional tree estimated per wind component and height level separately.

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is obtained by maximizing the sum of the corresponding log-likelihood contributions $\ell(\boldsymbol{\theta}; \mathbf{y}_i) = \log(f_{\text{MVN}}(\mathbf{y}_i; \boldsymbol{\theta}))$ based on the n observations in a given sample. The corresponding scores $s(\boldsymbol{\theta}, \mathbf{y}_i) = (\partial_{\theta_1} \ell(\boldsymbol{\theta}; \mathbf{y}_i), \ldots, \partial_{\theta_k} \ell(\boldsymbol{\theta}; \mathbf{y}_i))$ can be employed as a general goodness-of-fit measure. Hence, evaluating the scores at the individual observations and parameter estimates $s(\hat{\boldsymbol{\theta}}, \mathbf{y}_i)$ yields an $n \times k$ matrix that assesses how well each distribution parameter estimate $\hat{\boldsymbol{\theta}}$ fits one individual observation vector $\mathbf{y}_i$. If the scores change systematically along available covariates, the parameter instabilities are incorporated into the model by maximizing a partitioned likelihood. This procedure is repeated recursively until there are no significant parameter instabilities or until another stopping criterion is met (e.g., subgroup size or tree depth).

3 Nowcasting of wind profiles

To study the performance of the novel multivariate trees and forests, 1 h predictions of vertical wind profiles for 12 UTC are issued for a measuring tower in Karlsruhe. The response has $p = 10$ dimensions, consisting of zonal (u) and meridional (v) wind components at five different height levels (40 m, 60 m, 80 m, 160 m, 200 m). Numerical weather forecasts and

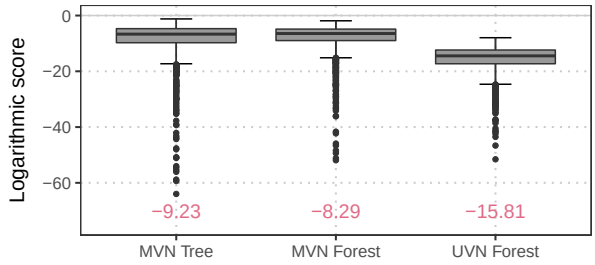


FIGURE 3. Out-of-sample predictive performance in terms of the logarithmic score based on the full predictive multivariate normal distribution for 1 h forecasts of the wind speed components at five different height levels of the measurement tower. In addition, the averaged performance over all dates is shown in red.

1-hourly lagged observations of various meteorological wind quantities are used as splitting variables, as well as derived quantities such as temporal means, minima, and maxima, or temporal and spatial differences. The terminal nodes of the tree in Fig. 1 depict the location parameters for the wind vectors at the different heights. Splits in the lagged observed u and v components (at 200 m) broadly distinguish four different regimes of wind directions: south-west (nodes 3, 4), south (nodes 8, 9), south-east (node 10), and north east (nodes 13, 14, 15). Within each regime splits in either lagged observed or predicted wind speed (ff), distinguishes low vs. high wind speeds in the same (or rather similar) directions.

To validate the estimated scale and correlation of the multivariate trees, these are compared to multivariate forests, as well as to a univariate distributional regression forest, employing the normal distribution, estimated for each wind component and height level separately. In the latter no correlation is assumed between the wind components at a single height level and between different levels. For a characteristic sample case, all three models capture the observed wind speed and direction comparably well (Fig. 2). The performance of the models, in terms of the logarithmic score, is assessed employing a yearly based four-fold cross-validation using daily data from 2014 to 2017 (Fig. 3). The box-and-whiskers show that the multivariate models outperform the univariate one, which seems to be too restrictive by the assumption of no correlation. Further, the multivariate forest is slightly superior to the multivariate tree by regularizing and smoothing the covariate effects.

The results show that the multivariate trees and forests are able to model all aspects contained in the univariate model, and further extend them by representing the correlation structure between the wind components at a single height level, as well as between the different levels. By fitting a single multivariate model for both wind components and all height levels, the

profile remains consistent and vertically coherent which allows to provide not only deterministic but rather probabilistic forecasts.

Computational details: The R package **disttree** implementing the proposed multivariate distributional regression trees and forests is available at `https://R-Forge.R-project.org/projects/partykit/`.

Acknowledgments: This project was partly funded by the Austrian Research Promotion Agency (FFG, grant no. 858537) and by the Austrian Science Fund (FWF, grant no. P31836).

References

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 1, 5 – 32.

Jones, P.J., Mair, P., Simon, T., and Zeileis, A. (2019). Network model trees. *OSF Preprints*, osf.io/ykq2a.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J. Comput. Graph. Stat.*, **15**, 3, 651 – 674.

Lang, M. N., Schlosser, L., Hothorn, T., Georg, J. M., Stauffer, R., and Zeileis, A. (2020). Circular Regression Trees and Forests with an Application to Probabilistic Wind Direction Forecasting. arXiv:2001.00412, *arXiv.org E-Print Archive*.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Ann. Appl. Stat.*, **13**, 1564 – 1589.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *J. Comput. Graph. Stat.*, **17**, 2, 492 – 514.

Multivariate Bayesian latent structure modeling of spatio-temporal health data

Andrew B. Lawson, Daniel R. Baer, Jane E. Joseph [1]

[1] Medical University of South Carolina,Charleston,USA

E-mail for correspondence: lawsonab@musc.edu

Abstract: Spatio-temporal health data is now routinely available. Often when time augments space, the focus is on modeling global spatio-temporal effects. However, temporal effects are often localized spatially and so it could be important to disaggregate these effects. This leads to spatial clustering of temporal effects. Often this disaggregation is approached via latent mixture component models. Extending this approach to multiple disease incidence is the focus of this presentation. The specific example that is explored, and motivates the detailed modeling, is incidence of mild cognitive impairment (MCI) and Alzheimers disease (AD). MCI is considered a pre-cursor of AD and so there is a temporal latent link between these outcomes. Our models address latent component mixtures for each disease but also coupled components shared between diseases. A case study in annual county level incidence in South Carolina is presented.

Keywords: Bayesian; Multivariate; spatio-temporal; Machine learning; AD-MCI modeling.

1 Background and Introduction

Alzheimer's disease (AD) is a serious neurological disorder with adverse effects on patient cognition and physical health. Moreover, compared to other leading causes of death, mortality related to AD has increased in recent years; from 2000 to 2015, AD has shown a 123% increase in mortality in the US. Methods which are able to elucidate the role of precursors or promoters on AD risk, and in particular the geographic variation in AD risk, will therefore be instrumental in characterizing AD risk at the aggregate patient level. To this end, we consider in this study the evaluation of novel Bayesian hierarchical models for disease mapping which will improve our ability to characterize AD risk. In particular, we consider an application

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

of our novel models to spatiotemporal AD count data arising from the counties of South Carolina, as well as to simulated data. Mild congnitive impairment (MCI) is regarded as a precursor of AD and its spatial incidence distribution can also be modeled. The link between MCI and AD may be important in assessing progression and establishing trends in risk in both diseases.

2 Disease Mapping models

In modeling counts of a disease collected over space and time, we can use a BHM, wherein we often assume a Poisson model for the observed disease counts y_{ij} at the first level of the model hierarchy:

$$y_{ij} \sim Pois(e_{ij}\theta_{ij}) \qquad i = 1, ..., n;\ j = 1, ..., J$$

where e_{ij} denotes the expected count of the disease and θ_{ij} the unknown relative risk parameter of the disease for the area at the j th measurement occasion.
We then model the unknown relative risk parameter, θ_{ij}, at the second level of the model hierarchy by decomposing the logarithm of the unknown relative risk parameter into the sum of spatial random effects, v_i, u_i, temporal random effects, γ_j, and space-time interaction terms, ψ_{ij} (Knorr-Held, 2000; Lawson, 2018). That is,

$$\log(\theta_{ij}) = \alpha_0 + v_i + u_i + \gamma_j + \psi_{ij}$$

where α_0 is an intercept term representing the baseline contribution to the log relative risk parameter over all areas and measurement occasions.

2.1 Space-time mixture (STM) models

The classic Knorr-held model (ST BHM) is limited in that it assumes 'global' temporal random effects, γ_j. That is, while the ST BHM provides estimates of the underlying smoothed spatial and temporal variation in disease risk, it does not account for the possibility that subsets of the areas may demonstrate homogenous temporal profiles in disease risk (Lawson et al , 2010; Napier et al , 2019). We therefore desire a model which allows for 'disaggregation' of these global random effects – thus allowing us to classify areas to descriptive latent temporal trends in disease risk.

2.2 Proposed Modeling Paradigm

We assume that both MCI and AD can be decomposed into temporal latent components with a spatial signature. Hence we have

$$\begin{cases} y_{ij}^{MCI} \sim Pois(e_{ij}^{MCI}\theta_{ij}^{MCI}) \\ y_{ij}^{AD} \sim Pois(e_{ij}^{AD}\theta_{ij}^{AD}) \end{cases}$$

and models for $\log(\theta_{ij}^{AD})$ and $\log(\theta_{ij}^{MCI})$. Different models can be assumed that link the diseases.
In general, the latent structure is decomposed as follows:

$$\log(\theta_{ij}) = \alpha + v + u + \sum_{l=1}^{L} \Lambda_{ijl} + \psi_{ij}$$

with $\sum_{l=1}^{L} \Lambda_{ijl} = \sum_{l=1}^{L} \omega_{il}\chi_{lj}$ where χ_{lj} is the l th stochastic latent trend and ω_{il} is a weight associating the trend to the i th area. Originally the components were assigned random walk prior distributions so that $\chi_{lj} \sim N(\chi_{l,j-1}, \tau_{\chi}^{-1})$. The weights can have a range of prior specifications, subject to $0 < \omega_{il} < 1$, and $\sum_{l=1}^{L} \omega_{il} = 1$. Commonly the weights are assumed to have a Dirichlet - singular multinomial prior distribution whereby $\underset{L\times 1}{\omega_i} \sim Mult\left(\mathbf{1}, \underset{L\times 1}{\mathbf{p}_i}\right)$ where $p_{il} = \frac{p_{il}^*}{\sum_{l=1}^{L} p_{il}^*}$ and $p_{il}^* \sim Gamma(1,1)$, but other alternatives exist (Lawson et al , 2010). Subsequently, user defined parametric latent trend components $f_l(j|\underset{q_l x 1}{\gamma_l})$, where q_l are a set of regression parameters, have been used to aid identification and minimise label switching (Napier et al, 2019).

AD-MCI modeling We utilise a non parametric non-linear measure of association (maximum information coefficient: MIC) (Reshef et al, 2011; Reshef et al 2016) to assess the degree of association between the AD and MCI risk within our models for the log risk. Hence,

$$\log(\theta_{ij}^{MCI}) = \alpha_0^{MCI} + \phi_i^{MCI} + \sum_{l=1}^{L} \Lambda_{ijl}^{MCI} + \psi_{ij}^{MCI}$$

$$\log(\theta_{ij}^{AD}) = \alpha_0^{AD} + \phi_i^{AD} + sign(\rho_i) I\{MIC_i > \delta\} A_{ijl}^{MCI} + [1 - I\{MIC_i > \delta\}] B_{ijl}^{AD} + \psi_{ij}^{AD}$$

where $A_{ijl}^{MCI} = \sum_{l=1}^{L_{MCI}} \Lambda_{ijl}^{MCI}$ and $B_{ijl}^{AD} = \sum_{l=1}^{L_{AD}} \Lambda_{ijl}^{AD}$. Also $\rho_i = corr(\underset{J\times 1}{\widehat{\theta}_i^{MCI}}, \underset{J\times 1}{\widehat{\theta}_i^{AD}})$, the Pearson correlation of the relative risks. This is used to provide a direction for the MIC. The $MIC_i = MIC(\underset{J\times 1}{\widehat{\theta}_i^{MCI}}, \underset{J\times 1}{\widehat{\theta}_i^{AD}})$, for a given region, is a time invariant non-parametric (potentially non-linear) measure of association. This formulation allows there to be inclusion of links to MCI components, for the AD risk, if the association is strong enough. Variants of these models have also been examined with different association patterns. Baer et al (2020) discuss these variants and also estimation issues (such as label switching, MCMC implementation, and identification). Nimble was used throughout for sampler construction and model fitting.

2.3 Results

Data from the SC Revenue and Fiscal Affairs (RFA) all-payers health data system is available for SC counties for the years 2007-2011. Both MCI incident counts and AD incident counts are available for ER visits. Figure 1 displays the SIRs for these data. Mapping of the MIC values displays the differential association depending on δ. The latent profiles found for AD and MCI are different, in that AD shows a cluster of increasing temporal risk and fluctuating MCI risk with lower temporal change. A variety of diagnostic plots based on the real data and simulations will be presented.

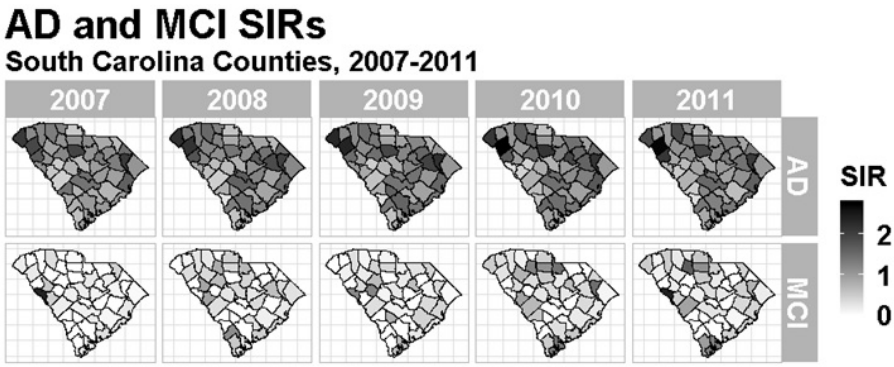


FIGURE 1. Standardised Incidence Ratios for AD and MCI for 2007-2011 for SC counties

Figure 2 displays the posterior mean temporal classes found for MCI and AD under the three different models with different trend assumptions. Model details can be found in Baer et al (2020).

Acknowledgments: Special Thanks to the NIH TL1 program at MUSC for supporting the research of DB.

References

Knorr-Held,L. (2000). Bayesian Modelling of Inseparable Space-Time Variation in Disease Risk. *Statistics in Medicine*, **19**, 2555 – 2567.

Lawson, A. B. (2018). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology.* New York: CRC Press 3rd Ed.

Lawson, A. B. et al (2010). Space-time latent component modeling of georeferenced health data. *Statistics in Medicine*, **29** (19), 2012 – 2027.

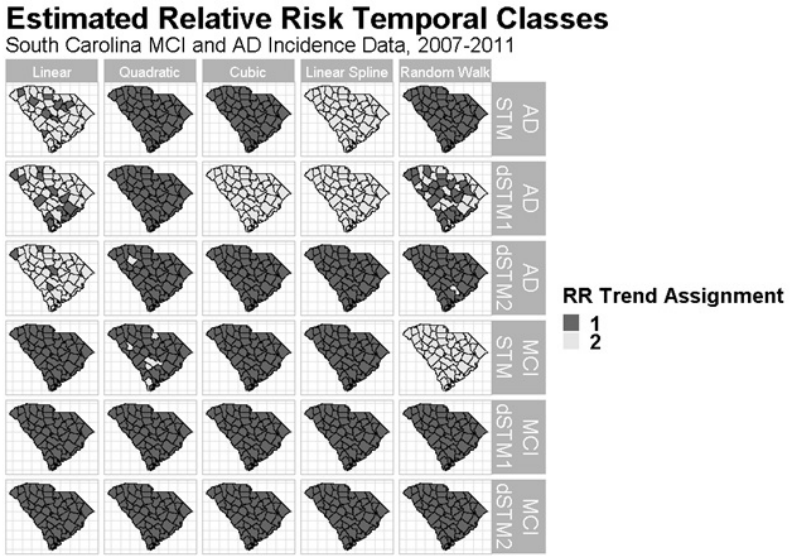


FIGURE 2. Temporal profile classes estimated under three different models for AD and MCI and different trend assumptions

Napier, G. et al (2019). A Bayesian space-time model for clustering areal units based on their disease trends. *Biostatistics.* **20** (4),681 – 697

Reshef, D.N., Reshef, Y. A. et al (2011) Detecting novel associations in large data sets, *Science.* **334** (6062), 1518 – 1524

Reshef, Y. A. et al (2016). Measuring dependence powerfully and equitably. *J Mach Learn Res.* **17** (212), 1 – 63.

Baer, D., Lawson, A. B. Joseph, J. (2020). Joint Space-Time Bayesian Disease Mapping via Quantification of Disease Risk Association. *Statistical Methods in Medical Research* (to appear).

Density-on-Scalar Regression Models with an Application in Gender Economics

Eva-Maria Maier[1], Almond Stöcker[1], Bernd Fitzenberger[21], Sonja Greven[1]

[1] Humboldt-Universität zu Berlin, Germany
[2] IAB (Institute for Employment Research), Nuremberg, Germany

E-mail for correspondence: eva-maria.maier@hu-berlin.de

Abstract: We provide a gradient boosting approach to estimate functional additive regression models with probability density functions as response variables and scalar covariates. To respect the special properties of densities, we formulate the regression model in a Bayes Hilbert space. This allows for a variety of applications, in particular for mixed densities, which have positive probability masses at some points of an interval. We illustrate how to handle this challenge by means of a motivating data set from the German Socio-Economic Panel Study (SOEP). In this application, we analyze the distribution of the woman's share in a couple's total labor income, which has positive probability masses at zero and one, using covariate effects for year, federal state, and age of the youngest child.

Keywords: Functional Regression; Gradient Boosting; Bayes Hilbert Spaces; Mixed Density Regression.

1 Introduction

We consider a special case of function-on-scalar regression (e.g., Brockhaus et al., 2015), namely density-on-scalar regression, where the responses are probability density functions and the covariates are scalar. Probability density functions have the special properties of being nonnegative and integrating to one, which are not preserved by the usual vector space structure of functions. Instead, we consider densities as elements of Bayes Hilbert spaces (Egozcue et al., 2006) and formulate our regression model using the respective operations. For estimation, we present a gradient boosting algorithm, which performs variable selection and allows for regularization, as well as for a large number of covariate effects, which can be modularly

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

and flexibly specified building on the model-based boosting framework of Bühlmann and Hothorn (2007). We use our approach to analyze the distribution of the woman's share in a couple's total labor income in Germany over the years and in dependence of different covariates. While these densities are defined on $[0, 1]$, they have positive probability mass at values 0 and 1, corresponding to one partner without labor income. This leads to a mixed reference measure, consisting of the Dirac measure at the boundary values and the Lebesgue measure in between.

An earlier approach by Talská et al. (2018) used Bayes Hilbert spaces for density-on-scalar regression, applying the centered log-ratio transformation to simplify estimation. They considered only linear regression models and only for densities defined on a finite interval and Lebesgue integrals. Other approaches to handle densities in regression often include different transformation approaches (e.g., Han et al., 2019), but only allow modeling and estimation on the transformed level without embedding the densities themselves in a vector space structure.

In Section 2, we provide a brief summary of Bayes Hilbert spaces. In Section 3, we present our approach for density-on-scalar regression. Section 4 contains the application of our methods for a mixed reference measure to the SOEP data.

2 Bayes Hilbert spaces

Bayes Hilbert spaces were first introduced for probability density functions defined on a finite interval by Egozcue et al. (2006), motivated by the approach of Aitchison (1986) for compositional data. We use the extension of Boogaart et al. (2014) to Bayes Hilbert spaces on finite measure spaces. Consider a measurable space $(\mathcal{T}, \mathcal{A})$ and a finite measure μ on it. Let $\mathcal{M}(\mu) = \mathcal{M}(\mathcal{T}, \mathcal{A}, \mu)$ be the set of measures with the same null sets as μ. We identify each measure $\nu \in \mathcal{M}(\mu)$ with its Radon-Nikodym derivative with respect to μ, denoted by f_ν. Proportionality defines an equivalence relation on the set $\{f_\nu \mid \nu \in \mathcal{M}(\mu)\}$. The corresponding set of equivalence classes is called *Bayes space (with reference measure μ)*, denoted by $\mathcal{B}(\mu) = \mathcal{B}(\mathcal{T}, \mathcal{A}, \mu)$. For the sake of readability we refrain from using squared brackets to denote the equivalence classes and simply write $f_\nu \in \mathcal{B}(\mu)$. If an equivalence class contains densities with finite integral, we choose the respective probability density as representative. Finally, the *Bayes Hilbert space (with reference measure μ)* is $B^2(\mu) = B^2(\mathcal{T}, \mathcal{A}, \mu) := \{f_\nu \in \mathcal{B}(\mu) \mid \int_\mathcal{T} (\log f_\nu)^2 \, \mathrm{d}\mu < \infty\}$. It is a vector space with addition $f_{\nu_1} \oplus f_{\nu_2} := f_{\nu_1} \cdot f_{\nu_2}$, $f_{\nu_1}, f_{\nu_2} \in B^2(\mu)$ and scalar multiplication $\alpha \odot f_\nu := (f_\nu)^\alpha$, $\alpha \in \mathbb{R}, f_\nu \in B^2(\mu)$. The additive neutral element $0_{B^2(\mu)}$ is the equivalence class of constant functions, the additive inverse element of $f_\nu \in B^2(\mu)$ is $\ominus f_\nu := \frac{1}{f_\nu}$, and the multiplicative neutral element is $1 \in \mathbb{R}$. Furthermore, $B^2(\mu)$ is a Hilbert

space with the inner product $\langle f_{\nu_1}, f_{\nu_2} \rangle_{B^2(\mu)} := \int_{\mathcal{T}} \operatorname{clr}[f_{\nu_1}] \cdot \operatorname{clr}[f_{\nu_2}] \, \mathrm{d}\mu$. Here, $\operatorname{clr}[f_\nu] := \log f_\nu - \frac{1}{\mu(\mathcal{T})} \cdot \int_{\mathcal{T}} \log f_\nu \, \mathrm{d}\mu$ is the *centered log-ratio (clr) transformation*. It is an isometric isomorphism, mapping functions from $B^2(\mu)$ to $L_0^2(\mu) = L_0^2(\mathcal{T}, \mathcal{A}, \mu) := \{\tilde{f} \in L^2(\mathcal{T}, \mathcal{A}, \mu) \mid \int_{\mathcal{T}} \tilde{f} \, \mathrm{d}\mu = 0\}$, which is a closed subspace of $L^2(\mu)$. The inner product induces a norm on $B^2(\mu)$ given by $\|f_\nu\|_{B^2(\mu)} := \sqrt{\langle f_\nu, f_\nu \rangle_{B^2(\mu)}}$.

3 Density-on-scalar regression

Let $B^2(\mu) = B^2(\mathcal{T}, \mathcal{A}, \mu)$ be a Bayes Hilbert space and $(y_i, \mathbf{x}_i) \in B^2(\mu) \times \mathbb{R}^K$, $K \in \mathbb{N}, i = 1, \ldots, N$, $N \in \mathbb{N}$, be data pairs. Motivated by structured additive regression models for function-on-scalar regression presented by Brockhaus et al. (2015), we consider the model

$$y_i = h(\mathbf{x}_i) \oplus \varepsilon_i = \bigoplus_{j=1}^{J} h_j(\mathbf{x}_i) \oplus \varepsilon_i,$$

where $\varepsilon_i \in B^2(\mu)$ are functional error terms with $\mathbb{E}(\varepsilon_i) = 0_{B^2(\mu)}$ and $h_j(\mathbf{x}_i) \in B^2(\mu), j = 1, \ldots, J$, $J \in \mathbb{N}$ are partial effects of a subset of $\mathbf{x}_i$, e.g., linear or smooth effects of one covariate, linear or smooth interaction effects of several covariates or group-specific effects. Each effect is described by a basis representation, which is the Kronecker product of two marginal bases – one in direction of the covariates, e.g., B-splines for smooth effects or the observations themselves for linear effects, and one over $\mathcal{T}$, e.g., transformed B-splines if $\mu = \lambda$ is the Lebesgue measure. A Ridge-type penalty term can be included for regularization. A suitable penalty matrix can be obtained from appropriate penalty matrices for the marginal bases (Brockhaus et al., 2015), e.g., the identity matrix (corresponding to a Ridge penalty) for linear effects or difference penalties for B-splines. Given these basis representations, we estimate the functions using a gradient boosting algorithm, where the empirical risk $\frac{1}{N} \sum_{i=1}^{N} \|y_i \ominus h(\mathbf{x}_i)\|_{B^2(\mu)}^2$ is minimized step-wise along the steepest gradient descent (with respect to the Fréchet differential). We show that this is equivalent to a minimization of the L^2-distance for the corresponding clr transformed model and base our algorithm on an extension of the one presented for functional data in Brockhaus et al. (2015), which was modified for functional data from Bühlmann and Hothorn (2007). Estimating the clr transformed model with this algorithm requires an additional integration to zero constraint and an extension to arbitrary finite measures. For mixed measures, we handle this using an orthogonal decomposition into continuous and discrete components.

4 Application

We apply our method to a data set generated from the German Socio-Economic Panel Study (SOEP) to analyze the distribution of the woman's

share in a couple's total gross labor income in Germany. Based on 154,924 individually observed couples living together in one household, where at least one partner has a labor income, we estimate 552 response densities of the woman's income share s – one for each combination of covariate values. The covariates are the *region* (one out of six) where a couple is living, *old_new* indicating whether the region contains old or new federal states, the *child group*, based on the age of the couple's youngest minor child living in its household (0 – 6/7 – 18/none), and the *year* of the observation, ranging from 1984 to 2016. 'New' federal states are the ones, which belonged to the German Democratic Republic (East Germany) after World War II and are only observed from 1991. The response densities of the share, which have to be estimated, are defined on $[0, 1]$ and have positive probability mass at values 0 and 1, corresponding to one partner without labor income. One exemplary barplot to confirm this statement is shown in Figure 1. The outmost bars for $s = 0$ and $s = 1$ have width zero, the ones

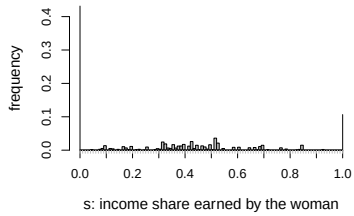


FIGURE 1. Barplot of share frequency for region north-east and child group 1 (0 – 6 years) in 2013.

in between have width 0.01. Thus, we consider the densities as elements of the Bayes Hilbert space $B^2(\mu) = B\left([0, 1], \mathfrak{B}_{[0,1]}, \mu\right)$, where $\mathfrak{B}_{[0,1]}$ is the Borel σ-algebra restricted on the interval $[0, 1]$ and the reference measure is $\mu := \delta_0 + \lambda + \delta_1$. Here, δ_x denotes the Dirac measure at $x \in \{0, 1\}$ and λ the Lebesgue measure (on $\mathfrak{B}_{[0,1]}$). For the response densities, we obtain the boundary values as the relative frequencies for $s = 0$ or $s = 1$. The density values in between are estimated using kernel density estimation and multiplying it by the relative frequency for $s \in (0, 1)$.

We estimate a model with an intercept, group-specific intercepts for the categorical covariates old_new, region, and child group, a smooth effect of the year and several interaction effects. The region effects are centered around the corresponding old_new effect for identifiability. The estimated effects can be interpreted in different ways, e.g., similar to log odds ratios, corresponding to differences on clr transformed level, or by examining ceteris paribus predictions on density level. We illustrate some results of our analysis in this way in Figures 2 and 3. Horizontal dashes at 0 and 1 correspond to the proportions of couples with no or all labor income earned by the woman, respectively. Note that all regions containing old and all

regions containing new federal states each show a strong similarity. Thus, we didn't include the effects for the regions in the discussion and focus on the coarser spatial effect old_new instead.

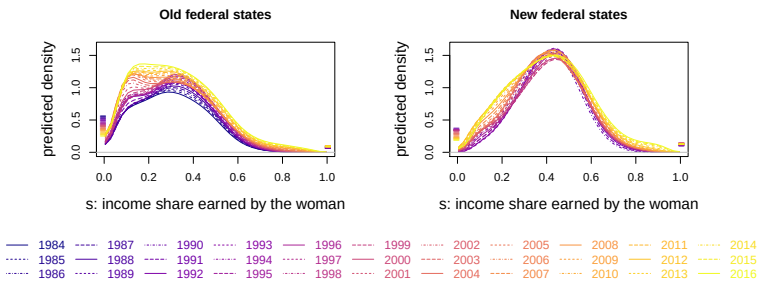


FIGURE 2. Predicted densities for old (left) and new (right) federal states.

Figure 2 shows the predictions for old vs. new federal states over the years. Both parts of Germany show a decrease of women without labor income, i.e., $s = 0$, over time. In contrast, for $s = 1$ corresponding to women who are the sole earner there is a (clearly weaker) increase overall. However, for the new federal states the maximal value is reached in the early 2000s with a slight decrease afterwards. The level of expected density values at $s = 0$ is smaller for the new federal states than for old federal states (and slightly larger for $s = 1$). This might derive from the socialist form of government in the German Democratic Republic, where it was more common that women were working compared to the Federal Republic of Germany. For dual-earner households, i.e., $s \in (0, 1)$, we find considerably different distributions in the two parts of Germany. For old federal states, the main part of probability mass is spread in the area of small shares of the couple's labor income (ca. $0.1 - 0.5$) with local maxima reached at about 0.1 or 0.35 depending on the year. In contrast, for new federal states, the predicted densities are closer to symmetric with all of them reaching their maximum at about 0.45. Regarding the development over the years, we see an increase for all $s \in (0, 1)$ for the old federal states, while the new federal states only show an increase for small and large shares ($s < 0.3$ or $s > 0.5$) and tend to decrease for intermediate shares.

In Figure 3, the predicted densities for the three child groups are illustrated. Unsurprisingly, the expected proportion of women without labor income is a lot higher for couples whose youngest child is at most six years old compared to couples with older or without children while the proportion of women being the sole earner is the highest for couples without minor children. For $s \in (0, 1)$ the shapes of the expected densities for child groups 1 and 2 are similar to each other – with the density values for child group 2 being about 1.6 times the respective values for child group 1 – and to the predicted densities for old federal states, see Figure 2. Overall, we expect

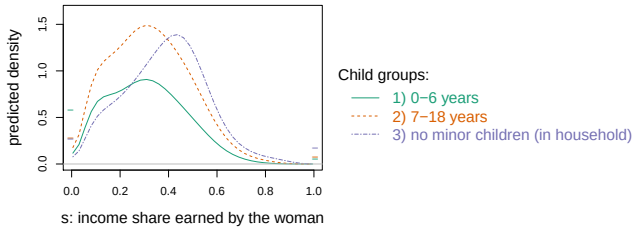


FIGURE 3. Predicted densities for the three child groups.

more dual-earner couples in child group 2 than in child group 1. The shape of the expected density for child group 3 is more alike the densities for new federal states. I.e., for couples without (minor) children the main part of the probability mass is shifted towards higher income shares compared to couples with minor children.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Boogaart, K. G. van den, Egozcue, J. J., and Pawlowsky-Glahn, V. (2014). Bayes Hilbert Spaces. *Australian & New Zealand Journal of Statistics*, **56**, 171 – 194.

Brockhaus, S., Scheipl, F., Hothorn, T., and Greven, S. (2015). The functional linear array model. *Statistical Modelling*, **15**, 279 – 300.

Bühlmann, P., and Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, **22**, 477 – 505.

Egozcue, J. J., Díaz-Barrero, J. L., and Pawlowsky-Glahn, V. (2006). Hilbert Space of Probability Density Functions Based on Aitchison Geometry. *Acta Mathematica Sinica*, **22**, 1175 – 1182.

Han, K., Müller, H.-G., and Park, B. U. (2020). Additive functional regression for densities as responses. *Journal of the American Statistical Association*, **115**, 997 – 1010.

Socio-Economic Panel Study (SOEP) (2018). Data for years 1984 – 2016, version 33, *doi: 10.5684/soep.v33.*

Talská, R., Menafoglio, A., Machalová, J., Hron, K., and Fišerová, E. (2018). Compositional regression with functional response. *Computational Statistics & Data Analysis*, **123**, 66 – 85.

Introducing non-stationarity to wrapped Gaussian spatial responses with an application to wind direction

Isa Marques[1], Nadja Klein[2], Thomas Kneib[1]

[1] University of Göttingen, Göttingen, Germany
[2] Humboldt-Universität zu Berlin, Berlin, Germany

E-mail for correspondence: imarques@uni-goettingen.de

Abstract: Circular data, i.e., data consisting of observations on the unit circle, can be found across many areas of science, for instance meteorology (wind directions), biology (animal movement directions), or medicine. The special nature of such data means that conventional methods for non-periodic data are no longer valid. As a consequence the analysis of such data is more challenging and the literature scarcer. In this paper, we introduce a spatial model for circular data that allows for non-stationarity in the mean and covariance structure of random fields. For this, we use the computationally efficient stochastic partial differential equation approach. Moreover, we develop tunable hyper-priors, inspired by the penalized complexity prior framework, that shrink the model towards a base model with stationary covariance function. The performance of the proposed model is analyzed in detail in a simulation study, with a strong focus on the properties of hyper-priors considered. Finally, we evaluate the ability of our approach to estimate wind-directions during a wind storm in Germany.

Keywords: circular data; Markov chain Monte Carlo; penalized complexity priors; stochastic partial differential equations; wind direction.

1 Introduction

Environmental and geophysical processes, such as surface winds or waves, are characterized by spatial variability. However, these data is of a periodic nature and, due to the circular geometry of the sample space, it requires reassessing typical spatial models for non-periodic data. Statistical literature on circular data spans as far as the 1970's, but it was only in the early 2010's that spatial modelling of circular data really took off. Jona-Lasinio,

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

et al. (2012) anticipated structured spatial dependence in such data types and brought typical distributions for circular data, such as the wrapped Gaussian distribution, to the realm of spatial statistics. Nonetheless, circular data models are still behind on the latest advances in spatial statistics. In this paper, we try to cover part of this gap and propose a computationally efficient model for spatial circular data, that allows for non-stationary in mean and covariance structure of the responses. In what follows, we will first shortly introduce our spatial model, and meet paths along the way and introduce the basic concepts on how to wrap data on the unit circle, thus entering the circular data domain. We present the main simulation results in Section 4. In Section 5 we will apply our model to studying wind direction during a wind storm in Germany. To conclude, we discuss the main findings of the paper.

2 The model

Let $\mathbf{s}$ denote a spatial index variable representing the location of a observation $Y(\mathbf{s})$ within a spatial domain $\mathcal{S} \subset \mathbb{R}^2$. Let $Z(\mathbf{s})$ be a total of B spatially indexed covariates where, for convenience, $Z(\mathbf{s})$ includes a 1, and has an associated coefficient vector $\boldsymbol{\beta} = (\beta_0, ..., \beta_B)'$. Then, we obtain

$$Y(\mathbf{s}) = Z(\mathbf{s})'\boldsymbol{\beta} + \gamma(\mathbf{s}) + \varepsilon(\mathbf{s}),$$

where $\gamma(\mathbf{s})$ is a zero-mean GRF and $\varepsilon(\mathbf{s})$ is a *i.i.d.* non-spatial error.
As GRFs are vulnerable to "the big n problem", a issue that renders the resulting model inappropriate for large datasets or for more complex spatial dependence structures, we instead use a recent approach in spatial statistics that replaces the GRF by an empirical equivalent Gaussian Markov random field (GMRF) during computations. This way, we exploit the sparseness of the precision $\mathbf{Q}$ of the GMRF. Hence, we take the best of two words: the good theoretical properties of GRFs and the good computational properties of GMRFs. The link between the two is achieved via the SPDE (Lindgren et al., 2011),

$$(\kappa^2 - \Delta)^{\alpha_\nu/2}(\tau\,\gamma(\mathbf{s})) = \mathcal{W}(\mathbf{s}), \quad \mathbf{s} \in \mathbb{R}^2, \quad \alpha_\nu = \nu + 1, \quad \nu > 0,$$

where Δ is the Laplacian, $\mathcal{W}$ is a Gaussian spatial white noise innovation process, $\tau > 0$ is a precision parameter and $\kappa > 0$ controls the spatial range. The solution $\gamma(\mathbf{s})$ of the resulting SPDE is a stationary GRF with Matérn covariance function. Under the finite element representation used to solve the SPDE, we get $\boldsymbol{\gamma} \sim N(\mathbf{0}, \mathbf{Q}(\tau, \kappa)^{-1})$.
In the SPDE-approach, non-stationarity in the covariance of the GRF can be attained by allowing the parameters of the SPDE to be spatially varying functions. Here, we consider

$$\log(\tau) = \theta_0^\tau \text{ and } \log(\kappa(\mathbf{s})) = \theta_0^\kappa + Z^\kappa(\mathbf{s})'\boldsymbol{\theta}_z^\kappa,$$

where $\boldsymbol{\theta}^{\kappa} = (\theta_0^{\kappa}, \boldsymbol{\theta}_z^{\kappa})'$ and θ_0^{τ} are the model's hyper-parameters, and $\mathbf{Z}^{\kappa}$ is a matrix of the relevant covariates inducing the spatial dependence. As $\boldsymbol{\kappa}$ affects both the marginal variance and range of the GRF, with this parameterization we make both non-stationary. The parameters θ_0^{κ} and θ_0^{τ} represent the stationary covariance specification of the model and have proper uniform priors. We choose a Gaussian prior $N(0, \xi^2\mathbf{I})$ for $Z^{\kappa}(\mathbf{s})'\boldsymbol{\theta}_z^{\kappa}$, where the prior for ξ^2 is constructed such that it penalizes non-stationarity in the covariance function; i.e., it shrinks towards the stationarity GRF case (see Section 3).

Finally, we need to bring the model into the circular data domain. The wrapped Gaussian distribution for circular data takes the linear variable, $Y(\mathbf{s}) \in \mathbb{R}$, for all $\mathbf{s} \in \mathcal{S}$, and wraps it around the unit circle. The result is a circular, or *wrapped*, variable $X(\mathbf{s}) \in [0, 2\pi)$; i.e, $X(\mathbf{s}) = Y(\mathbf{s}) \mod 2\pi \in [0, 2\pi)$. This can be re-written as $Y(\mathbf{s}) = \gamma(\mathbf{s}) = X(\mathbf{s}) + 2\pi K(\mathbf{s})$, where the *winding number*, $K(\mathbf{s}) \in \mathbb{Z}$, measures the number of "turns" around the unit circle. This strategy allows one to adopt popular distributions for non-periodic data and simply wrap them around the unit circle. Consequently, if we assume $Y(\mathbf{s})$ has a Gaussian distribution, $X(\mathbf{s})$ has a wrapped Gaussian distribution.

3 The penalized complexity prior

We follow a Bayesian approach and develop a penalized complexity (PC) prior for the hyper-parameters $\boldsymbol{\theta}_z^{\kappa}$ of the model (Simpson et al. (2017)). In this setting, we define a model with stationary covariance as the base model. Hence, the prior will favor a stationary GRF unless the data indicates otherwise.

Consider a flat prior for θ_0^{τ} and θ_0^{κ} and let $\boldsymbol{\theta}_z^{\kappa} \sim N(\mathbf{0}, \xi^2\mathbf{I})$. One can show that if the base model is such that $\xi^2 \to 0$ and the prior is constructed according the PC-prior principles, then ξ^2 has a Weibull prior with shape $\frac{1}{2}$ and scale λ, i.e., $\xi^2 \sim \text{Weibull}(\frac{1}{2}, \lambda)$ (Klein and Kneib, 2016). We choose λ using a user-defined approach based on the probability statement

$$P\left(\max_{\mathbf{s}\in\mathcal{S}} \mid Z^{\kappa}(\mathbf{s})'\boldsymbol{\theta}_z^{\kappa} \mid \leq c\right) \geq 1 - \alpha,$$

i.e., we model the probability that the maximum norm of the non-stationary effect is smaller than a pre-specified level and determine the distribution of the maximum based on simulations from the prior $p(\mathbf{Z}^{\kappa}(\mathbf{s}_m)'\boldsymbol{\theta}_z^{\kappa}|\xi^2)$.

By design, the choice of c and α is an *ad-hoc*, problem specific, choice. Nonetheless, on practical terms, it is possible to restrict the values $\boldsymbol{\kappa}$ can take. Namely, as long as the minimum range of the non-stationary component is large enough; i.e., it really reflects spatial dependence, it should be possible to derive such a general bound. We investigate this in Section 4.

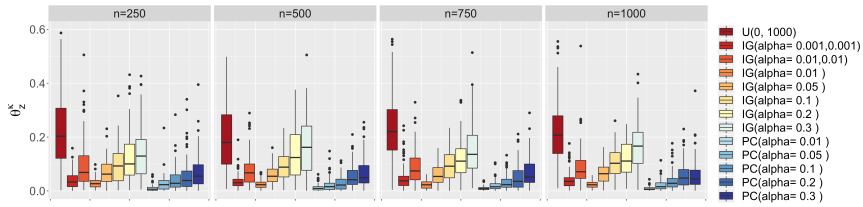


FIGURE 1. Posterior mean of θ_z^κ for data generated with stationary covariance function, i.e, true value is zero.

4 Simulation Study

Consider standardized covariates, $\mathbf{Z}$ and $\mathbf{Z}^\kappa$, and $\mathcal{S} \subseteq [0,1] \times [0,1]$. During simulations, we identified two important scenarios where a PC-prior outperforms typical priors used for variance parameters, e.g., inverse gamma prior with scale and shape in $\{0.001, 0.01\}$, or uniform priors. Namely:

1. **A PC-prior prevents overfitting:** this is clear in Figure 1, where we generate data with a stationary covariance function.
2. **A PC-prior improves estimation for complex spatial dependence structures:** the PC-prior, in general, reaches both the lowest root mean squared error (RMSE) for the posterior mean of the centered GRF and lower dispersion. This is particularly evident for scenarios with non-stationary behavior in the covariance function at the boundary of the domain.

When it comes to selecting c and α, the combination of $\alpha = 0.01$ with a c up to twice as large as the true maximum performs reasonably well. A general bound $c = 2$ works well for $\alpha = 0.01$. A more generalized approach could comprise a rescaling of $c = 2$ to new domain dimensions, keeping $\alpha = 0.01$. This gives us some flexibility when setting up PC-priors on a real dataset for which we do not know c. We use this in Section 5.

5 German wind direction data

In Germany, wind direction is characterized by predominant westerly winds, coming from the Atlantic Ocean and entering Germany through France. On the eastern side, wind is generated in the Caucasus, entering Germany through Poland and the Czech Republic. Both winds collide in the northern tip of Germany. The wrapped Gaussian distribution is unimodal and, consequently, we need to avoid situations in which over a large region, at a given time, a storm is rotating or two different weather systems

TABLE 1. CRPS for the two models considered.

Model	$\boldsymbol{\nu = 1}$	$\boldsymbol{\nu = 1.5}$
full stationary	0.204	0.240
full non-stationary	0.064	0.069

are meeting. Given this, we select data from the wind storm of September 30, 2019, which is characterized by predominantly eastern winds.
Model performance is evaluated using the circular continuous ranked probability score:

$$\text{CRPS}_{circ}(P, X) = E\{\alpha_{\text{CRPS}}(x, X)\} - \frac{1}{2}E\{\alpha_{\text{CRPS}}(x, x^*)\},$$

where α_{CRPS} represents the cosine distance, P is a forecast distribution on the circle, x and x^* are independent copies of a circular random variable with distribution P and X is the verifying direction. The results are expressed in units of angular distance, with a maximum allowed of π.
In our analysis, we consider the following covariates, $\mathbf{Z}_\kappa$ and $\mathbf{Z}$: maximum wind speed, altitude, average air temperature at 5 meters height, average air pressure at 10 meters height, longitude, latitude, an indicator for being at the northern German tip (state of Schleswig-Holstein) and an indicator for being close to the French border.
For the analysis, we randomly select a holdout set of data consisting of 20% of the locations and use the remaining 80% as training data. Here, we present the results for two models:

1. **Full stationary:** $\mathbf{Z} = \mathbf{Z}_\kappa = \mathbf{0}$.
2. **Full non-stationary:** $\mathbf{Z}$ includes all covariates and $\mathbf{Z}_\kappa$ includes all but the indicator variables.

Results show that the full non-stationary wrapped Gaussian spatial response model can approximate the true wind directions quite closely. This can be confirmed in Figure 2, for the test data considered. Additionally, we tested the results for $\nu \in \{0.5, 0.75, 1, 1.25, 1.5, 1.75\}$. In Table 1 we show the CRPS for the best performing models. The CRPS values attained are quite low.

6 Discussion

The developed model improves results over ordinary stationary GRF methods for modeling spatially wind direction data. Such interpolation properties for wind direction could serve as an input for other modeling tasks in the analysis of climate variables, which, in combination with the use of

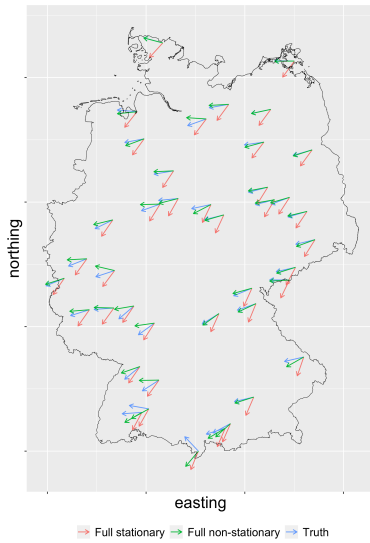


FIGURE 2. True and estimated mean wind directions for $\nu = 1$ for the test data.

sparse spatial precision matrices, would have great potential in the generation of efficient models for large scale datasets for meteorological data. The mere use of a reasonable number of interpretable covariates with understandable physical properties makes the model more intuitive and applicable by a broad range of researchers, in a wide spectrum of areas.

References

Jona-Lasinio, G., Gelfand, A., Jona-Lasinio, M. (2012). Spatial analysis of wave direction data using wrapped Gaussian processes. *The Annals of Applied Statistics*, **6**, 1478 – 1498.

Klein, N., Thomas, K. (2016). Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Analysis*, **11**, 1071 – 1106.

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B*, **73**, 423 – 498.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Soerbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, **32**, 1 – 28.

Continuous-time modelling of the hot hand effect in basketball free throws

Sina Mews[1], Marius Ötting[1], Houda Yaqine[1], Roland Langrock[1]

[1] Bielefeld University, Germany

E-mail for correspondence: sina.mews@uni-bielefeld.de

Abstract: We investigate the hot hand phenomenon in basketball using data on 110,513 free throws. As these occur at unevenly spaced time points within a game, we formulate a continuous-time state-space model to relate the actual throwing performance to the latent underlying form of a player. Our results reveal serial correlation in the latent throwing success probability, thus supporting the existence of a hot hand effect.

Keywords: Continuous-time model; Hot hand; Ornstein-Uhlenbeck process; State-space model.

1 Introduction

The existence of a hot hand effect, according to which sports athletes may temporarily enter a state during which they perform better than on average, is a much-debated topic among sports commentators, fans, and journalists. In the academic literature, the hot hand has gained great interest since the seminal paper by Gilovich et al. (1985), in which the hot hand effect was dismissed as a cognitive illusion. Driven by the increased accessibility of large sports data sets, there is an ever-growing body of research investigating the existence of the hot hand effect, yet the evidence remains inconclusive (see Bar-Eli et al., 2006, for a review). Moreover, there is no universally accepted definition as to what exactly constitutes a hot hand effect: while some people regard it as serial correlation in *outcomes* (see, e.g., Miller and Sanjurjo, 2018), others consider it as serial correlation in *success probabilities* (see, e.g., tting et al., 2020). The latter definition translates into a latent (state) process underlying the observed performance — intuitively

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

speaking, a measure for a player's "hotness" — which can be elevated without the player necessarily making every shot. As shots, or similar events in sports with game clocks, usually occur at points that are unevenly spaced in time, modelling such events requires a continuous-time approach. We thus develop a state-space model in continuous time to investigate the hot hand effect for free throws in basketball.

2 Data

We extracted data on more than 9,000 basketball games in the NBA for all seasons and playoffs between October 2013 and June 2019 from `https://www.basketball-reference.com/`. For our analysis, we only consider data on free throw attempts as these constitute highly standardised settings without any interaction between players, which is usually hard to account for when modelling field goals in basketball. In our analysis, we include all players who took at least 2,000 free throws in the period considered, totalling in 110,513 free throws from 44 players. There is considerable heterogeneity in the players' throwing success, with the corresponding empirical proportions for making a free throw ranging from 45.1% to 90.6%. As free throws occur irregularly within a basketball game, the information on whether an attempt was successful needs to be supplemented by its time t, corresponding to the time already played in minutes. For each player p in his n-th game, we thus observe an irregular sequence of binary variables $\{x_t^{p,n}\}_{t\geq 0}$, with

$$x_t^{p,n} = \begin{cases} 1 & \text{if free throw attempt at time } t \text{ is successful;} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the sequential data we model looks as follows:

$x_t^{p,n}$:	1	1	1	0	1	1	1	1
t:	8.55	8.55	10.33	10.33	19.64	24.97	24.97	24.97

These example data, from one match played by James Harden, illustrate that free throw attempts often appear in clusters of 2 or 3 attempts at the same time (depending on the foul), followed by a time period without any free throws. Therefore, it is important to take into account the different lengths of the time intervals between consecutive attempts, which is why we formulate our model in continuous time.

3 Model formulation and estimation

Following the idea that the observed throwing success depends on a player's current (latent) form, we model the observed free throw attempts $x_t^{p,n}$ using a state-space model formulation with a binary response. The associated

predictor on the logit scale comprises information on the player's average throwing success as well as his current state $s_t^{p,n}$ — additional explanatory variables can easily be added (see below). Dropping the superscripts p and n for notational simplicity, we hence have

$$x_t \sim \text{Bern}(\pi_t), \quad \text{logit}(\pi_t) = \alpha_p + s_t, \tag{1}$$

where α_p is a player-specific intercept and s_t is the underlying latent state, which can be interpreted as the player's current form (or "hotness"). The stochastic process $\{s_t\}_{t\geq 0}$, which is formulated in continuous time to address the temporal irregularity of the observation times, ought to be continuous-valued to allow for gradual changes in a player's form, and stationary such that in the long run it returns to the average form. The natural candidate for a corresponding stationary, continuous-time and continuous-valued stochastic process is the Ornstein-Uhlenbeck (OU) process,

$$ds_t = -\beta s_t dt + \sigma dW_t,$$

where $\beta > 0$ is the drift parameter indicating the strength of reversion to the long-term mean 0, $\sigma > 0$ controls the strength of fluctuations, and W_t denotes the Wiener process. Since we model the hot hand effect as a serial correlation in success probabilities, our main parameter of interest is the drift parameter β governing the speed of reversion (to the average form). The smaller β, the longer it takes for the OU process to return to its mean and thus the higher the serial correlation.

As the model's likelihood involves intractable integration over all possible realisations of the continuous-valued s_t, at each observation time, we approximate the integral by finely discretising the state space. This approximation can be seen as a reframing of the model as a continuous-time hidden Markov model with a large but finite number of states, enabling us to apply the corresponding efficient algorithms. In particular, we use the forward algorithm to calculate the likelihood, making use of the limiting distribution as well as the conditional distribution of the OU process to compute the initial state probabilities as well as the state transition probabilities. The model parameters are then estimated by numerically maximising the (approximate) joint likelihood over all games and all players.

4 Preliminary results

To investigate any potential hot hand effect for sequences of free throws, we fit two models to the data. First, we consider a benchmark model (*Model 1*) without any hot hand effect, i.e. a model without the underlying state process s_t in Equation (1) and as such, without any serial dependence. The second model (*Model 2*) includes the underlying state process as described in Section 3. Besides the player-specific intercepts, in both models we control for additional covariates in the linear predictor in Equation (1) which

may affect the players' throwing success (namely the current score difference, a home vs. away dummy, a dummy indicating whether the free throw occurred in the last 30 seconds of the quarter, and dummies indicating if it was the second or third throw in a row).
Model 2, i.e. the formulation including a potential hot hand effect, is clearly favoured over Model 1 (ΔBIC = 41.97). The estimated parameters of the OU process for Model 2 are $\hat{\beta} = 0.042$ (95% CI $[0.016; 0.109]$) and $\hat{\sigma} = 0.101$ (95% CI $[0.055; 0.185]$), respectively. The estimated drift parameter $\hat{\beta}$ is fairly small, indicating serial correlation of the state process over time, and thus providing evidence for a hot hand effect. This is highlighted also by simulated state trajectories based on the fitted model (Figure 1).

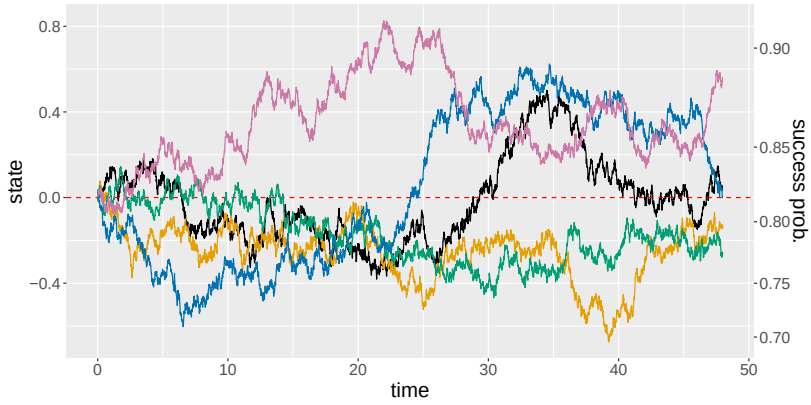


FIGURE 1. Simulation of possible state trajectories for the length of an NBA game based on the estimated parameters of the OU process. The red dashed line indicates the intercept (here: the median throwing success over all players), around which the processes fluctuate. The right y-axis shows the success probabilities resulting from the current state (left y-axis).

References

Bar-Eli, M., Avugos, S., and Raab, M. (2006). Twenty years of "hot hand" research: review and critique. *Psychology of Sport and Exercise*, **7**, 525 – 553.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: on the misperception of random sequences. *Cognitive Psychology*, **17**, 295 – 314.

Miller, J.B. and Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, **86**, 2019 – 2047.

tting, M., Langrock, R., Deutscher, C., and Leos-Barajas, V. (2020). The hot hand in professional darts. *Journal of the Royal Statistical Society, Series A*, **183**, 565 – 580.

New statistical model for misreported data

David Moriña[1,2], Amanda Fernández-Fontelo[1,3], Alejandra Cabaña[1], Pedro Puig[1]

[1] Barcelona Graduate School of Mathematics (BGSMath), Departament de Matemàtiques, Universitat Autònoma de Barcelona, Spain
[2] Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona, Spain
[3] Chair of Statistics, School of Business and Economics, Humboldt-Universität zu Berlin, Germany

E-mail for correspondence: dmorina@mat.uab.cat

Abstract: The main goal of this work is to present a new model able to deal with potentially misreported continuous time series. The proposed model is able to handle the autocorrelation structure in continuous time series data, which might be misreported. Its performance is illustrated through a real data application on COVID-19 disease in a chinese province, and a comprehensive simulation study is also discussed.

Keywords: continuous time series; mixture distributions; under-recorded data; ARMA models; public health.

1 Introduction

There is a growing interest in the last years to deal with data that is only partially registered or underreported in the time series literature (Fernndez-Fontelo et al. (2016)). This phenomenon is very common in many fields, and has been previously explored by different approaches in epidemiology, social and biomedical research among many other contexts. Many approaches to deal with underreported data have been suggested with a growing level of sophistication from the usage of multiplication factors to spatio-temporal modelling.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Model definition and properties

Consider an unobservable process with an AutoRegressive Moving Average ($ARMA$) structure defined by

$$\begin{aligned} X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \theta_1 \epsilon_{t-1} + \ldots + \\ + \theta_q \epsilon_{t-q} + \epsilon_t, \end{aligned} \tag{1}$$

where ϵ_t is a white noise process with $\epsilon_t \sim N(\mu_\epsilon, \sigma_\epsilon^2)$. In our setting, this process cannot be directly observed, and all we can see is a part of it, expressed as

$$Y_t = \begin{cases} X_t \text{ with probability } 1-\omega \\ q \cdot X_t \text{ with probability } \omega \end{cases} \tag{2}$$

The interpretation of the parameters in Eq. (2) is straightforward: q is the intensity of misreporting (if $0 < q < 1$ the observed process Y_t would be underreported while if $q > 1$ the observed process Y_t would be overreported). The parameter ω can be interpreted as the frequency of misreporting (proportion of misreported observations).

If the unobserved process X_t follows an $ARMA(p, q)$ model as defined in Eq. (1), the observed process has mean $\mathbb{E}(Y_t) = \frac{\mu_\epsilon}{1-\alpha_1-\ldots-\alpha_p} \cdot (1 - \omega + q \cdot \omega)$ and variance $\mathbb{V}(Y_t) = \left(\left(\frac{\sigma_\epsilon^2 \cdot (1+\theta_1^2+\ldots+\theta_q^2)}{1-\alpha_1^2-\ldots-\alpha_p^2} \right) + \frac{\mu_\epsilon^2}{(1-\alpha_1-\ldots-\alpha_p)^2} \right) \cdot (1 + \omega \cdot (q^2 - 1)) - \frac{\mu_\epsilon^2}{(1-\alpha_1-\ldots-\alpha_p)^2} \cdot (1 - \omega + q \cdot \omega)^2$. The autocorrelation function of the observed process can be written in terms of the properties of the hidden process X_t as

$$\begin{aligned} &\rho_Y(k) = \\ &= \frac{V(X_t) \cdot \rho_X(k) \cdot (1-\omega+q\cdot\omega)^2}{(V(X_t)+E(X_t)^2) \cdot (1+\omega\cdot(q^2-1)) - E(X_t)^2 \cdot (1-\omega+q\cdot\omega)^2} = \\ &= c(\alpha_1, \ldots, \alpha_p, \theta_1, \ldots, \theta_q, \mu_\epsilon, \sigma_\epsilon^2, \omega, q) \cdot \rho_X(k), \end{aligned} \tag{3}$$

where ρ_X is the autocorrelation function of the unobserved process X_t.

The likelihood function of the observed process Y_t is not directly obtainable but the parameters of the model can be estimated by means of an iterative algorithm based on its marginal distribution, using R packages *mixtools* (Benaglia et al. (2009)) and *forecast* (Hyndman et al. (2008)). The main steps are described in detail below:

(i) Following Eq. (2), the observed process Y_t can be written as $Y_t = (1 - Z_t) \cdot X_t + q \cdot Z_t \cdot X_t$, where Z_t is an indicator of the underreported observations, following a Bernoulli distribution with probability of success ω ($Z_t \sim Bern(\omega)$), its marginal distribution is a mixture of two normal random variables $N\left(\mu, \sigma^2\right)$ and $N\left(q \cdot \mu, q^2 \cdot \sigma^2\right)$ respectively, where $\mu = \frac{\mu_\epsilon}{1-\alpha_1-\ldots-\alpha_p}$ and $\sigma^2 = \frac{\sigma_\epsilon^2 \cdot (1+\theta_1^2+\ldots+\theta_r^2)}{1-\alpha_1^2-\ldots-\alpha_p^2}$. This fact can be used to obtain initial estimates for q and ω. Using the E-M algorithm (specifically on the E-step), the package *mixtools* calculates

the posterior probabilities (conditional on the data and the obtained estimates) of each observation to come from one of these two normals.

(ii) Using the indicator $\hat{Z}_t$ obtained in the previous step, the series is divided in two: One including the underreported observations (treating the non-underreported values as missing data) and another with the non underreported observations (treating the underreported values as missing data). An $ARIMA$ model is fitted to each of these two series and a new $\hat{q}$ is obtained by dividing the fitted means.

(iii) A mixture of two normals is fitted to the observed series Y_t with mean and standard deviation fixed to the corresponding values obtained from the previous step, and a new ω is estimated.

(iv) Steps (ii) and (iii) are repeated until the quadratic distance between two consecutive iterations $(\hat{q}_i - \hat{q}_{i-1})^2 + (\hat{\omega}_i - \hat{\omega}_{i-1})^2 + \sum_j (\hat{\alpha}_{j_i} - \hat{\alpha}_{j_{i-1}})^2 + , + \sum_k (\hat{\theta}_{k_i} - \hat{\theta}_{k_{i-1}})^2$ is below a fixed tolerance level.

(v) Once the parameter estimates are stable according to the previous criterion, the underlying process X_t is reconstructed as $\hat{X}_t = (1 - \hat{Z}_t) \cdot Y_t + \frac{1}{\hat{q}} \cdot \hat{Z}_t \cdot Y_t$, and an $ARIMA$ model is fitted to the reconstructed process to obtain $\hat{\alpha}_j$, $j = 1, \ldots, p$, $\hat{\theta}_k$, $k = 1, \ldots, r$ and $\hat{\sigma_\epsilon}^2$.

To account for potential trends or seasonal behaviour, covariates can be included in the described estimation process. Additionally, a parametric bootstrap procedure with 1000 replicates is used to estimate standard errors and build confidence intervals based on the percentiles of the distribution of the estimates.

3 Results

3.1 Simulation study

A thorough simulation study has been conducted to ensure that the model behaves as expected in several situations, including $AR(1)$, $MA(1)$ and $ARMA(1,1)$ structures for the hidden process X_t with values for the parameters α, θ, q and ω ranging from 0.1 to 0.9 for each parameter. For each autocorrelation structure and parameters combination, a random sample of size $n = 1000$ has been generated using the function *arima.sim* from R package *forecast*. Absolute average bias is similar regardless of the sample size, while average interval lengths (AIL) are higher and interval coverages are poorer (around 75% for $n = 50$) for lower sample sizes as could be expected. The average absolute bias, interval coverage and 95% confidence interval length are reported in Table 1. These values are averaged over all combinations of parameters. Additionally, standard $AR(1)$, $MA(1)$

TABLE 1. Model performance measures (average absolute bias, average interval length and average coverage) summary based on a simulation study.

Structure	Parameter	Bias	AIL	Coverage (%)
$AR(1)$	$\hat{\alpha}$	0.003	0.099	94.92%
	$\hat{q}$	$< 10^{-3}$	$< 10^{-3}$	95.47%
	$\hat{\omega}$	-0.001	0.052	92.46%
Standard $AR(1)$	$\hat{\alpha}$	0.501	0.124	0.69%
$MA(1)$	$\hat{\theta}$	$< 10^{-3}$	0.117	94.38%
	$\hat{q}$	$< 10^{-3}$	$< 10^{-3}$	94.38%
	$\hat{\omega}$	$< 10^{-3}$	0.052	93.28%
Standard $MA(1)$	$\hat{\theta}$	0.502	0.124	1.10%
$ARMA(1,1)$	$\hat{\alpha}$	0.002	0.165	96.02%
	$\hat{\theta}$	0.007	0.210	96.56%
	$\hat{q}$	$< 10^{-3}$	0.002	94.56%
	$\hat{\omega}$	$< 10^{-3}$	0.059	93.22%
Standard $ARMA(1,1)$	$\hat{\alpha}$	0.456	3.558	59.08%
	$\hat{\theta}$	0.579	3.496	56.00%

and $ARMA(1,1)$ models were fitted to the same simulated series without accounting for their underreporting structure.
It is clear from Table 1 that ignoring the underreported nature of data (labeled as *Standard* models in the table) leads to highly biased estimates with extremely low coverage rates, even with larger average interval lengths. This is especially relevant when the intensity or frequency of underreported observations is high.

3.2 COVID-19 incidence in the region of Heilongjiang

SARS-CoV-2 is a betacoronavirus that affects the lower respiratory tract and often manifests as pneumonia in humans which was identified as the causative agent of an unprecedented outbreak of pneumonia in Wuhan City, Hubei province in China starting in December 2019. Considering that many cases run without developing symptoms beyond those of MERS-CoV, SARS-CoV or pneumonia due to other causes, it is reasonable to assume that the incidence of this disease has been underregistered, especially at the beginning of the outbreak (Zhao et al. (2009)).
Heilongjiang is a province in north-east China. Although in general the behavior of this kind of diseases is far from being stationary, this province is far enough from the focus in Hubei province (south central China) so the

incidence is much lower and less explosive, and in the study period of time (2020/01/22-2020/02/26) it can be considered stationary, as can be seen in Figure 2, which shows registered and estimated evolution of COVID-19 incidence within the considered period of time.
A disease with a similar behavior (MERS-CoV) was modeled as an $ARMA(3,1)$ in Alkhamis et al. (2019), so we checked this model and similar ones. However, in our case the best performing model was an $MA(1)$ (AIC of -151.31 against -148.82 for the $ARMA(3,1)$), consistently with the residuals profile shown in Figure 1, obtained from fitting an $MA(1)$ model to the most likely process X_t reconstructed as described before.

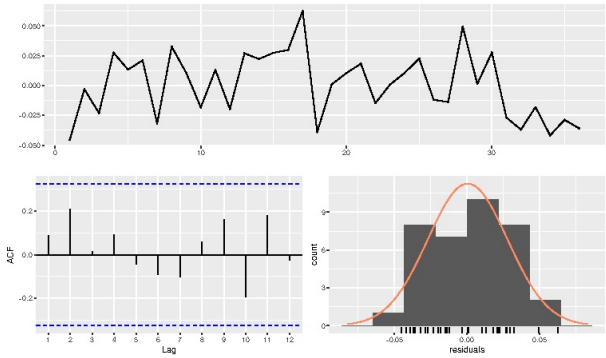


FIGURE 1. Residuals analysis of the residuals from an $MA(1)$ model.

By means of the described estimation method, it can be seen that the estimated model for the hidden process is $X_t = 0.481 \cdot \epsilon_{t-1} + \epsilon_t$, being the observed process Y_t,

$$Y_t = \begin{cases} X_t & \text{with probability } 0.507 \\ 0.195 \cdot X_t & \text{with probability } 0.493 \end{cases} \tag{4}$$

The estimated parameters are reported in Table 2.

Parameter	Bootstrap mean	Bootstrap SE
$\hat{\theta}$	0.481	0.179
$\hat{\omega}$	0.493	0.168
$\hat{q}$	0.195	0.089

TABLE 2. Bootstrap means and standard errors of the proposed model.

Acknowledgments: David Moriña acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the María

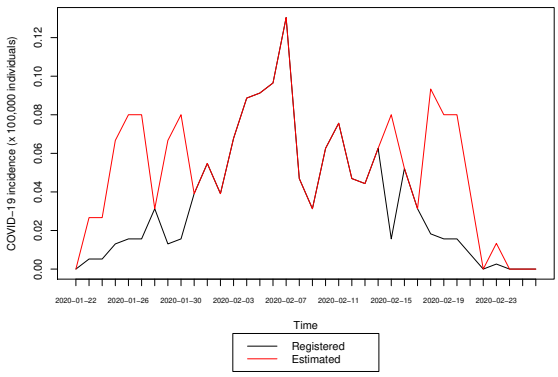


FIGURE 2. Registered and estimated COVID-19 incidence in the region of Heilongjiang in the period 2020/01/22-2020/02/26.

de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445), Instituto de Salud Carlos III (COV20/00115) and Fundación Santander Universidades.

References

Alkhamis, M. A., Fernández-Fontelo, A., VanderWaal, K. et al. (2019). Temporal dynamics of Middle East respiratory syndrome, 2012-2017 coronavirus in the Arabian Peninsula. *Epidemiology & Infection*, **147**, 1 – 10.

Fernández-Fontelo, A., Cabaña, A., Puig, P., Moria, D. (2016). *Statistics in Medicine*, **35(26)**, 4875 – 4890.

Benaglia, T., Chauveau, D., Hunter, D.R., Young, D.S. (2009). mixtools : An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, **32(6)**, 1 – 29.

Hyndman RJ and Khandakar Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, **27(3)**, 1 – 22.

Zhao, S., Musa, S. S., Lin, Q., Ran, J., Yang, G. et al. (2020). Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: A data-driven modelling analysis of the early outbreak. *Journal of Clinical Medicine*, **9(2)**, 388.

Regularisation in hidden Markov models with an application to football data

Marius Ötting[1], Andreas Groll[2]

[1] Bielefeld University, Germany
[2] TU Dortmund University, Germany

E-mail for correspondence: marius.oetting@uni-bielefeld.de

Abstract: We propose a modelling framework for dealing with a large amount of covariates in hidden Markov models (HMMs) by considering a LASSO penalty. This modelling framework is, for example, useful in sports for analysing a potential hot hand effect, as several existing studies on the hot hand consider HMMs. However, with most studies analysing data from basketball or baseball, there are several confounding factors which have to be taken into account, leading to a potential large number of covariates. Hence, in those settings regularisation methods are suitable to allow for implicit variable selection. As a case study we investigate a potential "hot shoe" effect among penalty-takers.

Keywords: hidden Markov model; LASSO; hot hand; sports analytics; football.

1 Introduction

An often discussed phenomenon in different sports is the "hot hand", meaning that players may enter a state where they experience extraordinary success. This phenomenon is also discussed in the media, where commentators and journalists — e.g. in football — commonly refer to players as being "on fire" when they score in consecutive matches. Academic research on the hot hand started by Gilovich et al. (1985). In their seminal paper, they analysed basketball free-throw data and found no evidence for the hot hand, arguing that people tend to belief in the hot hand due to memory bias.
More recent studies challenge the findings of Gilovich et al. (1985), often by analysing data from basketball or baseball with regard to a hot hand effect. In addition, these studies often consider hidden Markov models (HMMs), which constitute a natural modelling approach for the hot hand as they accommodate the idea that players potentially may enter a state where they

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

experience extraordinary success. However, when modelling a potential hot hand effect, there is hardly any sport where no potential confounding factors exist, such as weather conditions in baseball or the performance of opponents in basketball. Accounting for those factors leads to a large number of covariates, and often multicollinearity issues occur, making model fitting and interpretation of parameters difficult. To tackle these problems and to obtain sparse and interpretable models, we propose to conduct variable selection in HMMs by considering a LASSO penalisation approach (see Tibshirani, 1996).

The performance of LASSO-HMMs is first investigated in a simulation study. As a case study, we investigate a potential "hot shoe" effect of penalty takers in the German Bundesliga ($n = 3,482$ penalties). Figure 1 shows all penalties taken by Bayern Munich's attacker Gerd Müller, indicating that there are periods (e.g. between 1975 and 1976) where he scored several penalties in a row, but also periods (e.g. around 1971) where he missed a few consecutive penalties.

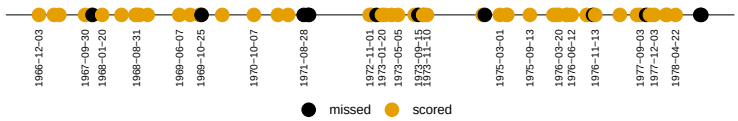


FIGURE 1. Penalty history over time of the player Gerd Müller for the time period from 1964 until 1979 (successful penalties in yellow, failures in black).

2 Methods

In HMMs, the observations y_t are assumed to be driven by an underlying state process s_t, in a sense that the y_t are generated by one of N distributions according to the Markov chain. In our application, the state process s_t serves for the underlying varying form of a player. State switching is modelled by the transition probability matrix (t.p.m.) $\boldsymbol{\Gamma} = (\gamma_{ij})$, with $\gamma_{ij} = \Pr(s_t = j | s_{t-1} = i)$, $i, j = 1, \ldots, N$. We further allow for additional covariates at time t, $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{Kt})^\top$, each of which assumed to have the same effect in each state, whereas the intercept is assumed to vary across the states, leading to the following linear state-dependent predictor:

$$\eta_t^{(s_t)} = \beta_0^{(s_t)} + \beta_1 x_{1t} + \ldots + \beta_k x_{Kt} \,.$$

For our response variable y_t, indicating whether the penalty attempt t was successful or not, we assume $y_t \sim \text{Bern}(\pi_t^{(s_t)})$ and link $\pi_t^{(s_t)}$ to our state-dependent linear predictor $\eta_t^{(s_t)}$ using the logit link function, i.e. $\text{logit}(\pi_t^{(s_t)}) = \eta_t^{(s_t)}$. Defining an $N \times N$ diagonal matrix $\boldsymbol{P}(y_t)$ with i–th diagonal element being equal to $\Pr(y_t | s_t = i)$, and assuming that the

initial distribution $\boldsymbol{\delta}$ of a player is equal to the stationary distribution, i.e. the solution to $\boldsymbol{\Gamma\delta} = \boldsymbol{\delta}$ subject to $\sum_{i=1}^{N} \delta_i = 1$, the likelihood for a single player p is given by

$$L_p(\boldsymbol{\alpha}) = \boldsymbol{\delta}\mathbf{P}(y_{p1})\boldsymbol{\Gamma}\mathbf{P}(y_{p2}) \dots \boldsymbol{\Gamma}\mathbf{P}(y_{pT_p})\mathbf{1}\,,$$

with vector $\boldsymbol{\alpha} = (\gamma_{11}, \gamma_{12}, \dots, \gamma_{1N}, \dots, \gamma_{NN}, \beta_0^{(1)}, \dots, \beta_0^{(N)}, \beta_1, \dots, \beta_k)^\intercal$ collecting all unknown parameters, and column vector $\mathbf{1} = (1, \dots, 1)^\intercal \in \mathbb{R}^N$ (see Zucchini et al., 2016). To obtain the likelihood for the complete data set, i.e. for multiple players, we assume independence between the observations of different players (here: $p = 310$), so that the likelihood is given by the product of the individual likelihoods:

$$L(\boldsymbol{\alpha}) = \prod_{p=1}^{310} L_p(\boldsymbol{\alpha}) = \prod_{p=1}^{310} \boldsymbol{\delta}\mathbf{P}(y_{p1})\boldsymbol{\Gamma}\mathbf{P}(y_{p2}) \dots \boldsymbol{\Gamma}\mathbf{P}(y_{pT_p})\mathbf{1}\,.$$

Parameter estimation is done by maximising the likelihood numerically using `nlm()` in R. However, considering a large amount of covariates leads to a rather complex model, which is hard to interpret and, in addition, multicollinearity issues might occur. Hence, we propose to employ a penalised likelihood approach based on a LASSO penalty.

The basic idea is to maximise a penalised version of the log-likelihood $\ell(\boldsymbol{\alpha}) = \log(L(\boldsymbol{\alpha}))$. More precisely, one maximises the penalised log-likelihood

$$\ell_{\text{pen}}(\boldsymbol{\alpha}) = \log(L(\boldsymbol{\alpha})) - \lambda \sum_{k=1}^{K} |\beta_k|\,, \tag{1}$$

where λ represents a tuning parameter, which controls the strength of the penalisation. To fully incorporate the LASSO penalty in our setting, the non-differentiable L_1 norm $|\beta_k|$ in (1) is approximated as suggested by Oelker and Tutz (2017). Specifically, $|\beta_k|$ is approximated by $\sqrt{(\beta_k + c)^2}$, where c is a small positive number (say $c = 10^{-5}$). Practically, a coefficient is then selected if $|\hat{\beta}_k| \geq 0.001$. The optimal value for the tuning parameter λ is chosen by model selection criteria such as AIC and BIC. To estimate the required effective degrees of freedom, we consider all parameters in the model which are unequal to zero, i.e. all entries of the t.p.m., all state-dependent intercepts, and all selected β_j's.

3 Simulation study

We consider a simulation scenario similar to our real-data application, with a Bernoulli-distributed response variable, an underlying 2-state Markov chain and 50 covariates, 47 of which being noise covariates:

$$y_t \sim \text{Bern}(\pi_t^{(s_t)}), \quad \text{with}$$

$$\text{logit}(\pi_t^{(s_t)}) = \eta_t^{(s_t)} = \beta_0^{(s_t)} + 0.5 \cdot x_{1t} + 0.7 \cdot x_{2t} - 0.8 \cdot x_{3t} + \sum_{j=4}^{47} 0 \cdot x_{jt}\,.$$

We further set $\beta_0^{(1)} = \text{logit}(0.75)$ and $\beta_0^{(2)} = \text{logit}(0.35)$. The performance of three different fitting schemes is investigated, namely HMMs without penalisation (i.e. $\lambda = 0$) and the LASSO-HMM with λ selected by AIC and BIC, respectively. The fitting schemes are compared by the mean squared error (MSE) of the β_j (see Figure 2). The results of the simulation study suggest that, in terms of MSE, the LASSO-HMM with λ selected by BIC performs worst, with the MSE being higher than for the HMM without penalisation. The LASSO-HMM with λ selected by AIC outperforms the other fitting schemes considered in terms of MSE.

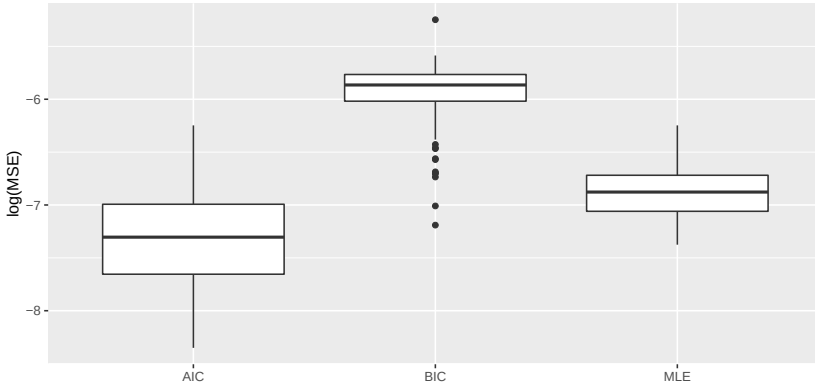


FIGURE 2. Boxplots of the MSE obtained in 100 simulation runs. "AIC" and "BIC" denote the LASSO-HMM fitting schemes with λ chosen by AIC/BIC. "MLE" denotes unpenalised HMM.

4 Application

As the LASSO-HMM with λ selected by the AIC showed the most promising results in the simulations, we focus on the results obtained by this fitting scheme. For modelling the hot shoe, we account for several factors potentially affecting the outcome of a penalty kick, namely a dummy indicating whether the match was played at home, the matchday, the minute of play the penalty was taken, the experience of both the penalty taker and the goalkeeper (quantified by the number of years the player played for a professional team), and the current match score difference. In addition, to account for player-specific abilities, we include dummy variables for all penalty takers and goalkeepers. This results in 656 covariates in total.

The parameter estimates obtained (on the logit scale) indicate that the baseline level for scoring a penalty is higher in the model's state 1 than in state 2 ($\hat{\beta}_0^{(1)} = 1.422 > -14.50 = \hat{\beta}_0^{(2)}$), thus indicating evidence for a hot shoe effect. State 1, hence, can be interpreted as a hot state, whereas state 2 refers to a cold state. In addition, with the t.p.m. estimated as

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.978 & 0.022 \\ 0.680 & 0.320 \end{pmatrix},$$

there is high persistence in state 1, i.e. in the hot state. However, when being in state 2 (cold state) switching to state 1 is most likely. Additionally, the model is slightly favoured by the AIC over a 1-state model, i.e. a standard logit model without a potential hot shoe effect ($\text{AIC}_{\text{hotshoe}} = 3664$, $\text{AIC}_{\text{1-state-model}} = 3670$). The coefficient paths of our model are shown in Figure 3. Out of the 656 covariates included in our model, only a single covariate is selected according to the AIC, namely the ability of Jean-Marie Pfaff with $\hat{\beta}_{\text{Pfaff}} = -0.0015$. The negative effect indicates that the odds for scoring a penalty decrease if Jean-Marie Pfaff is the goalkeeper of the opposing team — in fact he saved remarkable 9 out of 14 penalty kicks during his career in the Bundesliga. To further illustrate our variable selection approach, Figure 3 additionally highlights the covariates which would be selected next, namely the abilities of Gnther Herrmann (outfield player) and Rudolf Kargus (goalkeeper). As several existing studies provide evidence for a home advantage in football, we also highlight in Figure 3 the corresponding coefficient path of the dummy variable indicating whether a match was played at home (but note that it is also not selected here). For more detailed results of the application see Ötting and Groll (2019).

5 Outlook

Further research could focus on additional penalties to conduct variable selection within HMMs, such as the ridge penalty or the elastic net. In the case of multicollinearity, especially the elastic net may show a superior performance compared to the LASSO. Moreover, modifications of the standard LASSO such as the relaxed-LASSO could be considered.

Acknowledgments: We want to thank the group of researchers B. Bornkamp, A. Fritsch, L. Geppert, P. Gnändinger, K. Ickstadt, and O. Kuss for providing the German Bundesliga penalty data set.

References

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, **17**, 295 – 314.

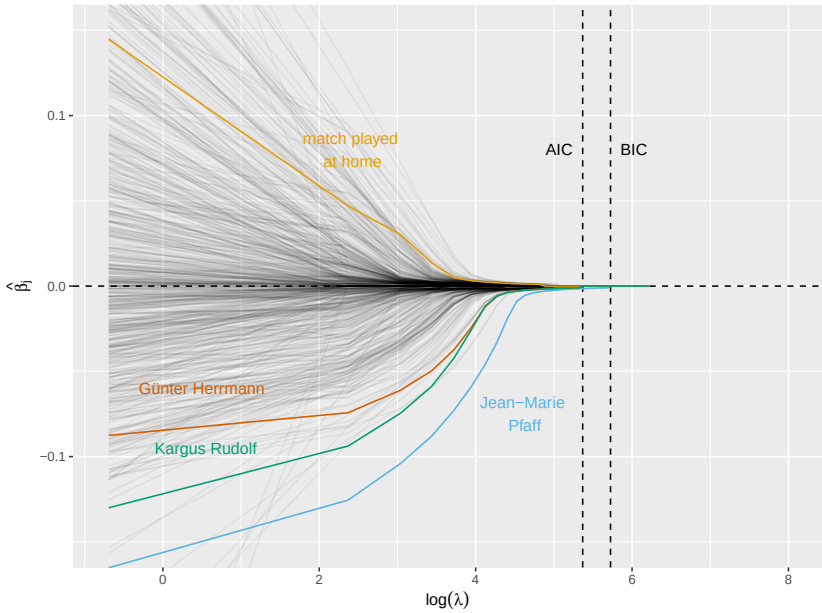


FIGURE 3. Coefficient paths of all covariates considered in the LASSO-HMM models. Dashed vertical lines indicate the penalty parameters λ as selected by AIC and BIC, respectively. For the BIC, no covariates are selected, whereas for the AIC the player-specific effect of Jean-Marie Pfaff is selected. The player-specific abilities of Günter Herrmann and Rudolf Kargus would be selected next.

Oelker, M. R. and Tutz, G. (2017). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification*, **11**, 97 – 120.

tting, M. and A. Groll (2019): A regularized hidden Markov model for analyzing the 'hot shoe' in football, *arXiv preprint* arXiv:1911.08138.

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

Zucchini, W., MacDonald, I.L., and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton: Chapman and Hall-CRC.

Intervention Analysis for INAR(1) Models

Xanthi Pedeli[1], Roland Fried[2]

[1] Athens University of Economics and Business, Greece
[2] TU Dortmund University, Germany

E-mail for correspondence: xpedeli@aueb.gr

Abstract: We study the problem of intervention effects generating various types of outliers in a first order integer valued autoregressive model with Poisson innovations. We concentrate on outliers which enter the dynamics and can be seen as effects of extraordinary events. We consider three different scenarios, namely the detection of an intervention effect of a known type at a known time, the detection of an intervention effect of unknown type at a known time and the detection of an intervention effect when the type and the time are both unknown. We develop F-tests and score tests for the first scenario. For the second and third scenarios we rely on the maximum of the different F-type or score statistics. The usefulness of the proposed approach is illustrated using simulated examples.

Keywords: Counts; Time series; Innovation outlier; Level shift; Transient shift.

1 Introduction

Time series of counts are observed in a broad variety of applications including economics, finance, epidemiology and meteorology, among others. Integer valued autoregressive (INAR) models have been introduced by McKenzie (1985) and Al-Osh and Alzaid (1987) and are widely used nowadays for this kind of data. We concentrate on the first order INAR model with Poisson innovations, denoted briefly as Poisson INAR(1),

$$Y_t = \alpha \circ Y_{t-1} + e_t,\ t \in \mathbb{N}, \tag{1}$$

where $\circ$ denotes the binomial thinning operator and (e_t) is an arrival process consisting of a sequence of independent identically distributed Poisson variables with parameter λ.
Detection of unusual events is important in any modeling framework but to the best of our knowledge it has not been investigated thoroughly in

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

the INAR framework, see e.g. Silva and Pereira (2015) and the references therein. We concentrate on outliers which enter the dynamics and can be seen as effects of extraordinary events. We aim at the detection of different types of effects including innovation outliers, transient shifts and level shifts at possibly unknown time points. More precisely, we extend model (1) as follows:

$$Y_t = \alpha \circ Y_{t-1} + e_t + \sum_{j=1}^{J} U_{t,j},\ t \in \mathbb{N}, \tag{2}$$

where J is the number of intervention effects and $(U_{t,j} : t \in \mathbb{N})$, $j = 1, \ldots, J$ are independent random variables denoting the effects of the different interventions on all time points. We assume $U_{t,j} \equiv 0$ for $t = 0, \ldots, \tau_j - 1$, and $U_{t,j} \sim Pois(\kappa_j \delta_j^{t-\tau_j})$ for $t = \tau_j, \tau_j + 1, \ldots$, with τ_j and κ_j denoting respectively the time point and size of the j-th intervention and $\delta_j \in [0, 1]$ controlling the effect of the intervention on the future of the time series after time τ_j. For $\delta_j = 1$ we get a permanent level shift starting at time τ_j, for $\delta_j = 0$ we get an innovation outlier, i.e., a single effect at time τ_j which spreads into the future according to the dynamics of the data generating process, and for $\delta_j \in (0, 1)$ we get a transient shift which decays with rate δ_j. The effect of each type of intervention on a realization of a stationary Poisson INAR(1) process is illustrated in Figure 1.

2 Test statistics

If the number of interventions J, the time points τ_j of their occurrence and the types δ_j of the interventions $j = 1, \ldots, J$ are known, then the conditional mean $E(Y_t|Y_{t-1})$ in our intervention model is linear in the remaining parameters α, λ and κ_j, $j = 1, \ldots, J$, leading to simple formulae for the conditional least squares (CLS) or conditional maximum likelihood estimates (CML). In the following, we define the F-statistic based on the residual sum of squares minimized by the CLS approach and the score test statistic based on maximization of the conditional log-likelihood function.

2.1 The *F*-statistic

The CLS estimates minimize the objective function,

$$RSS(J) = \sum_{t=2}^{n} \left[y_t - \lambda - \alpha y_{t-1} - \sum_{j=1}^{J} \kappa_j \delta_j^{t-\tau_j} I(t \geq \tau_j) \right]^2,$$

and can be calculated using simple explicit formulae and software for ordinary least squares estimation in linear models. The residual sum of squares can also be used when we want to decide whether a certain type of intervention effect is present at a given time point. A common measure for

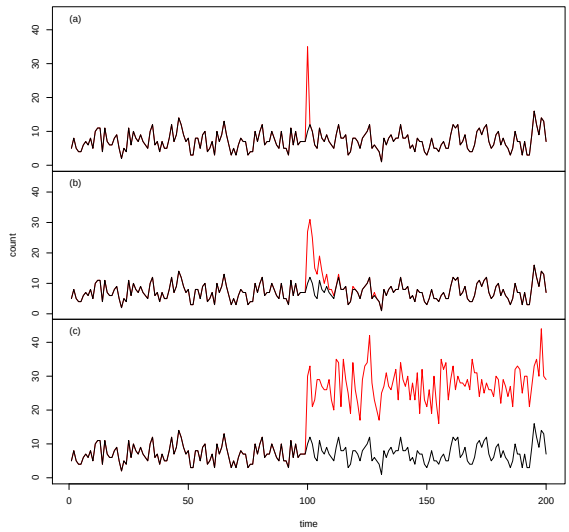


FIGURE 1. Effects of different types of outliers of size $\kappa = 20$ at time point $\tau = 100$ on a realization of a Poisson INAR(1) process generated with $\alpha = 0.3$, $\lambda = 5$ and $n = 200$. The black and red lines correspond to the clean and contaminated processes respectively, where contamination is due to (a) an innovation outlier, (b) a transient shift with $\delta = 0.8$ and (c) a level shift.

the goodness of fit of a linear model is the coefficient of determination. In case of Gaussian linear models one often prefers the F-type statistic $F = \frac{RSS(0)-RSS(1)}{RSS(1)/(n-4)}$ since it is F-distributed with 1 and $n-4$ degrees of freedom in Gaussian linear models if the simpler model without the additional (intervention) effect is correct. In our case, $n-4$ will usually be large so that such an F-distribution is close to the χ^2_1-distribution.

2.2 The score test statistic

The conditional log-likelihood function corresponding to model (2) is given by $\ell(\boldsymbol{\theta}) = \sum_{t=2}^{n} \log P(Y_t = y_t | Y_{t-1} = y_{t-1})$, where $\boldsymbol{\theta} = (\alpha, \lambda, \kappa_1, \ldots, \kappa_J)^T$. Provided that the solution of the score function $U(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$ exists, it yields the CML estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. The availability of the score function $U(\boldsymbol{\theta})$ and conditional information matrix $\mathcal{I}(\boldsymbol{\theta}) = Cov\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \middle| Y_{t-1} \right)$ allows us to define the score test statistic

$$S = U^T(\tilde{\alpha}, \tilde{\lambda}, 0)\mathcal{I}^{-1}(\tilde{\alpha}, \tilde{\lambda}, 0)U^T(\tilde{\alpha}, \tilde{\lambda}, 0), \tag{3}$$

for testing the presence of a single intervention effect ($J = 1$) of known type and time of occurrence, i.e. testing the null hypothesis $H_0 : \kappa = 0$ against the alternative $H_1 : \kappa \neq 0$. In (3), $U^T(\tilde{\alpha}, \tilde{\lambda}, 0)$ and $\mathcal{I}(\tilde{\alpha}, \tilde{\lambda}, 0)$ are the score function and conditional information matrix evaluated at the maximum likelihood estimators $(\tilde{\alpha}, \tilde{\lambda}, 0)$ computed under the null hypothesis of a clean INAR(1) process. Under H_0, (2) reduces to a stationary Poisson INAR(1) process. For such a process and under certain regularity conditions that are satisfied by the Poisson law, the conditional maximum likelihood estimates are consistent and asymptotically normal, i.e. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, \mathcal{I}^{-1}(\boldsymbol{\theta}))$. Therefore, under H_0 and as $n \to \infty$, the score statistic (3) converges to a χ^2_1-distribution and derivation of critical values for an asymptotic test of the null hypothesis of no intervention against the alternative of an intervention of certain type δ at known time τ is straightforward: we reject the null hypothesis at a given significance level a if the value of S is larger than the $(1 - a)$-quantile of the χ^2_1-distribution.

3 Some empirical results

Empirical rejection rates were obtained by analyzing 5000 time series of the same length $n \in \{100, 200\}$ for each of different INAR(1) models with $\alpha \in \{0.3, 0.6, 0.9\}$ and $\lambda \in \{2, 5\}$. Simulation results indicated that the score statistics perform better than the F-type statistics in detecting transient shifts ($\delta = 0.8$) and permanent level shifts ($\delta = 1$), especially when the INAR(1) process is characterized by strong autocorrelation ($\alpha = 0.9$). However, the F-type statistics achieve rejection rates closer to the targeted ones when the objective is to detect an innovation outlier ($\delta = 0$). The performance of both tests is little affected by the time τ of the occurrence of the intervention.

Acknowledgments: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska–Curie grant agreement no. 699980.

References

Al-Osh,M.A. and Alzaid,A.A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, **8**, 261 – 275.

McKenzie, E. (1985). Some simple models for discrete variate time series. *Journal of the American Water Resources Association*, **21**, 645 – 650.

Silva, M. E. and Pereira, I. (2015). Detection of additive outliers in Poisson INAR(1) time series. *In Mathematics of Energy and Climate Change*, **2**, 377 – 388.

Disease mapping method comparing the spatial distribution of a disease with a control disease

Oana Petrof[1], Thomas Neyens[1,2] , Maren Vranckx[1], Valerie Nuyts[3], Kristiaan Nackaerts[4] , Benoit Nemery[3], Christel Faes[1]

[1] Hasselt University, Data Science Institute (DSI), The Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt, Belgium,
[2] KU Leuven, Leuven Biostatistics and Statistical Bioinformatics Centre (L-BioStat), Department of Public Health and Primary Care, Leuven, Belgium,
[3] KU Leuven, Centre for Environment and Health, Department of Public Health and Primary Care, Leuven, Belgium
[4] KU Leuven, Department of Pneumology, University Hospital Leuven, Leuven, Belgium

E-mail for correspondence: oana.petrof@uhasselt.be

Abstract: Traditional disease mapping models are based on relating the observed number of disease cases per spatially discrete area to an expected number of cases for that area. Expected numbers are calculated by internal standardisation, which requires both accurate population numbers and disease rates per age group. Confidentiality issues or the absence of high-quality information about the characteristics of a population-at-risk can hamper those calculations. Based on methods in point process analysis, we propose the use of a case-control approach in the context of lattice data, in which an unrelated spatially unstructured disease is used as a control disease. We apply our methods to a Belgian study of mesothelioma risk, where pancreatic cancer serves as the control disease. The analysis results are in close agreement with those coming from traditional disease mapping models based on internally standardised expected counts.

Keywords: BYM model; Case-control study; Disease mapping; Mesothelioma; Standardization.

1 Introduction

The classical hierarchical models for disease mapping make use of data including the population at risk or a local number of cases "expected" under

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

some null model of disease transmission. Due to medical confidentiality, it is often difficult to obtain accurate and detailed population data (Beale et al. 2008). Census data can be used to reflect the population data of a specific region. However, countries census areas can be large or population data are not available for some countries.
The objective of this paper is to propose a disease mapping method, where a control disease is used as a proxy for the population at risk, extending the case-control methods for point-pattern data towards lattice data. In this study, interest is in mesothelioma cancer, while pancreatic cancer is used as control disease.

2 Methodology

2.1 Classical Disease Mapping Method

The response $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ represents the observed number of disease cases per areal unit throughout the study period. A Poisson model is commonly assumed to estimate the disease risk per area:

$$Y_i \sim Poisson(E_i\theta_i), \quad i = 1, ..., n, \tag{1}$$

where E_i represents the expected number of disease cases in area i and θ_i expresses the disease risk for the i^{th} area.
The expected number of cases is defined as (Waller and Gotway, 2004):

$$E_i^I = \sum_g \frac{Y_g}{N_g} N_{i,g} = \sum_g r_g N_{i,g} \tag{2}$$

where r_g is the age-specific incidence rate in the standard population. This ratio is multiplied by $N_{i,g}$ representing the population size of municipality i in age group g.

2.2 Disease Mapping with Control Disease

An approach commonly used in the context of point-pattern data is to compare the location of disease cases with that of a set of carefully selected controls for the population at risk (Kelsall and Diggle, 1998). We extend this idea to the disease mapping context.
Only the aggregated number of cases for the disease of interest ($\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$) and the number of cases for the control disease ($\mathbf{Z} = (Z_1, Z_2, ..., Z_n)$) are available. The expected number of cases for the disease of interest can be represented by

$$E_i^C = \frac{Z_i}{\sum_{j=1}^N Z_j} \left(\sum_{j=1}^N Y_j \right) = r_i^Z Y_\cdot \tag{3}$$

where r_i^Z is the rate of the control disease in area i and $Y_.$ is the total number of cases of the disease of interest.
Any disease utilized as a control disease will introduce uncertainty in the model, as it represents a sample from the population data. The calculated expected values will have a lot of uncertainty if only a small number (or no) cases of the control disease are present.
To account for the uncertainty in the estimation of the expected number E_i^C we assume that the control disease follows a multinomial distribution

$$(Z_1, ..., Z_N) \sim Multinomial(Z_., (r_1^Z, ..., r_N^Z)),$$

where $Z_.$ represents the total number of controls. We make use of the multinomial-Poisson transformation developed by Baker (1994):

$$\begin{aligned} Z_i &\sim Poisson(\lambda_i) \\ \lambda_i &\sim Gamma(0.5, 0.05), \\ r_i^Z &= \frac{\lambda_i}{\sum_{j=1}^N \lambda_j}, \end{aligned} \tag{4}$$

where the number of control cases Z_i in municipality i follows a Poisson distribution with mean λ_i. The resulting expected value is denoted as E_i^{C2}.
A conditional autoregressive convolution model proposed by Besag et al. (1991) was used to analyse and compare the three methods presented above.

3 Data analysis

Residential information about all mesothelioma and pancreatic cancer patients diagnosed between 2004 and 2015 is available (Belgian Cancer Registry) as well as information about the population distribution in all areas during the period 2009-2015.
All methods show a cluster of municipalities in the Central Northern part of Flanders, and in the Central Eastern part of the country (Figure 1). However, on the center panel some areas with increased risk are more dispersed over the country, as compared to the classical method. The right hand side panel results show much less variability as compared to the to the model in which the expected number is considered to be a fixed value (center panel). By incorporating more variability for the expected values, a smoothed map is observed for the new method (right panel), leading to a more accurate approximation of the Poisson convolution model results.

4 Conclusion

In this paper, we have proposed a method similar to methods used in point-pattern data (Diggle et al. 2000), in which the incidence of the disease of

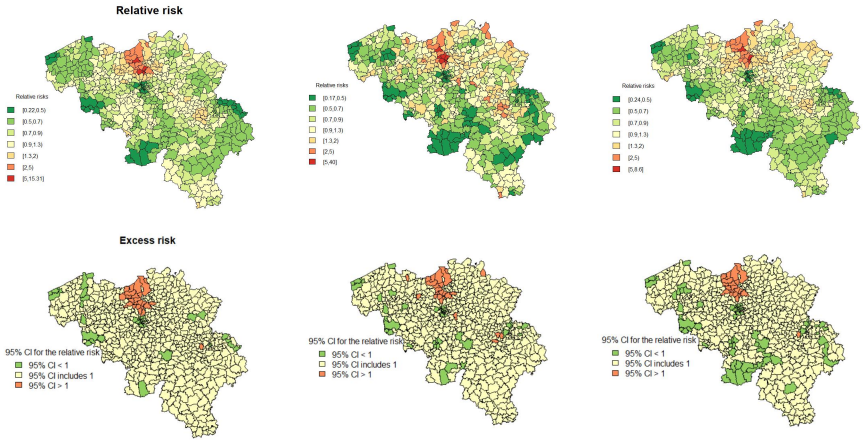


FIGURE 1. *Map of the relative risks and excess risk for the Poisson Convolution model. Left panel: using indirect standardized number; Center panel: using control-disease's standardized number; Right panel: using control-disease's standardized number accounting for uncertainty.*

interest is compared to the incidence of a control disease, in the context of lattice data. Allowing for extra variability through the use of a distribution for the expected values, leads to a control-disease approach used for a Poisson convolution model which had similar results with the classical methodology.

References

Baker, S. G. (1994). The multinomial-Poisson transformation. *Statistician*, **43**, 495 – 504.

Beale, L., Abellan, J. J., Hodgson, S., and Jarup, L. (2008). Methodologic Issues and Approaches to Spatial Epidemiology. *Environ Health Perspect.*, **116**, 1105 – 1110.

Besag, J., York, J., and Molli, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math.*, **43**, 1 – 20.

Diggle, P. J., Morris, S. E., and Wakefield, J. C. (2000). Point-source modelling using matched case-control data. *Biostatistics*, **1**, 89 – 105.

Kelsall, J. E., and Diggle, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *J R Stat Soc Ser C Appl Stat*, **47**, 559 – 573.

Waller, L. A. and Gotway, C. A. (2004). *Applied spatial statistics for public health data.* (Vol. 368). John Wiley & Sons.

Flexible estimation of the state dwell-time distribution in hidden semi-Markov models

Jennifer Pohle[1], Timo Adam[1,2], Roland Langrock[1], Larissa Beumer[3]

[1] Bielefeld University, Bielefeld, Germany
[2] University of St Andrews, St Andrews, UK
[3] Aarhus University, Aarhus, Denmark

E-mail for correspondence: jennifer.pohle@uni-bielefeld.de

Abstract: Hidden semi-Markov models (HSMMs) generalise hidden Markov models by explicitly modelling the time spent in a state, the so-called dwell-time distribution, using some distribution on the positive integers, e.g. the (shifted) Poisson or the negative binomial. In this paper, we propose a penalised maximum likelihood approach for fitting HSMMs without the need to specify a distributional assumption for the state dwell times. The feasibility and potential usefulness of the approach is illustrated using muskox animal movement data.

Keywords: Difference penalty; Penalisation; Smoothing; Time series modelling.

1 Introduction

Hidden Markov models (HMMs) are flexible time series models for observations driven by an underlying latent state sequence. For mathematical convenience, the state sequence is usually assumed to be a first-order Markov chain. This, however, implies that the state dwell time, i.e. the number of consecutive time points spent in a given state, follows a geometric distribution. Hidden semi-Markov models (HSMMs) overcome this limitation by allowing for an arbitrary dwell-time distribution. Within HSMMs, the dwell times are then usually modelled using standard parametric distributions, e.g. the Poisson or the negative binomial, which corresponds to a restrictive assumption on the shape of the dwell-time distribution, and hence on the way the state process evolves over time. To avoid such restrictive assumptions, we develop a fully data-driven penalised maximum likelihood

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

approach for estimating HSMMs without prior specification of the class of distributions used for the state dwell times.

2 HSMMs with flexible dwell-time distributions

An HSMM is a doubly stochastic process comprising a latent N-state semi-Markov process $\{S_t\}_{t=1}^T$ and an observed state-dependent process $\{X_t\}_{t=1}^T$. At each time point, X_t is assumed to be generated by one out of N distributions $f_i(x_t)$, $i = 1, \ldots, N$, as selected by the current state S_t. Given the states, the observations are assumed to be conditionally independent of each other and past states. The underlying semi-Markov chain is characterised by two components: (i) whenever the chain enters a new state i at some time point, a draw from a dwell-time distribution on $\{1, 2, \ldots\}$, defined by its probability mass function d_i, determines the number of consecutive time points the chain spends in that state; (ii) state switches are determined by the conditional transition probabilities $\omega_{ij} = \Pr(S_t = j | S_{t-1} = i, S_t \neq i)$, summarised in the $N \times N$ matrix $\mathbf{\Omega}$. Thus, an HSMM is completely specified by the vector $\boldsymbol{\theta}$ comprising the parameters defining d_i and $f_i(x_t)$, for $i = 1, \ldots, N$, and ω_{ij}, for $i, j = 1, \ldots, N$, $i \neq j$. In case that all state dwell times are geometrically distributed, the HSMM reduces to the special case of an HMM.

Letting $d_i(r)$ denote the probability of a dwell time of length r in state i, we assign a parameter π_{ir} to each individual probability $d_i(r)$ for $r \in \{1, 2, \ldots, R_i\}$, where the upper boundary R_i needs to be chosen large enough to capture the main support of the dwell-time distribution. To further allow for dwell times $r > R_i$, a geometric tail is added:

$$d_i(r) = \begin{cases} \pi_{ir} & \text{if } 0 < r \leq R_i; \\ \pi_{iR_i} \left(\frac{1 - \sum_{r=1}^{R_i} \pi_{ir}}{1 - \sum_{r=1}^{R_i - 1} \pi_{ir}} \right)^{r - R_i} & \text{if } r > R_i, \end{cases}$$

with $0 < \pi_{ir} < 1$ and $\sum_{r=1}^{R_i} \pi_{ir} < 1$. Using a state space expansion and a suitable block structure in the resulting enlarged transition probability matrix, this HSMM can be represented exactly as an HMM (Langrock and Zucchini, 2001, Zucchini *et al.*, 2016), with the resulting state space of dimension $\sum_{i=1}^N R_i$. This trick renders the computational machinery available for HMMs applicable also to HSMMs, including numerical maximisation of the log-likelihood $\ell(\boldsymbol{\theta} | x_1, \ldots, x_t)$, which is evaluated using the forward algorithm. To avoid overfitting with respect to the probability mass functions, we enforce smoothness by penalising the squared third-order differences $\Delta^3 \pi_{ir}$ of adjacent state dwell-time probability parameters:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \ell(\boldsymbol{\theta} | x_1, \ldots, x_T) - \sum_{i=1}^{N} \lambda_i \sum_{r=4}^{R_i} (\Delta^3 \pi_{ir})^2.$$

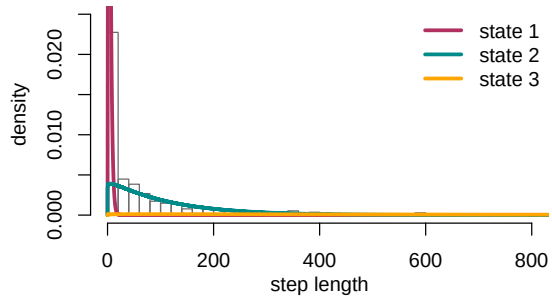


FIGURE 1. Estimated state-dependent gamma distributions for the 3-state HSMM, weighted according to the proportion of time the states are active.

The smoothing parameters λ_i control the balance between goodness-of-fit and smoothness of the $d_i(r)$ functions, and can be chosen for each state individually. For a data-driven selection of λ_i, cross-validation can be used.

3 Case study: modelling muskox movement

We illustrate our approach using $T = 1440$ hourly observed step lengths of a muskox in Greenland. Based on previous work that found the main behavioural modes to be resting, foraging and relocating (Pohle *et al.*, 2017, Beumer *et al.*, 2020), we fit 3–state HSMMs with state-dependent gamma distributions. The parameters are estimated by numerically maximising the penalised log-likelihood using the optimisation routine `nlm` in `R`. For simplicity, we use the same smoothing parameter λ for each state, testing $\lambda = 0, 10, 1000$, while fixing $R_1 = R_2 = R_3 = 10$.
Figure 1 displays the estimated state-dependent gamma distributions (for $\lambda = 1000$), which can reasonably be interpreted as corresponding roughly to resting (state 1), foraging (state 2) and relocating (state 3). The estimated dwell-time probability mass functions are displayed in Figure 2, for states 1-3 and the different values considered for λ. Irrespective of the choice of λ, the dwell-time distributions of the latent state process clearly differ from a geometric distribution, which suggests that a basic HMM would not correctly represent the dynamics in the state process. The necessity of penalisation becomes clear for example in view of $\hat{d}_3(r)$, the dwell-time distribution estimated for state 3: when increasing λ the distribution becomes smoother, and in particular the gaps in the probability mass function, as obtained when not penalising ($\lambda = 0$; top right panel in Figure 2), are filled due to the enforced smoothness.

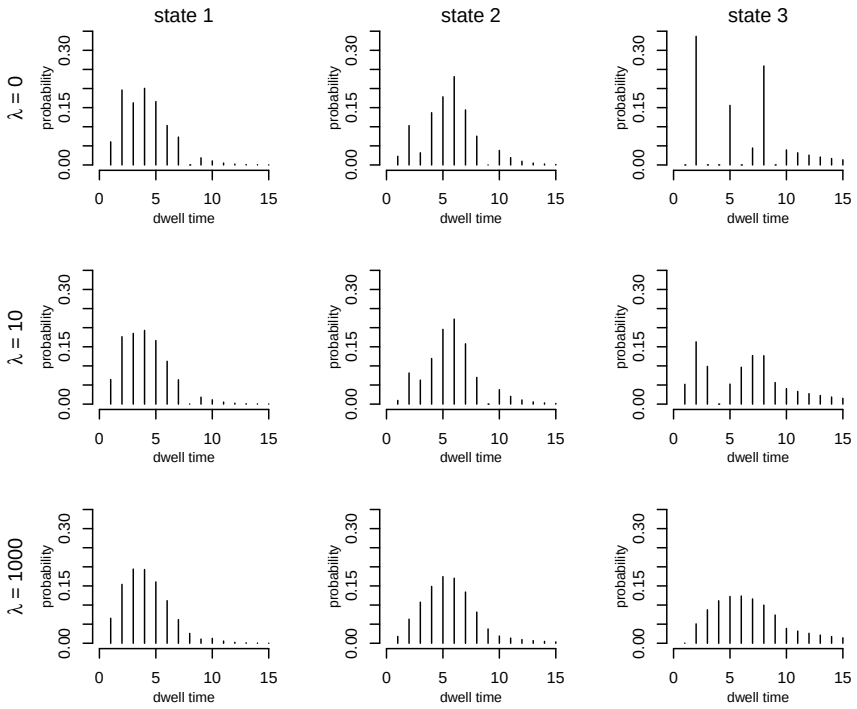


FIGURE 2. Estimated dwell-time probability mass functions for the 3-state HSMM, for states 1-3 and the different λ's considered.

4 Conclusions

As the state process is unobserved, it is often unclear how to select a model that appropriately reflects the underlying state dynamics. Our proposed penalisation estimation approach can be used as an exploratory tool to investigate the unknown shapes of the states' dwell-time distributions. The method can either be used for direct modelling purposes, or as a basis for subsequent modelling choices, for example in order to decide whether an HMM would be appropriate for the data at hand, or what distributional assumption may be adequate within a conventional HSMM.

References

Beumer, L.T., Pohle, J., Schmidt, N.M., Chimienti, M., Desforges, J.-P., Hansen, L.H., Langrock, R., Pedersen, S.H., Stelvig, M. and van Beest, F.M. (2020). An application of upscaled optimal foraging theory using hidden Markov modelling: year-round behavioural variation in a large arctic herbivore. *Movement Ecology*, **8**, doi: https://doi.org/10.1186/s40462-020-00213-x.

Langrock, R. and Zucchini, W. (2011). Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics and Data Analysis*, **55**, 715 – 724.

Pohle, J., Langrock, R., van Beest, F.M. and Schmidt, N.M. (2017). Selecting the number of states in hidden Markov models pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, **22**, 270 – 293.

Zucchini, W., MacDonald, I.L. and Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R*, 2nd Edition. Boca Raton: Chapman & Hall/CRC.

Growth curves for multiple-output response variables via Bayesian quantile regression models

Bruno Santos[1], Agatha Rodrigues[1], Thomas Kneib[2]

[1] Universidade Federal do Espírito Santo, Brazil
[2] Georg-August-Universität Göttingen, Germany

E-mail for correspondence: bruno.santos.31@ufes.br

Abstract: Reference fetal growth curves play an important role in identifying fetal growth restriction, macrosomia and other fetal malformations. This is verified based on percentiles of some biometric measurements at a specific gestational age using obstetric ultrasound. As an example, the diagnosis of microcephaly is based on a biparietal diameter smaller than the 10th percentile based on the reference curve. In practice, each biometric measurement reference curve is constructed independently of other measurements, even if they are correlated and some information about dependencies among them might be lost. Here we use these measurements to define growth curves modelling jointly more than one measurement. We consider structured additive quantile regression models for multiple-output response variables, where we are able to specify a nonlinear effect of time. We define a Markov Chain Monte Carlo (MCMC) procedure for model estimation, using ideas previously discussed in the literature. We examine four different ultrasound measurements and we show how one can retrieve more information when modelling these response variables jointly instead of individually. We illustrate the method with data from pregnancies from the University Hospital of the University of São Paulo (HU / USP) in the city of São Paulo, Brazil.

Keywords: Bayesian quantile regression; Multiple-output response variable; Growth curves.

1 Growth curves and Bayesian quantile regression models

A proper assessment of fetal growth is important to identify irregular growth that is related to fetal malformations and/or disease of the mother.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

This evaluation is usually made observing references growth curves which are build exactly for these purposes, where percentiles of a biometric measurement, for instance fetal biparietal diameter, is estimated for a specific population. The fetus biometric measures are obtained from ultrasonographic examinations measured throughout the pregnancy. Fujita et al. (2020) used four different measures to estimate the weight of the fetus during the pregnancy and then considered a mixed model to obtain growth curves taking into consideration other covariates to control for the heterogeneity in the data.

A possible strategy one can use to estimate these conditional percentiles, or more broadly conditional quantiles, of these measures is connected to quantile regression models. In these models one can write the conditional quantiles as

$$Q_Y(\tau|\boldsymbol{X} = \boldsymbol{x}) = f_\tau(z) + \boldsymbol{x}^{'}\boldsymbol{\beta}_\tau,$$

where f_τ is a nonlinear function related to covariate z and $\boldsymbol{x}^{'}\boldsymbol{\beta}_\tau$ is the typical linear quantile regression part. This nonlinear function could be used to estimate the time effect, for instance. One approach for multiple-output response variables was proposed by Hallin et al. (2010).

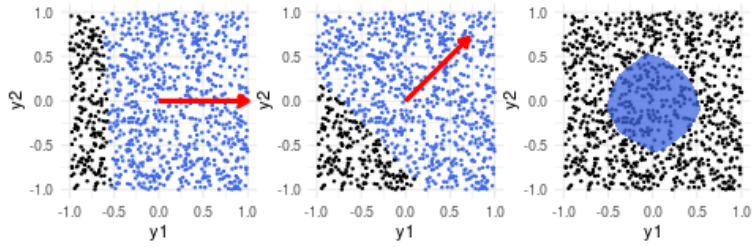


FIGURE 1. First two plots show examples of the separation by the τth directional quantile hyperplane, when $\tau = 0.2$ and the direction is denoted by the red arrow. The last plot show the quantile region obtained given the estimated models in 99 directions.

In Figure 1, we have the representation of a bivariate response variable where both components are uniformly distributed between -1 and 1. The proposal by Hallin et al. (2010) is based on directions represented by the red arrows, where the directional multiple-output quantile regression model divides the spaces in two halfspaces, as in the black and blue points. By construction, this hyperplane observes two subgradient conditions, though we focus here only on the first. This condition is related to quantile regression models, as the probability of the response variable belonging to the space represented by the black points is equal to τ. For instance, in this example where $\tau = 0.2$ one would expect 20% of the points to be denoted as black for all directions. If one defines multiple directions in this two-dimensional example and checks the intersection of all the blue points, we arrive at the

quantile region of the last plot in Figure 1. Given the importance of the chosen directions for our model, we discuss different methods for defining these in the next subsection. Regarding the quantile regions shown, we consider those to build our growth curves in the next section. See Hallin et al. (2010) and Santos and Kneib (2020) to check on how to estimate these hyperplanes using a frequentist and Bayesian approach, respectively.

1.1 Defining the directions

In Figure 2 we show the result of defining 512 directions based on two different ideas. The first one considered by Santos and Kneib (2020) splits the intervals [-1, 1] in equidistant points, that depend on how many directions one wants to estimate. For instance, if we define 8 marginal points then we arrive at $8^3 = 512$ models, because we consider all possible combinations of points, defininining vectors in $[-1, 1]^3$. Each vector is then divided by its norm, so we have unit vectors for directions. The result using this method is shown in black points in the left of Figure 2.

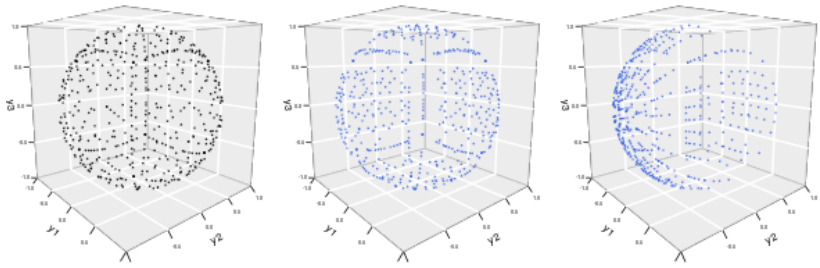


FIGURE 2. Different directions chosen for estimation given marginal quantities for each dimension: (left) equidistant points marginally; (middle) standard normally distributed in each marginal; (middle) standard normally distributed in two marginal, while Y_2 is exponentially distributed with mean 1.

The second method takes into account the marginal quantiles to define points in the interval [-1, 1]. The quantiles $(q_1, q_2, \ldots, q_p)$ are calculated for each dimension, where q_1 is the minimum and q_p is the maximum, while p is the number of marginal points. These quantiles are then scaled into the interval [-1, 1] and the same rules applied to the previous idea are used in order to define these new directions. In Figure 2 we illustrate two cases with this method in blue dots. The middle plot shows the case when each marginal has a standard normal distribution and the plot in the right shows when one of the marginal a standard exponential distribution, while the others are still normally distributed. This approach is closely connected to the idea of selecting knots when estimating nonlinear functions.

2 Application to gestational data

Here we consider the data used in Fujita et al. (2020), where there are 1445 ultrasonographic examinations of 434 pregnancies at 12-42 gestational weeks, where the babies were born between July 1, 2014 and December 31, 2017, at the University Hospital of University of São Paulo, Brazil.
We consider the biometric measurements: femur length (F), head circumference (HC), abdominal circumference (AC), biparietal diameter (BPD). We use as covariates the gestational age, fetal sex and mother's height and mother's body mass index. For each direction of interest, the nonlinear effect of gestational age was modelled with cubic p-splines functions with 20 equidistant knots. All results were obtained with chain size of 110,000 samples, after discarding the first 10,000 draws and recording every 100th value. For this two dimensional example, we used 99 directions to calculate the quantile regions.

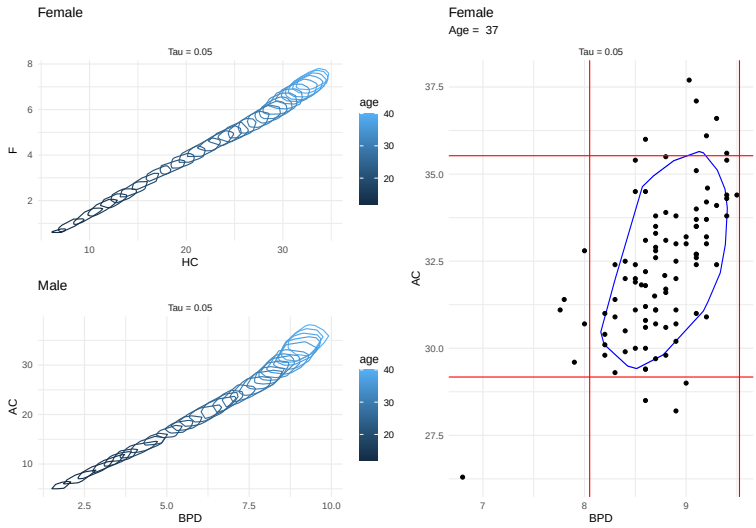


FIGURE 3. Left side: quantile contours for the combination of two measurements, FEMUR and HC for female fetuses on the top, with AC and BPD for male fetuses on the bottom, $\tau = 0.05$; Right side: quantile contour for $\tau = 0.05$, for female fetuses, 37 weeks of gestational age, while the red lines represent the marginal conditional quantiles ($\tau = 0.05, 0.95$) for the same variables, but considering a linear function on age.

Initially, we investigated all possible $\binom{4}{2}$ models using these measurements. On the left side of Figure 3, one can check the quantile contours for two cases, "F-HC" and "AC-BPD", where it is noticeable the nonlinear variation given gestational age. We plot the values for the quantile contours for values between 12 and 40 weeks of pregnancy, in intervals of 1 week.

The jumps seem to be bigger in the beginning of the pregnancy, while the variance appear to be larger closer to 40 weeks of pregnancy. There are also differences in the shape of the plots when we compare the plots in the top and in the bottom. This shows how it is important to take into consideration all the different types of correlation between the measurements.

On the right side of Figure 3, we have the quantile contour for one specific age, 37 weeks, a female fetus, for measures of BPD and AC. We also plot the marginal conditional quantiles based on a quantile regression model, with covariates age and fetal sex, for $\tau = \{0.05, 0.95\}$, with a linear function on age. It is easy to see how the quantile contour is able to capture the correlation between the measurements way more naturally. This result motivates the investigation of how these contours could be used to check for fetal growth restrictions, instead of estimating the weight based on those measurements and then assessing some form of conditional model, as in Fujita et al. (2020).

We are also able to obtain model estimates when considering the four measurements altogether as the response variable. For this case, for the directions we use the method defined in Section 1.1 that take into account marginal quantiles of each dimension, with 5 marginal points. Though we are unable to visualize the quantile regions as in Santos and Kneib (2020), given this four dimensional problem, we can still check which observations are inside or outside these regions. Then we can summarize some of this information to showcase interesting conclusions about this model.

For instance, if one observes the right side plot in Figure 3 there are a few observations, which would not be deemed to have an atypical value, given their gestational age and the fact that the fetal sex is female. Nevertheless, when examining its joint distribution of BPD and AC, then we would identify these observations being outside the quantile region.

A similar experiment can be done after marking all observation that are outside the quantile region in the four dimensional case. After that we can compare their respective conditional position, given its fetal sex and gestational age, for instance. We could also add mother's height and body mass index, that were also considered in the model, but for the sake of simplicity these are left out initially. For this illustration, we group gestational age in four approximately equally sized groups, based on its respective quantiles. Then for each combination of gestational age group and fetal sex, we order each measurement placing all observations in one of 5 groups, $\{0-20\%, \ldots, 80\%-100\%\}$. In Figure 4 we can compare then the information for observations placed outside the quantile region.

We can check that the regions where we find more observations outside the quantile region are both extremes, which is not surprising. Though lower values for each measurement are more likely to be classified in this way. Similarly to what we had seen using the two dimensional quantile region, here there are also points more center located in one of the dimensions, but still outside the quantile region.

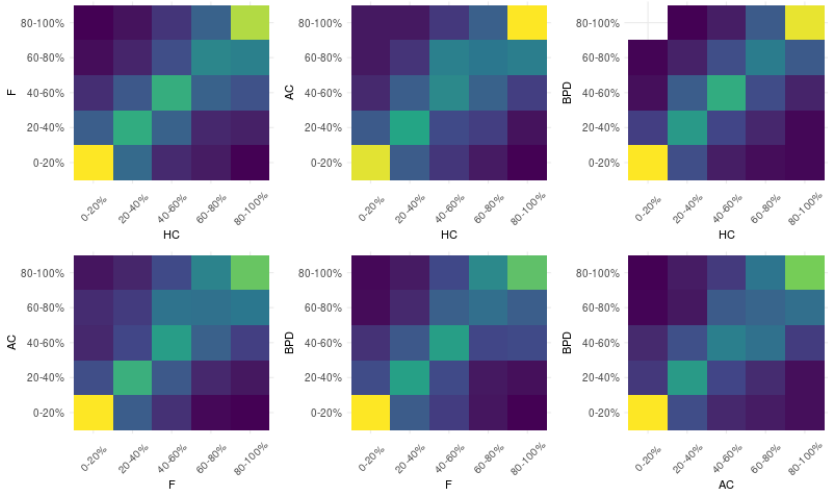


FIGURE 4. Conditional bivariate densities given fetal sex and gestational age groups for observations classified as being outside the quantile region for $\tau = 0.05$. Brightness levels for each tile represents its respective density, where higher brightness is related to a higher density.

3 Final remarks

In this paper we used Bayesian quantile regression models for multiple-output response variables to define fetal growth curves. We show how these models can better capture the correlation between the measurements in the fetus, instead of considering the conditional quantiles marginally for each measurement. One important advantage of this method is that we are able to study the variation of all these measurements without the need to assume a probability distribution in this four dimensionsal setting.

References

Fujita, M.M., Francisco, R.P.V., Rodrigues, A.S., Zugaib, M. (2020). Longitudinal study of individually adjusted fetal growth. *International Journal of Gynecology and Obstetrics*, **148**, 35 – 40.

Hallin, M., Paindaveine, D., Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: From l1 optimization to halfspace depth. *The Annals of Statistics*, **38**, 635 – 669.

Santos, B. and Kneib, T. (2020). Noncrossing structured additive multiple-output Bayesian quantile regression models. *Statistics and Computing*, DOI: 10.1007/s11222-020-09925-x.

Multivariate Ordinal Random Effects Models Including Subject and Group Specific Response Style Effects

Gunther Schauberger[1], Gerhard Tutz[2]

[1] Chair of Epidemiology, Department of Sport and Health Sciences, Technical University of Munich, Germany
[2] Department of Statistics, LMU Munich, Germany

E-mail for correspondence: gunther.schauberger@tum.de

Abstract: Common random effects models for repeated ordinal measurements account for the heterogeneity in the population by including subject-specific intercepts or variable effects. They do not account for the heterogeneity in answering tendencies. Extended models are proposed that account for the tendency to choosing extreme or middle responses where location effects as well as the tendency to extreme or middle responses are modeled as functions of explanatory variables. An example demonstrates the applicability of the method.

Keywords: multivariate ordinal response; random effects models; response styles.

1 Introduction

The cumulative model and the adjacent-categories model are popular models for univariate ordinal responses. In the following, we present the adjacent categories model for univariate and multivariate ordinal response as the basis for the newly proposed model that allows to account for response styles. For the sake of brevity we abstain from presenting the cumulative model and the corresponding extensions. Also, we restrict our presentation in Section 2 to the more common situation of an odd number of categories in the response variables.
The adjacent-categories model has the form

$$P(Y_i = r+1 | Y_i \in \{r, r+1\}, \boldsymbol{x}_i) = F(\gamma_{0r} + \boldsymbol{x}_i^T \boldsymbol{\gamma}),$$

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

where $F(.)$ is a distribution function. For the logistic distribution function one obtains

$$\log\left(\frac{P(Y_i = r+1)}{P(Y_i = r)}\right) = \gamma_{0r} + \boldsymbol{x}_i^T\boldsymbol{\gamma}. \quad (1)$$

Random effects models aim at explicitly modeling the heterogeneity of clustered responses. A cluster can be any statistical unit for which repeated measurements are available. In our applications a cluster typically refers to a person and repeated measurements refer to responses on a set of items. For such clustered data let the ordinal response $Y_{it} \in \{1, \ldots, k\}$ denote measurement t in cluster $i, i = 1, \ldots, n, t = 1, \ldots, T_i$. The simplest random effects model is a model that includes random intercepts only. Extending model (1) the inclusion of random intercepts gives

$$\log\left(\frac{P(Y_{it} = r+1)}{P(Y_{it} = r)}\right) = \gamma_{0r} + \boldsymbol{x}_{it}^T\boldsymbol{\gamma} + b_i$$

for the adjacent categories model where $b_i \sim N(0, \sigma^2)$ represents a random intercept for person i.

2 Accounting for Response Styles

In the adjacent categories model, the intercepts $\gamma_{0t1} \ldots \gamma_{0t,k-1}$ can be seen as threshold parameters. The threshold parameters determine the basic preference for specific categories. This property will be used in the following to model the subject-specific tendencies to choose specific categories. The main idea of the newly proposed model is to increase or decrease the distance between thresholds for specific persons with a centering at the middle category. In the predictor $\eta_{itr} = \gamma_{0tr} + \boldsymbol{x}_{it}^T\boldsymbol{\gamma} + b_i$ for the $t-$th variable we propose to replace the threshold γ_{0tr} by $\gamma_{0tr} + (k/2 - r)a_i$ where a_i is a subject-specific parameter. It is seen that the difference between adjacent linear predictors gives $\eta_{itr} - \eta_{it,r-1} = \gamma_{0tr} - \gamma_{0t,r-1} - a_i$, i.e., the difference between adjacent predictors changes by a_i. If a_i is positive the difference decreases, if it is negative the difference increases.

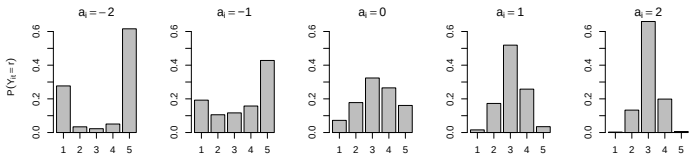


FIGURE 1. Probabilities of single categories (for an example with $k = 5$) depending on different values of response style parameter a_i.

Figure 1 illustrates how different values of the parameter a_i affect the probabilities of the single response categories. Positive values of a_i increase the

probabilities of the middle categories while negative values increase the probabilities of the extreme categories. For illustration let us consider the case $k = 5$ where one obtains the following thresholds

$$\gamma_{0t1} + 1.5a_i \qquad \gamma_{0t2} + 0.5a_i \qquad \gamma_{0t3} - 0.5a_i \qquad \gamma_{0t4} - 1.5a_i$$

The effect of explanatory variables is included by extending the response style effect a_i to the response style term $a_i + \boldsymbol{z}_i^T\boldsymbol{\alpha}$. Then, in the adjacent categories representation one obtains

$$\eta_{itr} = \gamma_{0tr} + (k/2 - r)(a_i + \boldsymbol{z}_i^T\boldsymbol{\alpha}) + \boldsymbol{x}_i^T\boldsymbol{\gamma} + b_i.$$

From a psychometric point of view the model can be seen as a generalization of the extended partial credit model proposed by Tutz, Schauberger and Berger (2018). In our proposal, their model is extended by covariate location effects $\boldsymbol{x}_i^T\boldsymbol{\gamma}$ and covariate response style effects $\boldsymbol{z}_i^T\boldsymbol{\alpha}$. The variables $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ can be distinct, overlapping, or identical.

The model contains two random effects, the subject-specific intercept b_i and the subject-specific response style effect a_i where a bivariate normal distribution $(b_i, a_i) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ is assumed. The diagonals of the matrix $\boldsymbol{\Sigma}$ contain the variance of the random intercepts σ_b^2 and of the response style parameters σ_a^2, the off diagonals are the covariances cov_{ba} between intercepts and the response style.

3 Application to pre-election data

The method is applied to data from the German Longitudinal Election Study (GLES) (Roteutscher et al., 2017). The participants were asked: "How afraid are you due to the ..."

1. refugee crisis?
2. global climate change?
3. international terrorism?
4. globalization?
5. political developments in Turkey?
6. use of nuclear energy?

The answers were measured on Likert scales from 1 (not afraid at all) to 7 (very afraid). As explanatory variables in the model we used *Abitur* (1: Abitur/A levels; 0: else), *Age*, *EastWest* (1: East Germany/former GDR; 0: West Germany/former FRG), *Gender* (1: female; 0: male) and *Unemployment* (1: currently unemployed; 0: else). In order to make the effect sizes easier to compare all variables were standardized before the respective analyses.

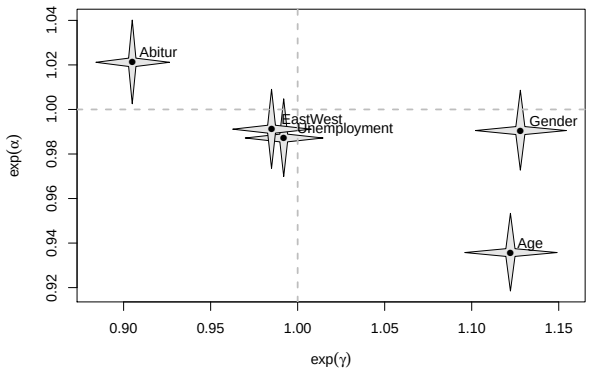


FIGURE 2. (Exponential) effects of explanatory variables in GLES data and 95% confidence intervals for location effects γ and response style effects α.

Figure 2 gives a visualization of the estimated effects and confidence intervals. It shows the exponentials of the covariate effects both for the location effects γ (abscissa) and the response style effects α (ordinate) together with the respective 95% confidence intervals.
The location effects for Abitur, Age and Gender and the response style effects for Age and Abitur are significant. According to these estimates, the overall level of political fears is increased with increasing age and for women in comparison to men while people with Abitur tend to have a lower level of fears than other respondents. On the other hand, with growing age people have an increasing tendency toward extreme categories while people with Abitur show a (slight) tendency towards middle categories.
The random effects components are seen from the estimated (co-)variance matrix

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.166 & -0.002 \\ -0.002 & 0.077 \end{pmatrix}. \tag{2}$$

There appears to be no strong correlation between the random location effects and the random response style effects.

References

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Weßels et al. (2017). Pre-election cross section (GLES 2017). GESIS Data Archive, Cologne, ZA6800 Data file Version 2.0.0.

Tutz, G., Schauberger, G., and Berger, M. (2018). Response styles in the partial credit model. *Applied Psychological Measurement*, **42(6)**, 407 – 427

Intensity Estimation on Geometric Networks with Penalized Splines

Marc Schneble[1], Göran Kauermann[1]

[1] Department of Statistics, LMU Munich, Germany

E-mail for correspondence: marc.schneble@stat.uni-muenchen.de

Abstract: In this article we consider so called geometric networks. Typical examples are road networks or other infrastructure networks. We observe network based point processes and our task is to estimate the intensity (or density) of the processes. Available routines that tackle this problem are commonly based on kernel smoothing methods. However, kernel based estimation in general exhibits some drawbacks such as suffering from boundary effects and the locality of the smoother. In an Euclidean space, the disadvantages of kernel methods can be overcome by using penalized spline smoothing. We here extend penalized spline smoothing towards smooth intensity estimation on geometric networks and apply the approach to both, simulated and real world data. The results show that penalized spline based intensity estimation outperforms kernel based methods.

Keywords: Intensity Estimation; Generalized Additive Models; Geometric Networks; Penalized Splines; Stochastic Point Processes.

1 Introduction

Okabe et al. (2009) introduced equal-split (dis-)continuous kernel density estimation on geometric networks, which was the first approach that respects the network geometry around the network's vertices. The idea is to split the mass of the kernel functions equally across all other segments that depart from a vertex when approaching this vertex from one side. The estimation procedure, including automatic bandwidth selection, is implemented in the R package `spatstat` (Baddeley et al, 2015). If the data is located on an Euclidean space, Eilers and Marx (1996) propose to estimate the density by making use of penalized splines. In this article we perform intensity estimation on geometric networks by extending the penalized spline approach to work on geometric networks and compare the results with a kernel based method.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Notation and Problem

Let $\boldsymbol{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_W\}$ be a set of vectors, which we denote as vertices, with $\boldsymbol{v}_i \in \mathbb{R}^q$ for $i = 1, \ldots, W$ and $\boldsymbol{E} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_M\}$ be a set of line segments $\boldsymbol{e}_m$ for $m = 1, \ldots, M$ with each $\boldsymbol{e}_m \subset \mathbb{R}^q$ being the connection between two vertices $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$. Furthermore, $\deg(\boldsymbol{v})$ denotes the count of segments which have an endpoint equal to a vertex $\boldsymbol{v}$. For simplicity of notation, we generally assume that $\boldsymbol{e}_m$ is a straight line such that $\boldsymbol{e}_m = \{\boldsymbol{v}_i + (1-t)\boldsymbol{v_j} \mid 0 \leq t \leq 1\}$ with length $d_m = |\boldsymbol{e}_m| = ||\boldsymbol{v}_i - \boldsymbol{v}_j||_2$, where $||\cdot||$ denotes the Euclidean distance. We now define the geometric network $\boldsymbol{L}$ through the set of line segments $\boldsymbol{E}$ by $\boldsymbol{L} = \bigcup_{m=1}^{M} \boldsymbol{e}_m \subset \mathbb{R}^q$. The lengths d_m imply a metric $d_{\boldsymbol{L}} : \boldsymbol{L} \times \boldsymbol{L} \to [0, \infty)$ on $\boldsymbol{L}$ and with $[\boldsymbol{z}_1; \boldsymbol{z}_2] \subset \boldsymbol{L}$ or with $(\boldsymbol{z}_1; \boldsymbol{z}_2) \subset \boldsymbol{L}$ we denote the path between $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. Let now $\mathcal{X}$ be a stochastic point process on the geometric network $\boldsymbol{L}$ with continuous intensity $\varphi_{\mathcal{X}} : \boldsymbol{L} \to [0, \infty)$. The expected number of points in a set $\boldsymbol{K} \subset \boldsymbol{L}$ is then defined through $\int_{\boldsymbol{K}} \varphi_{\mathcal{X}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \sum_{m=1}^{M} \int_{\boldsymbol{K} \subset \boldsymbol{e}_m} \varphi_{\mathcal{X}}(\boldsymbol{z}) \mathrm{d}\,_{|m}\boldsymbol{z}$, where $\mathrm{d}\,_{|m}\boldsymbol{z}$ denotes integration with respect to $\boldsymbol{e}_m$. Our aim is to estimate the intensity of the point process $\mathcal{X}$ on $\boldsymbol{L}$ given that we observe realizations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of this process.

3 Methodology

In this section we show how to extend the penalized spline estimation approach of Eilers and Marx (1996) to allow for intensity estimation on geometric networks. For simplicity of presentation, we restrict ourselves to linear B-spline bases. Such a basis on a geometric network can be constructed straightforwardly from the well-known one-dimensional setting.

3.1 Linear B-splines on a Network

On every line $\boldsymbol{e}_m$, which has endpoints $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, we specify an equidistant sequence of I_m knots $\boldsymbol{v}_i = \boldsymbol{\tau}_{m,1}, \ldots, \boldsymbol{\tau}_{m,I_m} = \boldsymbol{v}_j$ with $\boldsymbol{\tau}_{m,k} \in \boldsymbol{e}_m$ for $k = 1, \ldots, I_m$, where $d_{\boldsymbol{L}}(\boldsymbol{\tau}_{m,k}, \boldsymbol{\tau}_{m,k-1}) = \delta_m$. We choose the δ_m to be rather small and about the same size and construct $J_m = I_m - 2 \geq 1$ linear B-splines $B_{m,1}, \ldots, B_{m,J_m}$, which are defined by

$$B_{m,k}(\boldsymbol{z}) = \frac{d_{\boldsymbol{L}}(\boldsymbol{z}, \boldsymbol{\tau}_{m,k})}{\delta_m} I_{[\boldsymbol{\tau}_{m,k}, \boldsymbol{\tau}_{m,k+1})}(\boldsymbol{z}) + \frac{d_{\boldsymbol{L}}(\boldsymbol{\tau}_{m,k+2}, \boldsymbol{z})}{\delta_m} I_{[\boldsymbol{\tau}_{m,k+1}, \boldsymbol{\tau}_{m,k+2})}(\boldsymbol{z}) \quad (1)$$

for $\boldsymbol{z} \in \boldsymbol{L}, m = 1, \ldots, M$ and $k = 1, \ldots, J_m$. Furthermore, we construct a single B-spline around each vertex $\boldsymbol{v}_i \in \boldsymbol{V}$. Therefore, we numerate the $\deg(\boldsymbol{v}_i)$ segments which have an endpoint equal to $\boldsymbol{v_i}$ with $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{\deg(\boldsymbol{v}_i)}$. Again, without loss of generality, let $\boldsymbol{\tau}_{m_{1,1}} = \boldsymbol{v_i}, \ldots, \boldsymbol{\tau}_{\deg(\boldsymbol{v}_i),1} = \boldsymbol{v_i}$. Then, we define the vertex specific B-spline $B_{(i)}$ for vertex $\boldsymbol{v}_i$ by

$$B_{(i)}(\boldsymbol{z}) = \sum_{k=1}^{\deg(\boldsymbol{v}_i)} \left[1 - \frac{d_{\boldsymbol{L}}(\boldsymbol{v}_i, \boldsymbol{z})}{\delta_k}\right) I_{[\boldsymbol{v}_i; \boldsymbol{\tau}_{k,2})}(\boldsymbol{z}). \quad (2)$$

for $\boldsymbol{z} \in \boldsymbol{L}$ and $i = 1, \ldots, W$. Hence, a linear B-spline basis on $\boldsymbol{L}$ consists of the B-splines defined by (1) and (2) and has dimension $J = |B| = \sum_{m=1}^{M} J_m + |\boldsymbol{V}|$. For simplicity of presentation, we index from now on the B-spline Basis by $1, \ldots, J$ and by construction, it holds that $\sum_{j=1}^{J} B_j(\boldsymbol{z}) = 1$ for $\boldsymbol{z} \in L$ as in the one-dimensional setting.

3.2 Intensity Estimation on a Network

On our geometric network $\boldsymbol{L}$, we specify a bin width h_m on every segment $\boldsymbol{e}_m$ and then divide $\boldsymbol{e}_m$ into $N_m = \frac{d_m}{h_m}$ bins of the same length such that $\boldsymbol{L}$ is partitioned into $N = \sum_{m=1}^{M} \frac{d_m}{h_m}$ bins in total. Let $\boldsymbol{z}_{m,k}$ denote the midpoints of these bins and assume that data on n independently observed points $\boldsymbol{x}_i, i = 1, \ldots, n$, of the point process $\boldsymbol{L}$ have been observed. We define with $y_{m,k} \in \mathbb{N}_0$ the number of observations which are contained in the k-th bin of the m-th segment and assume a Poisson distribution for the counts $y_{m,k}$ such that we have $y_{m,k} \mid \boldsymbol{z}_{m,k} \overset{\text{ind}}{\sim} \text{Poi}(\lambda_{m,k})$, where $\lambda_{m,k}$ is approximated through $\lambda_{m,k} = \varphi_{\mathcal{X}}(\boldsymbol{z}_{m,k}) \cdot h_m = \exp\left(\nu_{\mathcal{X}}(\boldsymbol{z}_{m,k}) + \log h_m\right)$. We can consider $\log h_m$ as offset and aim to estimate $\nu_{\mathcal{X}}(\boldsymbol{z})$ as continuous log-intensity for $\boldsymbol{z} \in \boldsymbol{L}$. Therefore, we replace $\nu_{\mathcal{X}}(\boldsymbol{z})$ through the B-spline basis representation $\nu_{\mathcal{X}}(\boldsymbol{z}) = \sum_{j=1}^{J} B_j(\boldsymbol{z})\gamma_j$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)^\top$ is the vector of coefficients that needs to be estimated from the data. Imposing a penalty on the resulting Poisson likelihood leads to the penalized log-likelihood (constant terms are ignored)

$$\ell_{\mathcal{P}}(\boldsymbol{\gamma}; \rho) = \sum_{m=1}^{M} \sum_{k=1}^{N_m} \left[y_{m,k} \sum_{j=1}^{J} B_j(\boldsymbol{z}_{m,k})\gamma_j - \exp\left(\sum_{j=1}^{J} B_j(\boldsymbol{z}_{m,k}) + \log h_m \right) \right] - \rho \mathcal{P}_r(\boldsymbol{\gamma}),$$

where $\mathcal{P}_r(\boldsymbol{\gamma})$ is a penalty which is defined in the next section and ρ is the smoothing parameter, which we estimate by exploiting the generalized Fellner-Schall method (Wood and Fasiolo, 2017).

3.3 Penalties on a Network

We can view the B-splines on $\boldsymbol{L}$ itself as a network graph L_B which is defined through a $J \times J$ adjacency matrix $\boldsymbol{A}$, where $\boldsymbol{A}(i,j) = 1$, if $\text{supp}(B_i) \cap \text{supp}(B_j) \neq \emptyset$ and else $\boldsymbol{A}(i,j) = 0$. Furthermore, $\boldsymbol{S_A}$ denotes the $J \times J$ shortest path matrix of L_B. Now, let $\mathcal{D}_1 = \{(i,j) \mid \boldsymbol{S_A}(i,j) = 1, 1 \leq i < j \leq J\}$. According to Eilers and Marx (1996) we penalize neighboring coefficients. A first order penalty is then defined by

$$\mathcal{P}_1(\boldsymbol{\gamma}) = \sum_{\mathcal{D}_1} (\gamma_i - \gamma_j)^2 = \boldsymbol{\gamma}^\top \boldsymbol{K}_1 \boldsymbol{\gamma}, \tag{3}$$

where $\boldsymbol{K}_1 \in \mathbb{Z}^{J \times J}$ defines the resulting quadratic form according to the pairwise differences in (3). Penalties of order $r \geq 2$ can be defined in a similar way via the shortest path matrix $\boldsymbol{S_A}$.

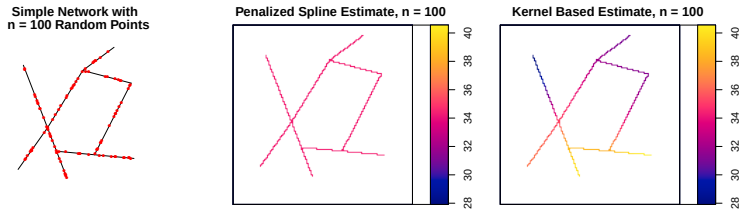


FIGURE 1. Network with $n = 100$ uniform random points (left panel), penalized spline intensity estimate (left panel) and kernel intensity estimate (right panel).

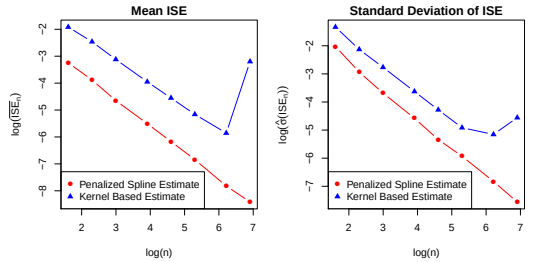


FIGURE 2. Mean and standard deviation of penalized spline based and kernel based ISE for $n = 5, 10, 20, 50, 100, 200, 500, 1000$ on a log-log-scale, where $\varphi_{\boldsymbol{X}_n}$ is defined according to a uniform intensity on $\boldsymbol{L}$.

4 Simulation Study

We independently simulate n random points according to a specified intensity function $\varphi_{\mathcal{X}_n}$ on a small network $\boldsymbol{L}$, which is shown in Figure 1, and estimate the intensity of the simulated point process with penalized spline smoothing. This is opposed to a kernel based intensity estimate with automatic bandwidth selection as implemented in the `spatstat` package.
To begin with, we specify a uniform intensity on $\boldsymbol{L}$, i.e. $\varphi_{\mathcal{X}_n}(\boldsymbol{z}) = n/|\boldsymbol{L}|$ for all $\boldsymbol{z} \in \boldsymbol{L}$. For $n = 100$ an exemplary simulated point process on $\boldsymbol{L}$ is shown in the left panel of Figure 1. The middle (right) panel shows the intensity estimate when exploiting the penalized spline (kernel based) approach. For the former method, we use $\delta = 0.05, h = 0.01$ with a first-order penalty. The penalty on the log-likelihood of the model affects $\widehat{\gamma}_1 \approx \cdots \approx \widehat{\gamma}_P$. Therefore, when using the penalized spline method we here estimate nearly a constant intensity.
We extend the above setting with multiple simulations per sample size and set $\delta = 0.05, h = 0.01, r = 1$. We also use several values of n to assess consistency. To do so we simulate $S = 1000$ networks for $n = 5, 10, 20, 50, 100, 200, 500, 1000$ and quantify the estimation error of

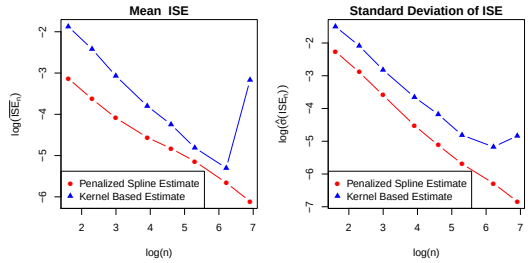


FIGURE 3. Mean (bottom left panel) and standard deviation (bottom right panel) of penalized spline based and kernel based ISE for $n = 5, 10, 20, 50, 100, 200, 500, 1000$ with true intensity (4) on a log-log-scale.

the s-th simulation through

$$\mathrm{ISE}_n(s) = \frac{1}{n^2} \int_{\boldsymbol{L}} \left(\varphi_{\mathcal{X}_n}(\boldsymbol{z}) - \widehat{\varphi}_{\mathcal{X}_n}(\boldsymbol{z}; s)\right)^2 d\boldsymbol{z},$$

where $\widehat{\varphi}_{\mathcal{X}_n}(\cdot; s)$ denotes the estimate of $\varphi_{\mathcal{X}_n}$ based on the s-th sample. That is, we quantify the estimation error through the integrated squared error (ISE) between the true density $f_{\mathcal{X}}(\boldsymbol{z}) = \varphi_{\mathcal{X}_n}(\boldsymbol{z})/n = 1/|\boldsymbol{L}|$ and the estimated density $\widehat{f}_{\mathcal{X}}(\boldsymbol{z}) = \widehat{\varphi}_{\mathcal{X}_n}(\boldsymbol{z})/n$. The resulting means $\overline{\mathrm{ISE}_n}$ over all $S = 1000$ simulations and the corresponding standard deviations $\widehat{\sigma}(\mathrm{ISE}_n)$ for penalized spline and kernel based estimation are shown in Figure 2. We can clearly see that the penalized spline approach performs distinctly superior to the kernel based approach in terms of ISE. For the penalized spline approach we see that the mean and the standard deviation of ISE_n decrease nearly linearly on a log-log-scale with the number of random points n.

In order to investigate the performance of penalized spline based intensity estimation also for non-uniform intensities, we consider now the intensity function

$$\varphi_{\mathcal{X}_n}(\boldsymbol{z}) = \sqrt{y} \exp(-xy) \cdot \frac{n}{C} \tag{4}$$

on the network from above. Here, $\boldsymbol{z} = (x, y)^\top$ is the plane-coordinate representation of a point $\boldsymbol{z} \in \boldsymbol{L}$ with $0 \leq x, y \leq 1$ and $C \approx 1.558$ is the normalization constant such that $\int_{\boldsymbol{L}} \varphi_{\mathcal{X}_n} d\boldsymbol{z} = n$. In the bottom row of Figure 3 we oppose the means (bottom left panel) and standard deviations (bottom right panel) of $S = 1000$ ISE samples for both, penalized spline and kernel based intensity estimation using the same parameter setting as in the uniform case from above. We can see that also here the penalized spline method is distinctly superior to the kernel based approach in terms of ISE. Furthermore, the decrease of the mean ISE is still linear on a log-log-scale but the relationship is not as strong as in the uniform case.

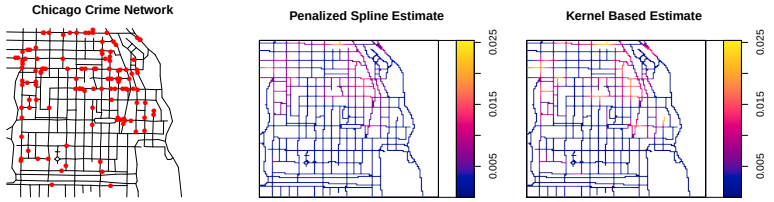


FIGURE 4. Chicago crime network (left panel), penalized spline intensity estimate (left panel) and kernel intensity estimate (right panel).

5 Application – Crimes in a District of Chicago

As application with real data we consider the Chicago crimes network which is also implemented in the `spatstat` package and elaborated in detail in Baddeley et. al (2015). The top panel of Figure 4 shows the location of 116 crimes recorded over a two-week period in 2002 in a district of Chicago (neglecting the kind of crime). The lower panels of Figure 4 show the estimates resulting from the data using a penalized spline based approach (left panel) or a kernel based estimation (right panel). We find that the high-intensity regions are similarly located for both methods. However, the peaks of the kernel intensity estimate are more explicit and the intensity fitted with the penalized spline approach is more smooth. When considering the few events in the southern area of the map extract, a peak occurs in the kernel intensity estimate which is not visible in the spline approach. Taking the performance of the kernel density estimates as shown in the simulated data above into account, it seems plausible to consider crime events to be uniformly distributed over the southern part of the network, that is, the peaky behavior of the kernel methods seems misleading.

References

Baddeley, A., Rubak, E. and Turner, R. (2015). *Spatial point patterns: Methodology and applications with R.* Chapman and Hall/CRC.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89 – 102.

Okabe, A., Satoh, T. and Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a gis-based tool. *International Journal of Geographical Information Science*, **23(1)**, 7 – 32.

Wood, S. N. and Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to tweedie location, scale and shape models. *Biometrics*, **73(4)**, 1071-1081.

Scaleable distributional regression

Thorsten Simon[1], Nikolaus Umlauf[1]

[1] Department of Statistics, Universität Innsbruck, Austria

E-mail for correspondence: Thorsten.Simon@uibk.ac.at

Abstract: Estimation of distributional regression models using datasets beyond 10^6 observations is a difficult task. We propose a novel optimizer which is based on the ideas of stochastic gradient descent and can easily deal with large data sets. Moreover, the algorithm performs automatic variable and smoothing parameter selection and its performance is in most cases superior or at least equal to other implementations for distributional regression. An implementation is provided in the R package **bamlss**. We illustrate the usefulness of the approach by implementing a state-of-the-art prediction model for lightning occurrence and counts in complex terrain.

Keywords: GAMLSS; optimization; stochastic approximation; gradient descent; count data.

1 Introduction

Fitting distributional regression models of high complexity to large data $n > 10^6$ is computationally challenging. Moreover, in many applications solving the problem also requires automatic selection of variables. Retrospective lightning analysis is a problem of this kind. For a region of $300\,000\,km^2$ in Europe we face $n \approx 4 \times 10^6$ observations and $p \approx 80$ covariates, each of which should be modelled as smooth term.
For this purpose we propose the novel *batchwise backfitting* optimizer which is based on the *stochastic gradient descent* algorithm (Sakrison, 1965) and approximates the regression coefficients stochastically. The computation of the gradients of the log-likelihood in each iteration is only based on a batch of observations. In contrast, within a Newton-Raphson type algorithm the gradients are computed on the full data in each iteration. This design principle makes the batchwise backfitting optimizer computationally simple and thus scaleable. Toulis and Airoldi (2015) provide background of estimation algorithms based on stochastic approximations.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Batchwise backfitting

This section follows the notation introduced in Umlauf et al. (2018). A standard approach to fit the posterior mode of a distributional regression model is to apply a backfitting algorithm that updates the regression coefficients iteratively,

$$\boldsymbol{\beta}_{jk}^{[t+1]} = (\mathbf{X}_{jk}^\top \mathbf{W}_{kk} \mathbf{X}_{jk} + \mathbf{G}_{jk}(\boldsymbol{\tau}_{jk}))^{-1} \mathbf{X}_{jk}^\top \mathbf{W}_{kk} (\mathbf{z}_k - \boldsymbol{\eta}_{k,-j}^{[t+1]}),$$

with working observations $\mathbf{z}_k = \boldsymbol{\eta}_k^{[t]} + \mathbf{W}_{kk}^{-1}\mathbf{u}_k$, working weights $\mathbf{W}_{kk}^{-1}$ and score vector $\mathbf{u}_k$, derived from derivates of the log-likelihood w.r.t. $\boldsymbol{\eta}_j$.
Now, instead of using all observations of the data, we only use a randomly chosen subset denoted by the subindex $[\mathbf{s}]$ in one updating step

$$\begin{aligned} \boldsymbol{\beta}_{jk}^{[t+1]} &= (1-\nu) \cdot \boldsymbol{\beta}_{jk}^{[t]} + \\ &\nu \cdot (\mathbf{X}_{[\mathbf{s}],jk}^\top \mathbf{W}_{[\mathbf{s}],kk} \mathbf{X}_{[\mathbf{s}],jk} + \mathbf{G}_{jk}(\boldsymbol{\tau}_{jk}))^{-1} \mathbf{X}_{[\mathbf{s}],jk}^\top \mathbf{W}_{[\mathbf{s}],kk} (\mathbf{z}_{[\mathbf{s}],k} - \boldsymbol{\eta}_{[\mathbf{s}],k,-j}^{[t+1]}) \\ &= (1-\nu) \cdot \boldsymbol{\beta}_{jk}^{[t]} + \nu \cdot \boldsymbol{\beta}_{jk,[\mathbf{s}]} \end{aligned} \tag{1}$$

and introduce a step length control parameter ν (or *learning rate*) specifying the amount of which $\boldsymbol{\beta}_{jk}^{[t+1]}$ is updated. This iteration mimics a classical second order stochastic gradient descent algorithm since

$$\boldsymbol{\beta}_{jk}^{[t+1]} = \boldsymbol{\beta}_{jk}^{[t]} + \nu \cdot (\boldsymbol{\beta}_{jk,[\mathbf{s}]} - \boldsymbol{\beta}_{jk}^{[t]}) = \boldsymbol{\beta}_{jk}^{[t]} + \nu \cdot \boldsymbol{\delta}_{jk}^{[t]} \tag{2}$$

where the difference $\boldsymbol{\delta}_{jk}^{[t]}$ between parameter updates from iteration t and batch $[\mathbf{s}]$ is composed from first and second order derivative information with

$$\begin{aligned} \boldsymbol{\delta}_{jk}^{[t]} &= \boldsymbol{\beta}_{jk,[\mathbf{s}]} - \boldsymbol{\beta}_{jk}^{[t]} \\ &= \left[\boldsymbol{\beta}_{jk}^{[t]} - \mathbf{H}_{[\mathbf{s}],kk} \left(\boldsymbol{\beta}_{jk}^{[t]}\right)^{-1} \mathbf{s}_{[\mathbf{s}]} \left(\boldsymbol{\beta}_{jk}^{[t]}\right)\right] - \boldsymbol{\beta}_{jk}^{[t]} \\ &= -\mathbf{H}_{[\mathbf{s}],kk} \left(\boldsymbol{\beta}_{jk}^{[t]}\right)^{-1} \mathbf{s}_{[\mathbf{s}]} \left(\boldsymbol{\beta}_{jk}^{[t]}\right) \end{aligned}$$

Hence, in each iteration the update step length is adaptive, because of the curvature information provided in $\boldsymbol{\delta}_{jk}^{[t]}$.
The working weights $\mathbf{W}_{[\mathbf{s}],kk}$, the working responses $\mathbf{z}_{[\mathbf{s}],k}$ and the predictors $\boldsymbol{\eta}_{[\mathbf{s}],k}$ are computed based on the current states $\boldsymbol{\beta}^{[t]}$. For one batch $[\mathbf{s}]$, the algorithm subsequently cycles over all parameters of the response distribution and all model terms in the typical backfitting manner, i.e., the predictors $\boldsymbol{\eta}_{[\mathbf{s}],k}$ are updated instantly. After all model terms $f_{jk}(\mathbf{X}_{[\mathbf{s}],jk}; \boldsymbol{\beta}_{jk})$ are updated the algorithm proceeds with the next batch. Hence, the batchwise backfitting algorithm updates in a memory efficient manner from batch

to batch either until all observations were included once, or the algorithm runs through the data a prespecified number of *epochs*.
The smoothness of model terms $f_{jk}(\cdot)$ is controlled by the smoothing parameters (variances) $\boldsymbol{\tau}_{jk}$. In the proposed implementation these parameters are either estimated according to an information criterion like the AIC, which is computed on an *out-of-sample* batch $[\tilde{\mathbf{s}}]$, or by using slice sampling under the information criterion. If $\nu = 1$, the algorithm can be interpreted as a resampling method and each update $\boldsymbol{\beta}_{jk}^{[t+1]}$ is a *sample* of the regression coefficients. In this case, convergence is achieved similar to MCMC algorithms, i.e., if the iterations start fluctuating around a certain level. The final estimate $\hat{\boldsymbol{\beta}}$ is then computed by taking the means or medians of the chains.
Moreover, equation (2) can also be utilized to enforce complete variable selection in a boosting type algorithm, if only the model term with the best improvement in the *out-of-sample* log-likelihood is updated.
In addition to commonly used penalties $\mathbf{G}_{jk}(\boldsymbol{\tau}_{jk})$, complete model term selection can also be incorporated by an additional lasso type penalty for coefficients $\boldsymbol{\beta}_{jk}$ (Groll et al., 2019).
Convergence of the algorithm is controlled in two ways: (1) by the step length control parameter ν, and (2) by the working weights, working responses and predictors, which are computed based on all previous batches. This means by setting, e.g., $\nu = 1/2$, the algorithm will converge after visiting m batches $[\mathbf{s}]$. If $\nu = 1$, the algorithm can be interpreted as a resampling algorithm and each update $\boldsymbol{\beta}_{jk}^{[t+1]}$ is so to say a "sample". In this case, convergence is achieved similar to MCMC algorithms, i.e., if the iterations start fluctuating around a certain level. The final estimate $\hat{\boldsymbol{\beta}}$ is then computed by taking the means or medians of the chains.
If reasonably good starting values are chosen the algorithm will converge in any case, as can be seen by changes of the expected bias at iteration t

$$\begin{aligned} E(\boldsymbol{\beta}^{[t]} - \boldsymbol{\beta}^{\star}) &= E(\nu \cdot \boldsymbol{\beta}_{[\mathbf{s}]} + (1-\nu) \cdot \boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^{\star}) \\ b^{[t]} &= E(\nu \cdot \boldsymbol{\beta}_{[\mathbf{s}]} + \boldsymbol{\beta}^{(t-1)} - \nu \cdot \boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^{\star}) \\ &= E(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^{\star}) + \nu \cdot E(\boldsymbol{\beta}_{[\mathbf{s}]} - \boldsymbol{\beta}^{(t-1)}) \\ &= b^{[t-1]} - \nu \cdot E(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}_{[\mathbf{s}]}). \end{aligned}$$

Hence, the improvements of the algorithm will be large in the beginning. The closer $\boldsymbol{\beta}^{(t-1)}$ is to the true parameters $\boldsymbol{\beta}^{\star}$ the smaller will be the improvements in the expected bias from iteration to iteration.
A simple way to find good starting values for complex models is to first estimate a model with only intercept parameters using the boosting-type flavor of the batchwise backfitting algorithm. This can be done quite quickly even with huge amounts of data.

3 Retrospective lightning analysis

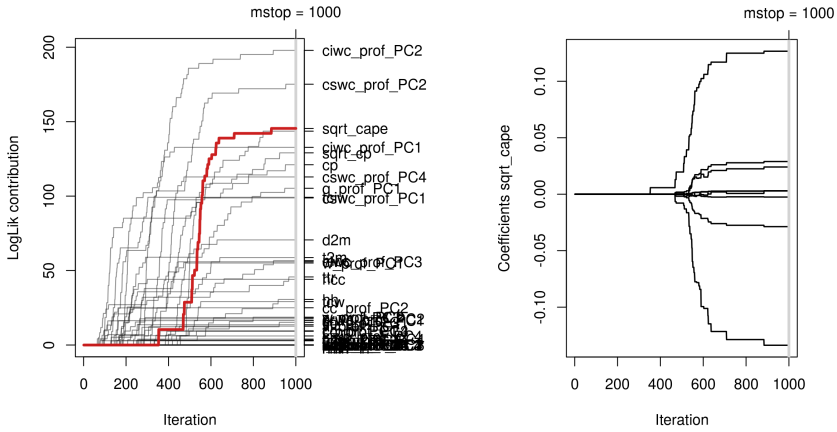


FIGURE 1. Left panel: Log-likelihood contributions of model terms, the contributions of the term `sqrt_cape` are highlighted. Right panel: Paths of the regression coefficients of the term `sqrt_cape`.

Lightning and associated atmospheric processes affect many scientific areas as well as industry and everyday life. Thus a consistent and long-term (several decades) data describing the occurrence and intensity of lightning events would be of great interest. However, observational networks that detect lightning discharges homogeneously in space and time are in the order of only a decade. For instance, lightning detection data with such properties over Austria and surrounding (ALDIS; Schulz et al., 2005) are available for the period after 2010. On the other hand atmospheric re-analyses (ERA5; Copernicus Climate Change Service, 2017) model the state of the atmosphere for several decades retrospectively, but they neither model lightning occurrence nor counts explicitly.

With a statistical model—linking lightning detection data and atmospheric re-analyses—one could (retrospectively) predict lightning for the time before 2010 and thus analyse lightning events in the past for which no observations are available. To train the model we employ ALDIS and ERA5 data, respectively, with $n \approx 4 \times 10^6$ observations and $p \approx 80$ covariates each of which entering the model as univariate smooth term. Lightning discharges are counted on a grid with mesh size of 32 km $\times$ 32 km $\times$ 1 h. The covariates are derived from ERA5 single level and pressure level output.

We employ the proposed batchwise backfitting algorithm to model lighting over Austria. One potential distribution for this type of data is a discrete generalized Pareto. 1 000 batches, each of size 4 000, are drawn from the

data. A step length control of $\nu = 0.01$ allows only slight updates of the coefficients from step to step. In each iteration only the *best fitting* term is updated. Further, the smoothing parameters in each iteration are selected using the AIC on an out-of-sample batch.
During the iterative fitting procedure the log-likelihood contributions of the different terms are tracked (Fig. 1, left). These depict patterns compareable to results known from gradient boosting that would evaluate the gradients on the whole data. In each iteration a single term is updated increasing its contribution. The coefficient paths—shown for the exemplary term of `sqrt_cape` (Fig. 1, right)—reveal a slight update of the estimates in each step. These results indicate stability of the algorithm.

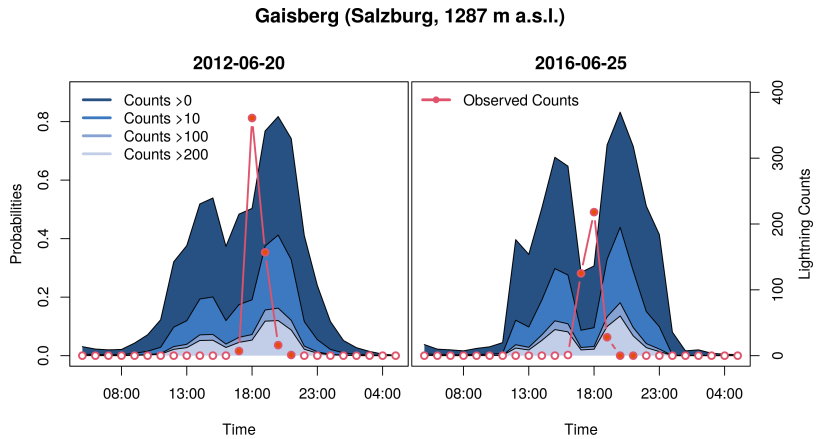


FIGURE 2. Sample application for the 32 km × 32 km grid box containing the orographic peak *Gaisberg* (47° 48′ 20″ N, 13° 6′ 45″ E) for two dates 2012-06-20 (left, in-sample) and 2016-06-25 (right, out-of-sample). Contours show the probability of exceedance. Circles show the observed flash counts. Filled circles indicate counts greater than zero.

The final model can, for instance, be used study the evolution of single events along the diurnal cycle (Fig. 2). Probabilities that lightning counts exceed thresholds of 0, 10, 100, and 200 are derived from the predicted distributions. The observed counts exhibit intense lightning events. Further, applications such as investigating the spatial propagation are possible.

Computational details: The R package **bamlss**, implementing i.a. the proposed batchwise backfitting optimizer `bbfit()`, is available at `https://CRAN.R-project.org/package=bamlss`.

Acknowledgements: We thank Gerhard Diendorfer and Wolfgang Schulz from ALDIS for data support. Thorsten Simon acknowledges the support of the Austrian Science Fund (FWF): Project number P 31836.

References

Belitz and Lang (2008). *Simultaneous selection of variables and smoothing parameters in structured additive regression models.* Comput. Stat. Data An. **53**, 61–81. doi: 10.1016/j.csda.2008.05.032.

Copernicus Climate Change Service (2017). *ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate.* Copernicus Climate Change Service Climate Date Store (CDS). Date of access: June 2019.
url: https://cds.climate.copernicus.eu/cdsapp#!/home.

Groll, Hambuckers, Kneib, and Umlauf (2019). *LASSO-type penalization in the framework of generalized additive models for location, scale and shape.* Computational Statistics & Data Analysis, **140**, 59–73, doi: 10.1016/j.csda.2019.06.005

Prieto (2014). *Modelling road accident blackspots data with the discrete generalized Pareto distribution.* Accident Anal. Prev., **71**, 38–49, doi: 10.1016/j.aap.2014.05.005.

Sakrison (1965). *Efficient recursive estimation: Application to estimating the parameters of a covariance function.* Int. J. Eng. Sci., **3(4)**, 461–483, doi: 10.1016/0020-7225(65)90029-7.

Schulz, Cummins, Diendorfer and Dorninger (2005). *Cloud-to-ground lightning in Austria: A 10-year study using data from a lightning location system.* J. Geophys. Res., **110(D9)**. doi: 10.1029/2004JD005332.

Simon, Mayr, Umlauf and Zeileis (2019). *NWP-based lightning prediction using flexible count data regression.* Adv. Stat. Clim. Meteorol. Oceanogr., **5**, 1–16. doi: 10.5194/ascmo-5-1-2019

Rigby and Stasinopoulos (2005). *Generalized additive models for location, scale and shape (GAMLSS).* J. Roy. Stat. Soc. C, **54(3)**, 507–554. doi: 10.1111/j.1467-9876.2005.00510.x

Toulis and Airoldi (2015). *Scalable estimation strategies based on stochastic approximations: Classical results and new insights.* Stat. Comput, **25(4)**, 781–795. doi: 10.1007/s11222-015-9560-y.

Tutz and Binder (2006). *Generalized additive modeling with implicit variable selection by likelihood-based boosting* Biometrics, **62(4)**, 961–971. doi: 10.1111/j.1541-0420.2006.00578.x.

Umlauf, Klein and Zeileis (2018). *BAMLSS: Bayesian additive models for location, scale, and shape (and beyond).* J. Comput. Graph. Stat, **3**, 612–627. doi: 10.1080/10618600.2017.1407325.

Elastic analysis of irregularly and sparsely sampled curves

Lisa Steyer[1], Almond Stöcker[1], Sonja Greven[1]

[1] Humboldt-Universität zu Berlin

E-mail for correspondence: lisa.steyer@hu-berlin.de

Abstract: We provide methods and algorithms to approximate the elastic distance between irregularly and sparsely sampled curves and to fit smooth elastic means for collections of such curves. Moreover, we illustrate both methods by applying them to a dataset comprising GPS tracks, where we first cluster the tracks based on the elastic distance between them and then estimate elastic means for each cluster.

Keywords: square-root-velocity; elastic distance; re-parametrisation; shape.

1 Analysis of functional data modulo re-parametrisation

We are interested in studying statistical properties of collections of observed curves, for example the outline of objects or movement patterns of subjects. Although these curves are modelled as functions $\boldsymbol{\beta} : [0,1] \to \mathbb{R}^d$, only their image represents the curve. This means the analysis should be independent of the parametrisation. To deal with this invariance, Srivastava et al. (2011) proposed a distance on the quotient space of absolutely continuous curves modulo parametrisation.
For two absolutely continuous curves $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, this elastic metric is defined as the minimal L_2-distance between the corresponding square-root-velocity (SRV) functions after alignment: For monotonically increasing, onto and differentiable warping functions $\gamma : [0,1] \to [0,1]$ define

$$d(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \inf_{\gamma} \|\mathbf{q}_1 - (\mathbf{q}_2 \circ \gamma) \cdot \sqrt{\dot{\gamma}}\|_{L_2}, \tag{1}$$

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

with SRV transformations $\mathbf{q}_1$ and $\mathbf{q}_2$ of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ defined via

$$\mathbf{q}_i(t) = \begin{cases} \frac{\dot{\boldsymbol{\beta}}_i(t)}{\sqrt{\|\dot{\boldsymbol{\beta}}_i(t)\|}} & \text{if } \boldsymbol{\beta}_i(t) \neq 0 \\ 0 & \text{if } \boldsymbol{\beta}_i(t) = 0 \end{cases} \quad \text{for } i = 1, 2.$$

Thus, $(\mathbf{q}_2 \circ \gamma) \cdot \sqrt{\dot{\gamma}}$ is the SRV transformation of $\boldsymbol{\beta_2} \circ \gamma$. A solution to the variational problem (1) is usually approximated using a dynamic programming algorithm (for instance in Srivastava et al. (2011)). This works well in the case of densely observed curves.

Nevertheless, in real-world applications, we usually observe curves only at a finite (and often small) number of discrete points, where even the number of points might differ between curves. We present an algorithm to approximate the elastic distance (1) when at least one of the curves is discretely observed via interpreting them as polygons with constant speed parametrisation between its corners. This enables us to apply distance-based methods like clustering and classification. Moreover, we use spline functions to compute a smooth representative of the Frchet mean with respect to (1) for a collection of observed curves.

We demonstrate both methods on a dataset comprising GPS waypoints tracked on Tempelhof Field, a recreation area in Berlin (see Figure 1). The dataset consists of 55 paths with 15 to 45 waypoints each. Clustering and smooth mean estimation allows us to find new paths on Tempelhof field not yet included in OpenStreetMap.

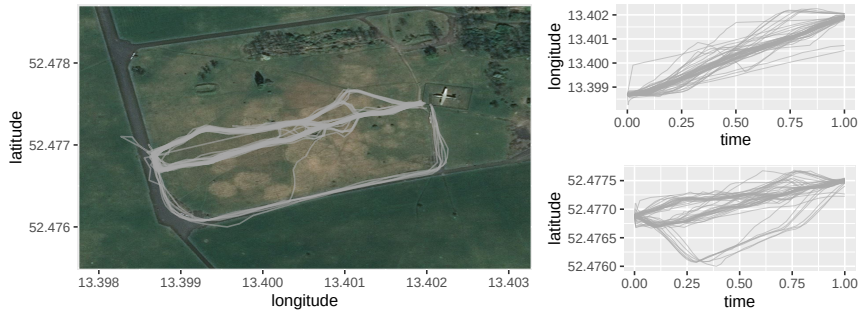


FIGURE 1. Left: GPS paths tracked on Tempelhof Field and plotted on OpenStreetMap. Right: longitude and latitude over relative time.

We are solely interested in analysing the paths the subjects walked on, not the trajectories over time. Separately looking at longitude and latitude over time suggests that the individuals had quite different walking patterns, namely did not move with constant speed. This implies a classical functional analysis of the trajectories is not suitable to study the paths used by the test subjects.

2 Methods

Both methods presented rely on treating unobserved curves $\boldsymbol{\beta}$ with SRV transformation $\mathbf{q}$ as polygons parametrised with constant speed between the observed corners $\boldsymbol{\beta}(s_1), \ldots, \boldsymbol{\beta}(s_m)$. In this case, the problem of finding an optimal re-parametrisation $\boldsymbol{\beta} \circ \gamma$ of $\boldsymbol{\beta}$ to another curve with SRV transformation $\mathbf{p}$ can be simplified (similar as in Lahiri et al. (2015)). More precisely, we can show that instead of solving the minimisation problem (1) over the function space of all suitable warping functions γ, we only need to solve a maximisation problem over a subset of $\mathbb{R}^{m-1}$, the new parametrisations $t_1 = \gamma^{-1}(s_1), \ldots, t_m = \gamma^{-1}(s_m)$ at the corners of the observed polygon.
Since $\boldsymbol{\beta}$ is assumed to be a polygon with constant speed parametrisation between its corners, its SRV transformation $\mathbf{q}$ is piecewise constant with $\mathbf{q}|_{[s_j, s_{j+1}]} = \mathbf{q}_j \in \mathbb{R}^d$ for all $j = 1, \ldots, m$. We consider

$$\begin{aligned} \textbf{Maximise} \quad & \sum_{j=0}^{m-1} \sqrt{(s_{j+1} - s_j) \int_{t_j}^{t_{j+1}} \langle \mathbf{p}(t), \mathbf{q}_j \rangle_+^2 \, dt} \\ \textbf{w.r.t} \quad & 0 = t_0 \leq t_1 \leq \cdots \leq t_m = 1. \end{aligned} \tag{2}$$

Thereby, $\mathbf{p}$ can be the SRV function of any absolutely continuous curve with $\|\mathbf{p}\|_\infty < \infty$ and $\langle \cdot, \cdot \rangle_+$ denotes the positive part of the d-dimensional scalar product.
The way we handle the remaining optimisation problem depends on whether the second curve with SRV transformation $\mathbf{p}$ is a model-based smooth curve or the SRV transform of an observed polygon as well. In the following, we describe algorithms for both cases and show how a smooth mean for a sample of curves can be computed.

2.1 Elastic distance for two observed curves

For two SRV curves $\mathbf{p}$ and $\mathbf{q}$ that are piecewise constant, for instance the SRV transformations of observed polygons, we can even derive a closed form solution to the maximisation problem (2) with respect to the new parametrisation $t_j \in \mathbb{R}$ at each of the corners $\boldsymbol{\beta}(s_j)$. With this, we propose a coordinate wise maximisation procedure, where we iterate between two steps:

(i) Update all t_j with $j \in \{1, \ldots, m-1\}$ **odd**.

(ii) Update all t_j with $j \in \{1, \ldots, m-1\}$ **even**.

Parallel computation is possible, since the update for the odd (or even) indices does not depend on the other odd (even) indices. We can show that every accumulation point of the resulting sequence $\{(t_j^{(k)})_{j=1,\ldots,m-1}, k \in \mathbb{N}\}$ is a local maximiser. Moreover, it can easily be adapted for closed shapes by updating $t_0 \in [t_{m-1} - 1, t_1]$ and setting $t_m = t_0 + 1$.

2.2 Smooth means for samples of curves

Similar to approaches in functional data analysis we like to model characteristics of a distribution of curves using differentiable basis representations. We suggest considering piecewise linear SRV curves. This implies differentiable curves. Moreover we can show that such curves with piecewise linear square-root-velocity transformation have unique representatives modulo re-parametrisation.
With this in mind, we can use the space of linear spline SRV curves with fixed knots as a model space for the mean of a sample of curves with respect to the elastic metric. Precisely, we compute a smooth approximation of the **Frchet mean**, a generalisation of the sample mean for metric spaces. Here, the Frchet mean is a minimiser of the sum of quadratic elastic distances. Hence, for a set of observed SRV curves $\{\mathbf{q}_i \mid i = 1, \ldots, n\}$, we consider the following minimisation problem

$$\begin{aligned} \textbf{Minimise} \quad & \sum_{i=1}^{n} \inf_{\gamma_i} \left\| \mathbf{p} - (\mathbf{q}_i \circ \gamma_i)\sqrt{\dot{\gamma}_i} \right\|_{L_2}^2 \\ \textbf{w.r.t} \quad & \mathbf{p} : [0,1] \to \mathbf{R}^d \text{ piecewise linear and continuous.} \end{aligned} \tag{3}$$

To tackle this nested optimisation problem we propose to alternate between the outer minimisation over the current linear SRV mean function $\mathbf{p}$ and the inner minimisation over the warping functions γ_i. More precisely, we alternate the following steps:

(i) For a given spline curve $\mathbf{p}$ update the optimal re-parametrisations γ_i of the observed curves $\mathbf{q}_i$, $i = 1, \ldots n$. To do so we assume $\boldsymbol{\beta}_i$ is a parametrised polygon with observed values at its corners and constant speed between them. Hence we assume a piecewise constant SRV transformation $\mathbf{q}_i$ of $\boldsymbol{\beta}_i$ for all $i = 1, \ldots, n$. This means we are left with minimisation problem (2) which we tackle using a gradient descent method.

(ii) Update the least-squares estimates for the coefficients of the spline curve $\mathbf{p}$ for a given set of parametrisations γ_i via minimising the sum of squared L_2-distances

$$\sum_{i=1}^{n} \left\| \mathbf{p} - (\mathbf{q}_i \circ \gamma_i)\sqrt{\dot{\gamma}_i} \right\|_{L_2}^2 = \sum_{i=1}^{n} \int_0^1 \left\| \mathbf{p}(t) - (\mathbf{q}_i \circ \gamma_i(t))\sqrt{\dot{\gamma}_i(t)} \right\|^2 \, dt$$

In practice we need to replace the integrals by discrete approximations.

This algorithm can be adapted for closed shapes as well. Here we add a cost function penalising openness with increasing weight to the loss function in step (ii). See Figure 2 for an example on a toy dataset of heart shapes.

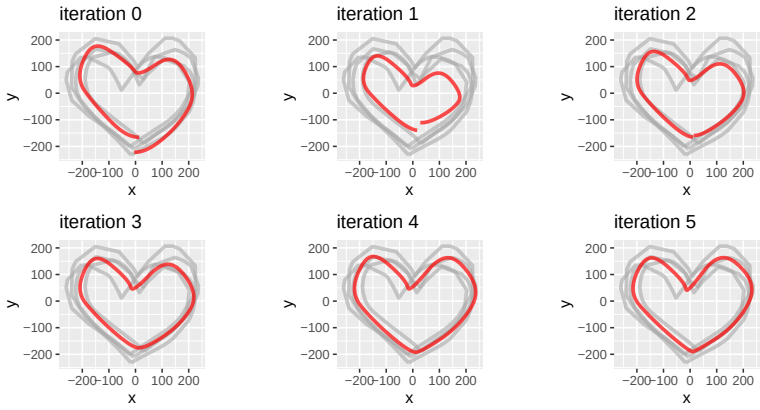


FIGURE 2. Five iterations of the algorithm for closed curves.

3 Data example: Clustering and modelling smooth means of GPS-tracks

From the GPS data described in section 1, we recover the paths the individuals walked on while tracking their trajectories. This is done in two steps. First, the tracks are clustered using average linkage based on the elastic distance (as described in 2.1) and the elbow criteria for stopping. Afterwards we compute a smooth elastic Frchet mean for each of the four largest clusters (as described in 2.2).

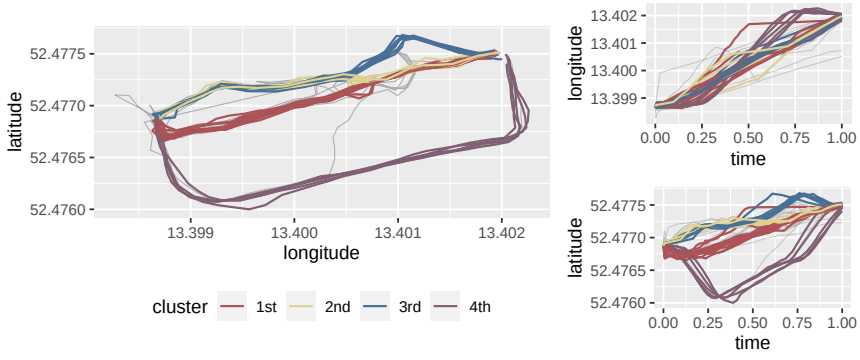


FIGURE 3. Left: The observed trajectories as elements of the four largest clusters. Right: Longitude and latitude for the trajectories of the four largest clusters.

The clustering result displayed on the top left part of Figure 3 is visually satisfying. Looking again at longitude and latitude separately clearly indicates that clustering based on the usual Euclidean distance would lead to

worse results. In particular, elements of the first and third largest clusters might be classified differently using a non-elastic distance.

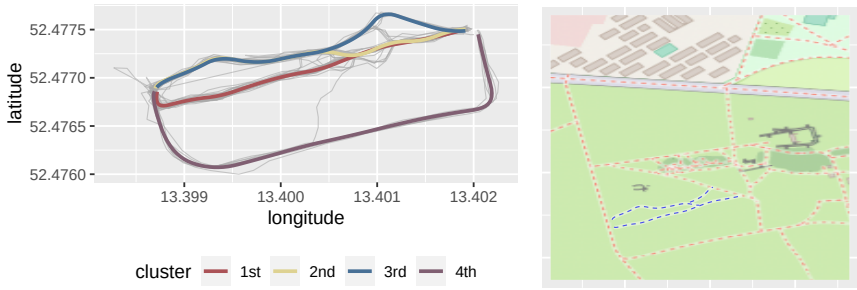


FIGURE 4. Left: Smooth means modelled as linear SRV curves with 9 inner knots for the four largest clusters. Right: The new paths (in blue) of the four largest clusters added to the existing OpenStreetMap.

After the classification step, we compute a smooth mean curve for each of the four largest clusters. The mean curves displayed in Figure 4 have been obtained using linear SRV spline curves with nine inner knots. They seem to describe the observed tracks well, although the number of estimated spline coefficients and therefore model parameters is low (22 coefficients per curve compared to at least 30 per observation). Thus, we obtain a smooth mean curve for irregularly sampled curves based on the elastic distance that captures the data well and allows dimensionality reduction.
One application of the procedure outlined above could be to identify new paths not yet included in an existing map. The smooth mean curves could be added to an OpenStreetMap, for instance, where we only add parts of our estimated means that are notable different from already existing paths. An example of the resulting map is displayed in Figure 4.

References

Lahiri, S., Robinson, D., Klasen, E. (2015). Precise matching of PL curves in $\mathbb{R}^N$ in the square root velocity framework. *Geometry, Imaging and Computing*, **3**, 133 – 186.

Srivastava, A., Klassen, E., Joshi, S.H., Jermyn, I.H. (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**, 1415 – 1428.

Dynamic Bayesian clustering of sport activities

Stival Mattia[1], Bernardi Mauro[12]

[1] Department of Statistical Sciences, University of Padova, Italy
[2] Institute for applied mathematics "Mauro Picone" (IAC) CNR, Rome, Italy

E-mail for correspondence: mattia.stival@phd.unipd.it

Abstract: The monitoring of sport activities through the use of smart–devices is assuming an increasing importance in several disciplines. In this context, data are collected as a sequence of activities, where each activity is represented by a partially–observed multivariate time series characterized by complex dependence structures. We propose a Bayesian matrix–variate dynamic mixture model for clustering trajectories of a large panel N of P–variate time series. The matrix state space formulation allows to consider for both longitudinal and cross sectional dependence, accounting also for missing values and other anomalies that characterize this kind of data. A fully conjugate approach is adopted, and the relative Gibbs sampler to sample from the full posterior distribution is available. Computational achievements can be obtained by performing Kalman recursions on a reduced form of the vectorized model, and simulating cluster allocations in one step, by using a MH within Gibbs algorithm. In the empirical application we analyze the running activities of one athlete.

Keywords: Bayesian dynamic clustering; matrix–variate; performance analysis.

1 Introduction

The monitoring of sport activities as well as the analysis of sport events is a topic of increasing interest in several disciplines such as biology, medicine, statistics and quantitative methods, engineering, material science and mathematics. The reason for such interest relies on the primary need to improve the knowledge and individualise the design of training activities and exercise programs to maximise the improvements, and avoid over–training, which may lead to impaired health, and typically under–performance (see, e.g., Cardinale and Varley, 2017). Nowadays, the use of

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

GPS–enabled tracking devices and heart rate monitors is common in several sporting disciplines, such as running, swimming, and cycling. In this context, data are collected as a sequence of N activities, where each activity is represented by a high frequency multivariate time series collecting P different variables, such as, GPS position, altitude, speed, and heart rate. Over the past years, the scientific interest of researchers has been catalysed toward the use of training data to monitor sport activities, and different solutions to many relevant problems have been proposed. Cardinale and Varley (2017), for example, pose their attention on recent technological advancements on the use of wearable technologies to quantify and monitor training load, which is relevant in sport sciences since it allows to optimise the training programmes, avoiding the risks related to overtraining and overreaching. They distinguish between data regarding internal and external load. In the first case they refer to data related to more physiological aspects, such as the heart rate responses to stimulation imposed by training activities. In the second case they refer to data related to the work completed by the athletes, measured independently of their internal characteristics, such as speed and duration. They then focus on validity and reliability of the usage of such data, highlighting the importance of analysing them individually, for each athlete. Many approaches to predict the performance of athletes are based on the original work published by Calvert et al. (1976). For example, Kolossa et al. (2017) propose the use of the so-called fitness-fatigue model for performance estimation, leveraging the Kalman filter algorithm. The model requires as input variable the training load and provides as output the performance, being dependent on the initial performance, the training load (which has to be estimated, somehow), and two unobserved variables, called fitness and fatigue. Although this approach considers the training process as a sequence of activities, it does not exploit the potential of the ever growing amount of data collected by athletes. A valuable contribution in this field was provided by Frick and Kosmidis (2017), who developed an R package aiming to fill the gap between the routine collection of data from sport devices and their analyses using the R statistical software. The package provides several user-friendly utilities for importing, managing, and analysing tracking data. Although the relevance of their contribution, the methods they propose do not account for the real-time usage of these data. Specifically, we think that providing feedback information on the effects of training results to be more effective if the feedback comes while the activity is performed, in order to make well-time decisions on it. For this reason we propose a Bayesian matrix–variate clustering model useful for classifying online the trajectories of multivariate time series, accounting also for missing values and other anomalies that characterize this kind of data. In this field, clustering trajectories allows to identify groups of activities which require similar effort, which is useful for understanding how one athlete is behaving during the activity.

2 The model

Let $\mathbf{Y}_t$ be the $P \times N$ matrix of observations storing, in the n–th column, the P–dimensional vector of observations for the n–th activity at time t, for $n = 1, \ldots, N$, and $t = 1, \ldots, T$. We assume that the N activities can be clustered into G different groups, and that the activities belonging to the same group share the trends for all the measured variables. More specifically, we assume that $\mathbf{Y}_t$ can be described by the following state space model

$$\mathbf{Y}_t = \mathbf{R}_t + \mathbf{Z}\mathbf{A}_t\mathbf{S}^{\mathrm{T}} + \boldsymbol{\Upsilon}_t, \quad \boldsymbol{\Upsilon}_t \sim \mathbf{MN}_{P,N}(\mathbf{0}, \boldsymbol{\Sigma}^R \otimes \boldsymbol{\Sigma}^C) \tag{1}$$

$$\mathbf{A}_{t+1} = \mathbf{T}\mathbf{A}_t + \boldsymbol{\Xi}_t, \quad \boldsymbol{\Xi}_t \sim \mathbf{MN}_{Q,G}(\mathbf{0}, \boldsymbol{\Psi}^R \otimes \boldsymbol{\Psi}^C) \tag{2}$$

for some starting value $\mathbf{A}_1$. In the above equations, the matrices $\mathbf{Z}$ and $\mathbf{T}$ are non-stochastic, $\mathbf{R}_t$ is a known matrix, $\mathbf{A}_t$ is a $Q \times G$ dimensional matrix that stores by columns the latent states of each group, $\mathbf{S}$ is a $N \times G$ selection matrix, independent of the time, such that its n–th row is $\mathbf{S}_n = \begin{bmatrix} I(S_n = 1) & I(S_n = 2) & \cdots & I(S_n = G) \end{bmatrix}$, where $I(\cdot)$ denotes the indicator function, and $\boldsymbol{\Upsilon}_t$ and $\boldsymbol{\Xi}_t$ follow a matrix–variate Normal distribution with row–column decomposable covariance matrices (Gupta and Nagar, 1999). It is worth noticing that in the matrix–variate state space formulation of the model in equations (1)–(2), the time series dependence of the observed variables $\mathbf{Y}_t$ is accounted for by the matrix autoregressive process described by the state equation (2), the matrices $\boldsymbol{\Sigma}^R$ and $\boldsymbol{\Sigma}^C$ capture the cross–sectional dependence between variables and activities, respectively, while $\boldsymbol{\Psi}^R$, $\boldsymbol{\Psi}^C$ capture the cross–sectional dependence within states and between groups. Therefore, the state space formulation accounts for all the complex sources of dependence structures among the observed variables. Moreover, conditional on the clustering variable $\mathbf{S}$ and the model parameters, standard routines for state space models (Durbin and Koopman, 2012) can be applied to the vectorised representation of the model $\mathbf{y}_t = \mathtt{vec}(\mathbf{Y}_t)$. Based on the idea of Jungbacker and Koopman (2008), effective performance gains can be achieved without loss of information by collapsing the observations using the transformation $\mathbf{y}_t^L = \mathbf{A}^L\mathbf{y}_t$, with $\mathbf{A}^L = (\mathbf{S}^{\intercal}(\boldsymbol{\Sigma}^C)^{-1}) \otimes (\boldsymbol{\Sigma}^R)^{-1}$, where we have assumed that $\mathbf{Z}$ is full row–rank. We notice that $\boldsymbol{\Sigma}^R \otimes \boldsymbol{\Sigma}^C = (c\boldsymbol{\Sigma}^R) \otimes (\boldsymbol{\Sigma}^C/c)$, as well as $\boldsymbol{\Psi}^R \otimes \boldsymbol{\Psi}^C = (c\boldsymbol{\Psi}^R) \otimes (\boldsymbol{\Psi}^C/c)$, which implies that different combinations of the parameters lead to the same likelihood. It is sufficient to require $\sigma^R_{11} = 1$ and $\psi^R_{11} = 1$ to achieve identifiability. Many different conjugate solutions have been proposed in literature for dealing with prior specification with these identifiability constraints. A fully conjugate approach can be adopted by considering

$$\boldsymbol{\Sigma}^R = \begin{bmatrix} 1 & \gamma_\sigma{}^{\mathrm{T}} \\ \gamma_\sigma & \Phi_\sigma + \gamma_\sigma\gamma_\sigma^{\mathrm{T}} \end{bmatrix} \quad \boldsymbol{\Psi}^R = \begin{bmatrix} 1 & \gamma_\psi{}^{\mathrm{T}} \\ \gamma_\psi & \Phi_\psi + \gamma_\psi\gamma_\psi^{\mathrm{T}}, \end{bmatrix} \tag{3}$$

where γ_σ and γ_ψ are multivariate normal, and Φ_σ and Φ_ψ are inverse Wishart random variables. In the Gibbs sampler, the elements are updated

separately, conditional of each others. If N is large, the step in which cluster allocation are updated results to be infeasible, since $N \times G$ Kalman filter recursions have to be run. The selection matrix $\mathbf{S}$ can be updated entirely in one step, by following the methods proposed by Titsias and Yau (2017) and Zanella (2020). As an alternative, we propose to consider a Metropolis-Hasting step. Let $\mathbf{S}^{(it-1)}$ be the selection matrix at iteration $it-1$. At iteration it, the proposed matrix $\mathbf{S}^{\star}$ with n–th row $\mathbf{S}^{\star}_n = \begin{bmatrix} I(S^{\star}_n = 1) & \cdots & I(S^{\star}_n = G) \end{bmatrix}$ is built by drawing N independent rows such that

$$Pr(S^{\star}_n = k \mid S^{(it-1)}_n) = \frac{e^{\beta I(S^{(it-1)}_n = k)}}{\sum_{g=1}^{G} e^{\beta I(S^{(it-1)}_n = g)}}, \tag{4}$$

for $k = 1, \ldots, G$, $n = 1, \ldots, N$, and β tuning parameter. It is worth noticing that the existing approaches can be combined together, in order to get a greater flexibility in the prior specification, as well as alternative identifiability constraint can be taken into account to better manage the computational costs. We refer to Wang and West (2009) and McCulloch et al. (2000) for further details and alternative specifications. Furthermore, the model can be easily extended for considering the effect on the measurement of time-dependent and activity-specific covariates.

3 Application

In the application we cluster a sequence of $N = 90$ running activities 600 seconds long into $G = 3$ groups based on the variables Heart Rate (in beats per minute), Speed (in meter per seconds), and Cadence (in steps per minute) collected by one athlete in the period between 2017-07-20 and 2018-09-24. For each activity, the variable Altitude (in meter) for each second is measured. We consider its first difference as time dependent and activity specific covariate, and assume that variation in Altitude has a group-specific and additive effect on the responses, by considering in equation (1) the term $\mathbf{R}_t = \mathbf{B}\mathbf{S}^{\intercal}\mathbf{X}_t$. In this notation, $\mathbf{X}_t = \texttt{diag}(X_{1t}, \ldots, X_{Nt})$ is a matrix, storing in the diagonal the variation in altitude at time t of the N activities, and $\mathbf{B}$ is a $P \times G$ matrix, storing in the g–th column, a P-dimensional vector of additive effects on Heart Rate, Speed, and Cadence of variation in Altitude for activities belonging to the group g. In this way, the cluster allocation of one activity is a personalized and relevant summary which considers both internal and external load information (see, e.g., Cardinale and Varley, 2017). In Figure 1 and Figure 2 we present the results. Specifically, in Figure 1 the posterior mean of the signals (i.e., $\mathbf{\Theta}_t = \mathbf{Z}\mathbf{A}_t$) are represented with a thicker line, while in the background the activities are colored according to their cluster obtained by MAP. As we can see, the signals of the blue group appear to be different for all the variables, differently from the other groups that are characterized by a similar

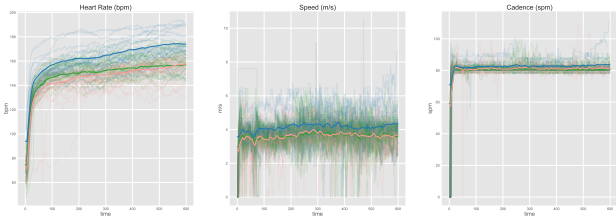


FIGURE 1. Posterior median of signals $\boldsymbol{\Theta}_t = \mathbf{Z}\mathbf{A}_t$ for the variables considered in the analysis. In the background the activities are colored according to their cluster, obtained by MAP.

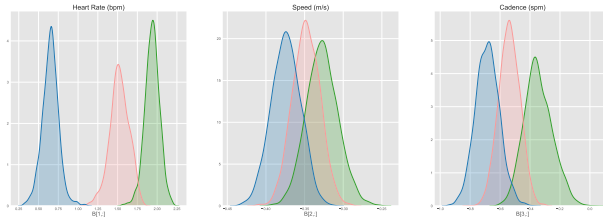


FIGURE 2. Posterior densities of coefficients in $\mathbf{B}$. These plots show the effect on the variable Heart Rate (bpm), Speed (m/s), and Cadence (spm) of 1 positive meter variation in Altitude (m/s) for the different groups, respectively.

behavior for the variables Speed and Cadence. These groups essentially differ for the behavior of the variable Heart Rate. By assuming that the effort of one activity can be described by behavior of Heart Rate over time, pink cluster require less effort than the green one, as well as the green cluster require less effort than the blue one. In this sense, we can affirm that activities in pink cluster are better then activities in green cluster, since they are characterized by similar behavior with respect to the variable Speed and Cadence, but they require less effort over time. The groups identified by the model differ also with respect to $\mathbf{B}$. Figure 2 shows the posterior density of the coefficients stored in $\mathbf{B}$. The plots show the effect on Heart Rate, Speed, and Cadence of 1 positive meter variation in Altitude (m/s) for the three groups, respectively. The effects are positive for the variable Heart Rate, and negative for the other variables. A positive variation in Altitude increases the Heart Rate, i.e., the activity require more effort, although both Cadence and Speed decrease instantaneously. Moreover, by looking at the absolute values of the effects in Figure 2, we notice that the Heart Rate behiavior of blue group activities is less sensible to variation in Altitude with respect to the other groups, differently from the variable

Speed and Cadence for which the absolute values are generally larger. We highlight the role of example of this application. Further developments will consider also lagged effects of variation in Altitude, as well as the presence of other covariates (such as, for example, Temperature) that might impact on $\mathbf{Y}_t$.

References

Calvert, T.W., Banister, E.W., Savage, M.V., and Bach, T. (1976). A systems model of the effects of training on physical performance. *IEEE Transactions on Systems, Man, and Cybernetics*, **6**(2),94 – 102

Cardinale, M. and Varley, M.C. (2017). Wearable training-monitoring technology: applications, challenges, and opportunities. *International Journal of Sports Physiology and Performance*, **12**(s2), 55 – 62.

Durbin, J. and Koopman, S.J. (2012) *Time Series Analysis by State Space Methods*. Oxford university press.

Frick, H. and Kosmidis, I. (2017) trackeR: Infrastructure for Running and Cycling Data from GPS-Enabled Tracking Devices in R. *Journal of Statistical Software*, **82**(7), 1 – 29.

Gupta, A.K. and Nagar K.N. (1999) *Matrix Variate Distributions*. Chapman and Hall/CRC.

Jungbacker, B. and Koopman, S.J. (2008) Likelihood-based analysis for dynamic factor models. Discussion paper Vrije Universiteit, Amsterdam.

Kolossa, D., Azhar, M.B., Rasche, C., Endler, S., Hanakam, F., Ferrauti, A. and Pfeiffer, M. (2017) Performance estimation using the fitness-fatigue model with kalman filter feedback. *International Journal of Computer Science in Sport*, **16**(2),117 – 129.

McCulloch, R.E., Polson, N.G. and Rossi, P.E. (2000) A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of econometrics*, **99**(1), 173 – 193.

Titsias, M.K. and Yau, C. (2017) The Hamming ball sampler. *Journal of the American Statistical Association*, **112**(520), 1598 – 1611.

Wang, H. and West, M. (2009) Bayesian analysis of matrix normal graphical models. *Biometrika*, **96**(4), 821 – 834.

Zanella, G. (2020) Informed Proposals for Local MCMC in Discrete Spaces. *Journal of the American Statistical Association*, **115**(530), 852 – 865.

A Coupled Hidden Markov Model for Daily Rainfall at Multiple Sites

Oliver Stoner[1] and Theo Economou[1]

[1] University of Exeter, UK

E-mail for correspondence: O.R.Stoner@exeter.ac.uk

Abstract: We present a simple but flexible modelling framework for simulation of rainfall time series at multiple sites. Coupled hidden Markov latent states capture spatio-temporal dependence in rainfall occurrence, while multivariate random effects capture correlation in intensity. The model is set in the Bayesian hierarchical framework for thorough quantification of parametric and predictive uncertainty, and for ease of implementation. We apply the model to three sites of varying distance, to illustrate flexibility in capturing different levels of correlation.

Keywords: Bayesian; Hierarchical; Latent; Multivariate; Spatio-temporal.

1 Introduction

Statistical rainfall models play a key role in environmental risk assessment. Probabilistic modelling of rainfall is, however, challenging due to high-levels of natural variability and a complex spatio-temporal dependency structure. Hidden Markov models (HMMs) simplify this endeavour by characterising the distribution of rainfall occurrence and intensity r_t at time step $t = 1, \ldots, n$ as a mixture of a finite, ideally small, number of latent states, so that $p(r_t) = \sum_{j=1}^{Z} \mathbf{1}(z_t = j) p(r_t | z_t)$.

In the simplest case, $Z = 2$ states can be taken to represent "dry" periods of zero or very little rainfall ($z_t = 1$) and "wet" periods of predominantly non-zero rainfall ($z_t = 2$). The persistence of these states is defined by a first-order Markov model, including an initial state probability vector $\mathbf{P_0}$ and a transition matrix $\mathbf{P}$, such that $p(z_t = i \mid z_{t-1} = j) = P_{i,j}$.

To characterise temporal structure in the occurrence and intensity of rainfall at multiple sites, we might desire one latent quantity $\mathbf{z}^{(s)}$ for each site $s = 1, \ldots, S$. However, assuming these are independent across sites means

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

that model simulations are unlikely to exhibit realistic between-site correlation in occurrence and intensity. Coupled HMMs introduce dependence between the latent chains, by assuming the probabilistic model for $z_t^{(s)}$ depends on at least $z_{t-1}^{(-s)}$ and potentially on $z_t^{(-s)}$ as well. This can be achieved in a number of ways, for example by expanding the transition matrix $\mathbf{P}$ to higher dimensions, but here we focus on one very flexible approach which specifies a full joint distribution for $\mathbf{z_t} = (z_t^{(1)}, \ldots, z_t^{(S)})$.
Suppose we have $S = 3$ sites and one HMM quantity $z_t^{(s)} = 1, 2$ for each. At time t there are $Z^S = 8$ possible combinations of the three latent quantities, so we can define a new quantity $z'_t = 1, \ldots, Z^S$ to represent each of these combinations, as illustrated in Table 1. This is then modelled as a hidden Markov quantity, where the associated $\mathbf{P}_0$ and $\mathbf{P}$ are to be estimated.

TABLE 1. A coupled hidden Markov model with two states and three sites.

	z'_t	1	2	3	4	5	6	7	8
Site 1	$z_t^{(1)}$	1	2	1	2	1	2	1	2
Site 2	$z_t^{(2)}$	1	1	2	2	1	1	2	2
Site 3	$z_t^{(3)}$	1	1	1	1	2	2	2	2

2 Model for Daily Rainfall

We apply this approach to 2019 daily rainfall data at three sites, illustrated in Figure 1. To investigate the flexibility of the model in capturing different levels of correlation between sites, two of the sites, Teignmouth and Camborne, are both located very close together, while the third site, East Kilbride, is located much further away, deep in the frozen wastelands of the North. As exhibited by the scatter plots, correlation in rainfall occurrence and intensity are considerably stronger between the two closer sites.

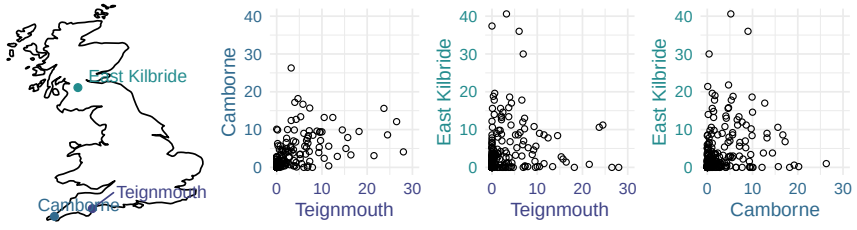


FIGURE 1. **Left:** Map showing the locations of the three rainfall sites. **Right:** Scatter plots showing 2019 daily rainfall observations at the three sites (mm).

HMMs for rainfall can be roughly separated into three components: 1)

a latent state model (which we have already identified); 2) a probabilistic model for rainfall occurrence, conditional on the latent state; and 3) a probabilistic model for rainfall intensity, conditional on both rainfall occurrence and the latent state. The probability of rainfall occurrence can generally be treated as a parameter to be estimated in some zero-inflated or hurdle model, which can vary with time and location (as in Stoner and Economou, 2020). For simplicity, we specify here that $P(r_t^{(s)} > 0 \mid z_t^{(s)} = 1) < 0.001$ and $P(r_t^{(s)} > 0 \mid z_t^{(s)} = 2) > 0.999$, such that the latent state drives the occurrence of rainfall almost entirely. For 3), we adopt a Gamma$(\mu_t^{(s)}, \sigma^{(s)})$ model, parametrised in terms of a mean parameter μ and standard-deviation parameter σ. To capture structured between-site heterogeneity and seasonal variability, the model for $\mu_t^{(s)}$ combines a site-specific intercept with a site-specific spline of time of year:

$$\log(\mu_t^{(s)}) \quad = \quad \alpha^{(s)} + \sum_{k=1}^{K} \beta_k^{(s)} x_t^{(k)} + \phi_t^{(s)}. \tag{1}$$

To capture between-site correlation in rainfall intensity, we also include the random effect $\phi_t = (\phi_t^{(1)}, \phi_t^{(2)}, \phi_t^{(3)}) \sim \text{Normal}(\mathbf{0}, \mathbf{\Sigma})$. The model is formulated in the Bayesian hierarchical framework, with weakly informative prior distributions for all parameters and implemented using NIMBLE, a fast and comprehensive package for flexible MCMC in R. We define the unobserved latent states z_t' as unobserved categorical quantities to be sampled.

3 Results

We assess model fit using posterior predictive checking, where for each MCMC sample we simulate a new set of random effects $\tilde{\phi}$ and then a set of rainfall values $\tilde{\mathbf{r}}$. From these "replicate" rainfall values we can then calculate summary statistics which describe important characteristics of the data. By computing these statistics from the original data $\mathbf{r}$ and comparing them to their respective replicate distributions, we assess whether they could have plausibly arisen from our model. This method is particularly appropriate for models intended to simulate new data, as we can investigate whether simulations display similar properties to the original data.

Here we focus on the model's ability to capture between-site correlation in occurrence and intensity. First we check the correlation in occurrence by computing the between-site Pearson correlation of $o_t^{(s)} = \mathbf{1}(r_t^{(s)} > 0)$. The posterior distributions of this statistic are shown for each site pair in the top-left of Figure 2. Secondly we check the overall between-site Pearson correlation of the rainfall values $r_t^{(s)}$, shown in the bottom-left of Figure 2. The model appears to slightly underestimate the correlation in occurrence across all three pairs but appears to capture the overall correlation quite

well. Notably, in both cases the distributions reflect the same ordering of strength of correlation as displayed by the data.
We can also check the set of probabilities of site i and site j both exceeding l, $p(r_t^{(i)} > l, r_t(j) > l)$, for varing l, as illustrated in Figure 2.(right). The model appears to capture the joint exceedance probabilities well for values less than 3mm (around 70% of the data) and for high values (e.g. >9mm). It does however underestimate the joint probabilities in-between (e.g. 5mm).

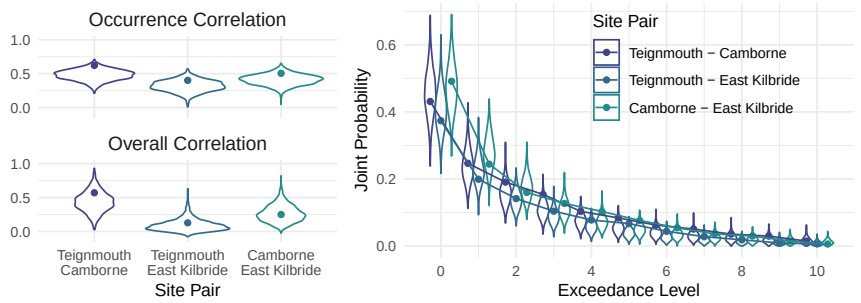


FIGURE 2. Replicate distributions of between-site Pearson correlations of rainfall occurrence (top-left), overall Pearson correlations (bottom-left) and joint exceedance probabilities (right). Points and lines show corresponding data values.

4 Conclusion

We presented a simple modelling framework for rainfall at multiple sites, where spatio-temporal dependence in rainfall occurrence is driven by a coupled hidden Markov model, and between-site dependence in rainfall intensity is driven by a Multivariate-Normal random effect. The model was sufficiently flexible to capture the different levels of correlation observed between the closer pair and the two more distant pairs. The model even captured the stronger correlation seen between Camborne and East Kilbride compared to the weaker correlation seen between Teignmouth and East Kilbride, which may be related to proximity to the Atlantic ocean. These statistics were not captured perfectly, but in most cases the absolute error was small, suggesting adaptation of the relatively simple model presented here, e.g. by introducing temporal non-homogeneity to the transition matrix, could well result in a capable model for rainfall at multiple sites.

References

Stoner, O. and Economou, T. (Under Review). An Advanced Hidden Markov Model for Hourly Rainfall Data. *Computational Statistics*

and Data Analysis.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Lang, D. and Bodik, R.(2017). Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *Journal of Computational and Graphical Statistics*, 26:2, 403-413.

Met Office (2020). MIDAS: UK Daily Rainfall Data. *NCAS British Atmospheric Data Centre*.

Enhanced variable selection for distributional regression

Annika Strömer[1], Leonie Weinhold[1], Christian Staerk[1], Stefanie Titze[2], Nadja Klein[3], Andreas Mayr[1]

[1] Department of Medical Biometrics, Informatics and Epidemiology, University of Bonn, Germany
[2] Department of Nephrology and Hypertension, FAU Erlangen-Nuremberg, Germany
[3] Humboldt-Universitat zu Berlin, Berlin, Germany

E-mail for correspondence: annika.stroemer@ukbonn.de

Abstract: We present an approach for enhanced variable selection for distributional regression via component-wise boosting. Boosting is an alternative method for fitting regression models and is applicable for high-dimensional data problems. Furthermore, the algorithm leads to data-driven variable selection. In practice, however, the algorithm still tends to select too many variables in some situations including false positives. This occurs particularly for low-dimensional data ($p < n$) in which case we observe a slow overfitting behavior. Due to the slow overfitting, the stopping iteration gets larger and more variables get included in the model. Many of the false positives are incorporated with a small coefficient and therefore have a small impact, but lead to a larger model with difficult interpretation. We try to overcome this issue by giving the algorithm the chance to de-select those variables. We consider the impact on variable selection and prediction and additionally compare the new approach to the One Standard Error Rule.

Keywords: Beta Regression; Variable Selection; Model-Based Boosting; GAMLSS.

1 Introduction

Beta regression is an alternative approach to model bounded outcome variables, as in this case the classical Gaussian regression may lead to biased results and requires variable transformations (Ferrari and Cribari-Neto, 2004). The beta distribution is characterized by the expected value μ and

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

the precision parameter ϕ. In context of distributional beta regression, which refers to a generalized additive model for location scale and shape (GAMLSS), we model μ and additionally ϕ in terms of several explanatory variables.
Here, we consider a component-wise gradient boosting algorithm with the great advantage of being able to process high-dimensional data problems with $p > n$. Furthermore, boosting leads to a data-driven variable selection which is controlled by the main tuning parameter: the number of boosting iterations m_{stop}. In each iteration only the best performing variable is selected and a base-learner that was once included in the model can not be de-selected. Different types of base learners can be used for each variable, reflecting the type of influence of the variable on the model. In the easiest case the base learners are simple linear models.

2 Enhanced variable selection

We address this issue with an approach that aims at eliminating those variables with a small impact and directly enforce the sparsity of the model. The general idea is to apply standard boosting which implicates early stopping and variable selection. We determine the variables with a minor importance for the model and de-select those components. Then we boost again with only the selected variables that survived. In our approach we consider the risk reduction as a measure for variable importance for every coefficient and de-select those variables with a small risk reduction. The selection is based on the likelihood for a variable in relation to the total difference. We de-select component j if

$$\sum_{k=1}^{m_{\text{stop}}} I(j = j^{*[k]})(r^{[k-1]} - r^{[k]}) < \tau \frac{r^{[0]} - r^{[m_{\text{stop}}]}}{\sum_{j=1}^{p} I(\hat{\beta}_j \neq 0)}$$

with the indicator function I and a threshold $\tau \geq 0$. Furthermore, $r^{[k-1]} - r^{[k]}$ represents the risk reduction and $j^{*[k]}$ denotes the updated variable index in iteration k. $\hat{\beta}$ is the estimated regression coefficient vector after m_{stop} boosting iterations.
An alternative approach is the One Standard Error Rule (oSE) which leads to an earlier stopping to obtain sparser models. The smaller m_{stop} the fewer variables are included in the model, because only one variable is updated in each iteration. This concept has already been used in context of penalized regression and regression trees and the aim here is to choose a smaller stopping iteration that is still within the range of one standard error of the risk for the optimal iteration (Breiman et al., 1984).

TABLE 1. Mean (sd) number of variables for the parameters μ and ϕ.

Model	μ	ϕ
boosted model	26.55 (6.97)	14.47 (5.74)
de-selected ($\tau = 0.5$)	12.87 (2.54)	6.23 (2.08)
de-selected ($\tau = 1$)	7.49 (2.54)	3.64 (1.65)
de-selected ($\tau = 1.5$)	4.12 (2.09)	2.30 (1.22)
oSE	7.92 (2.63)	2.78 (1.53)

3 Quality of life of chronic kidney disease patients

To illustrate the effect of the enhanced variable selection we consider the German Chronic Kidney Disease Study (GCKG). This is an ongoing cohort study with 3522 observations with stage III chronic disease and 54 explanatory variables. We want to select the most informative variables for the quality of life of chronic kidney disease patients (Mayr et al., 2018). For the analysis we use the R add-on package `betaboost`. For categorical variables we apply categorical effects and for continuous variables we incorporate spline effects as base learners. For the analysis we draw 500 bootstrap replicates and, on each sample, we fitted a beta regression model without and with enhanced variable selection for different threshold parameters τ. Additionally we consider the oSE for comparison.

Table 1 displays the mean (standard deviation) number of selected variables for μ and ϕ for the different models. We used three different values for the threshold parameter. One can observe that more variables are included for the expected value than for the precision parameter. Furthermore, the models with enhanced variable selection already show a significant decrease of selected variables for small threshold parameters. In comparison to the model with enhanced variable selection for $\tau = 1.5$ the oSE leads to larger models.

If one considers in this context the negative log-likelihood on test-data (out-of-bag observations) in Figure 1 then the oSE performs worse in relation to the other models. The smallest negative log-likelihood is obtained by the classical boosting model. As expected, the more variables are removed from the model, the worse the prediction accuracy. However, already for a small threshold parameter of $\tau = 0.5$ we can reduce the number of variables by more 50% without drastically decreasing the prediction accuracy of the resulting model.

4 Conclusions

The presented new approach for enhanced variable selection is a way to obtain sparser models with simpler interpretation. The prediction accu-

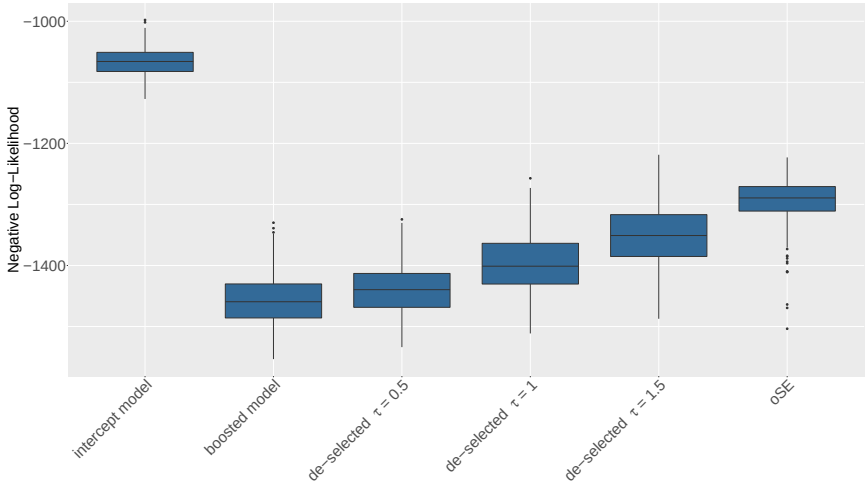


FIGURE 1. Negative log-likelihood of the different models on test-data.

racy usually does not improve but can lead to comparable accuracy with less predictors. In practice one needs to specify or optimize (e.g. via cross-validation) the additional parameter τ. We compared our approach with the One Standard Error Rule which does not only tackle the sparsity issue but also leads to additional shrinkage. The oSE can create smaller models and prevent the risk of overfitting, but it might also remove informative variables from the model that were selected late. This has a negative influence on the prediction accuracy.
In this data example we obtained sparser models with the enhanced variable selection and the oSE, but we even have a better accuracy for the enhanced variable selection for all considered threshold parameters than with the oSE.

Acknowledgments: The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, D-046.0078).

References

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Ferrari, S.L.P and Cribari-Neto, F. (2004). Beta-regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799 – 815.

Mayr, A., Weinhold, L., Hofner, B., Titze, S., Gefeller, O. and Schmid, M. (2018). The betaboost package - a software tool for modelling

bounded outcome variables in potentially high-dimensional epidemiological data. *International Journal of Epidemiology*, **47**(5), 1383 – 1388.

Serial correlation structures in latent linear mixed models for analysis of multivariate longitudinal ordinal responses

Trung Dung Tran[12], Emmanuel Lesaffre[12], Geert Verbeke[12], Geert Molenberghs[12]

[1] I-BioStat, KU Leuven, Leuven, Belgium
[2] I-BioStat, Universiteit Hasselt, Hasselt, Belgium

E-mail for correspondence: trungdung.tran@kuleuven.be

Abstract: We propose a latent linear mixed model to analyze multivariate longitudinal data of multiple ordinal variables, which are manifestations of fewer continuous latent variables. We focus on the latent level where the effects of observed covariates on the latent variables are of interest. We incorporate serial correlation into the variance component rather than assuming independent residuals. We show that misleading inference may be drawn when misspecifying the variance component. We apply our proposed model to examine the treatment effect on patients with the amyotrophic lateral sclerosis (ALS) disease. The result shows that the treatment can slow down the decline of cervical and lumbar functions.

Keywords: Linear mixed model; Ornstein-Uhlenbeck process; Serial correlation.

1 Introduction

In many multivariate longitudinal data analyses, the observed outcomes are considered to be manifestations of one or more underlying latent characteristics. Interest is then often in the effect of covariates (e.g. treatment) on the latent characteristics.
The progression of the disease ALS as a result of changes in latent neurological characteristics is an example. After being affected by the disease, patients gradually lose their ability of performing daily activities. This decline in ability is supposed to be related to impairments at three neurological regions of the central nervous system: bulbar region of the brain, cervical portion, and lumbar portion of the spinal cord (Wijesekera and

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

Nigel Leigh, 2009). Wang and Luo (2017) fit a random intercept model with independent residuals to examine the treatment effect on the latent neurological functions. The model makes a strong and probably unrealistic assumption, i.e., the correlations between latent repeated measurements is constant, regardless of time interval. However, Tran et al (2020) show that, for each neurological function, the correlation between any two (latent) measurements is a decreasing function of time interval.
Therefore, we propose a latent linear mixed model that takes serial correlation into consideration. Our model consists of two integrated components for two connected levels: (i) a polytomous item response theory (IRT) model is used to link the responses to the latent variables (LVs), and (ii) a linear mixed model (LMM) is used to connect continuous longitudinal LVs to observed covariates. Our extension incorporates an Ornstein-Uhlenbeck (OU) process (e.g. Tran et al, 2020) into the variance component of the LMM by decomposing the variance component into random effects, serial correlation, and independent residuals.

2 Model specification

Suppose that K variables (items) on N individuals are recorded repeatedly over time. Let Y_{ijk} be the observed response for the k^{th} item of the i^{th} individual at time t_{ij} where $i = 1, ..., N$, $j = 1, ..., n_i$, $k = 1, ..., K$ with n_i the number of occasions for individual i. The observed items are assumed to manifest R LVs, $\xi_{ij} = (\xi_{ij1}, ..., \xi_{ijr}, ..., \xi_{ijR})^T$, linked together as follows:

$$h(P(Y_{ijk} \leq m)) = \theta_{km} - \lambda_k^T \xi_{ij} + a_{ik},$$

where $h(.)$ is a link function (typically a logit or probit), m ($0 \leq m \leq c_k - 2$) is some score of item k with c_k the number of categories, and θ_{km} and λ_k are item-specific cut-point and factor loading parameters, respectively. The cut-points $\{\theta_{km}\}$ are non-decreasing in m. The R-vector λ_k contains the factor loadings of the k^{th} variable on the LVs. Denote by Λ, a $K \times R$ matrix with λ_k^T as the k^{th} row, the factor loading matrix. Finally, $a_{ik} \sim N(0, \sigma_{ak}^2)$ is the random effect for item k of individual i. The incorporation of a_{ik} is to take local dependence into account (Tran et al, 2019).
For the LVs, we start with the following LMM:

$$\xi_{ijr} = \beta^{(r)T} \mathbf{x}_{ij} + \mathbf{b}_i^{(r)T} \mathbf{z}_{ij} + \varepsilon_{ijr} \tag{1}$$

for the r^{th} latent variable with $\beta^{(r)}$ and $\mathbf{b}_i^{(r)}$ a p- and q-vector representing fixed and random effects, respectively. Further, $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are a p- and q-vector of covariates, respectively. Joining over r in (1), we obtain:

$$\xi_{ij} = \mathcal{B}\mathbf{x}_{ij} + B_i \mathbf{z}_{ij} + \varepsilon_{ij}, \tag{2}$$

where $\mathcal{B}$ is a $R \times p$ matrix with r^{th} row equal to $\beta^{(r)T}$, B_i is a $R \times q$ matrix with r^{th} row equal to $\mathbf{b}_i^{(r)T}$, $\varepsilon_{ij} = (\varepsilon_{ij1}, ..., \varepsilon_{ijr}, ..., \varepsilon_{ijR})^T$, and $\text{vec}(B_i^T) \sim$

$N(\mathbf{0}, D)$, where vec(M) turns matrix M into a vector and D is a $Rq \times Rq$ covariance matrix. Denote by ε_i the partitioned $n_i R$-vector consisting of stacked ε_{ij} vectors. To allow serial correlation, we decompose ε_{ij} as $\varepsilon_{ij} = \varepsilon_{ij(1)} + \varepsilon_{ij(2)}$, where $\varepsilon_{ij(1)}$ arises from an OU process and $\varepsilon_{ij(2)}$ are independent residuals. Further, $\Omega_2 := \text{Cov}(\varepsilon_{ij(2)})$ is diagonal.
The OU process is then specified for $\varepsilon_{ij(1)}$:

$$\varepsilon_{i1(1)} = \delta_{i1} \sim N(\mathbf{0}, \Omega_1), \quad \varepsilon_{ij(1)} = e^{-\Gamma d_{ij}} \varepsilon_{i,j-1,(1)} + \delta_{ij} \ (\forall j > 1)$$
$$\text{with } \delta_{ij} \sim N(\mathbf{0}, \Omega_1 - e^{-\Gamma d_{ij}} \Omega_1 e^{-\Gamma^T d_{ij}}),$$

where $d_{ij} = t_{ij} - t_{i,j-1}$ and $e^M = I + \sum_{j=1}^{+\infty} \frac{M^j}{j!}$ denotes the matrix exponential where M is a square matrix with $M^j = M \times ... \times M$ (j times). Furthermore, Ω_1 and Γ are two $R \times R$ matrices, satisfying the following conditions: the real part of each eigenvalue of Γ is positive, $\Gamma\Omega_1 + \Omega_1\Gamma^T$ is a covariance matrix, and, Ω_1 is a covariance matrix (Tran et al, 2020). The specification above implies a mean structure, i.e., $\xi_{ij} \sim N(\mu_{ij}, \Omega_1 + \Omega_2)$ where $\mu_{ij} = \mathcal{B}\mathbf{x}_{ij} + B_i \mathbf{z}_{ij}$, and a dynamic structure for ξ_{ij}, i.e.,

$$\xi_{i,j} \mid \xi_{i,j-1} \sim N(\mu_{i,j} + e^{-\Gamma d_{ij}}(\xi_{i,j-1} - \mu_{i,j-1}),$$
$$\Omega_1 - e^{-\Gamma d_{ij}} \Omega_1 e^{-\Gamma^T d_{ij}} + \Omega_2 + e^{-\Gamma d_{ij}} \Omega_2 e^{-\Gamma^T d_{ij}}).$$

For identification, intercepts are not incorporated in $\mathcal{B}$ in (2) and the variances in Ω_1 are fixed at 1.

3 Application to the ALS dataset

The ALS Functional Rating Scale (ALSFRS) was developed to monitor disease progression by measuring those symptoms. The original ALSFRS contains ten items falling into four categories: bulbar (speech, salivation, and swallowing), fine motor (handwriting, cutting, and dressing), and gross motor (turning, walking, and climbing) function, and respiratory disability. Later, the respiratory category of ALSFRS was revised, replacing the single item by three new items.
A dataset of 300 subjects with 2911 observations was used for analysis. We considered nine items of three categories: bulbar, fine motor, and gross motor function, and Λ took the following form:

$$\begin{pmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_{42} & \lambda_{52} & \lambda_{62} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{73} & \lambda_{83} & \lambda_{93} \end{pmatrix}^T.$$

A number of models were fitted with the Stan software package. We started with model M1, analyzed in Wang and Luo (2017) where it only includes the random intercepts and independent residuals at the latent level. Time and treatment time interaction were included as covariates. We kept the

same mean structure and then fitted several extensions: M2 with local dependence, M3 with serial correlation, and M4 (our proposal) with these two components. Finally, we fitted M5 to check whether we can remove independent residuals in M4 and fitted M6 to see whether adding random slopes into M3 can represent the serial correlations.
Watanabe's information criterion (WAIC) for six models are 39818.2, 29826.5, 35841.6, 26903.0, 26898.6, and 27810.9, respectively. Comparing M1 to M2, the result asserts the incorporation of the local dependence component. Comparing M1 to M3, the result suggests that a serial correlation structure exists. Comparing M2 to M4, incorporation of a serial correlation component is still necessary for a model where local dependence is already included. Comparing M4 to M5, we might remove the independent residuals when the serial correlation is included. Finally, M5 fitted better than M6, i.e. the random effects cannot represent the serial correlations. Hence, inference was made based on M5.
The treatment effect on the evolution of the neurological functions are different across the models: no significant effects (M1 and M6), a significant effect on lumbar function (M2 and M3), and significant effects on cervical and lumbar functions (M4 and M5). Estimated values (not shown) for the treatment effect on the neurological functions under M5 show that the treatment significantly slows down the progression of cervical and lumbar functions by 0.11 and 0.12 unit per six months, respectively.

4 Discussion

By incorporating an OU process to model serial correlation, our proposal can be considered as a combination of Wang and Luo (2017) and Tran et al (2020) where it addresses both the mean and dynamic structures. As seen in Section 3, our proposal corrects for misleading inference that we may make when ignoring the serial correlation structure.

References

Tran, T.D., Lesaffre, E., Verbeke, G., and Duyck, J. (2019). Modeling local dependence in latent vector autoregressive models. *Accepted in Biostatistics.*

Tran, T.D., Lesaffre, E., Verbeke, G., and Duyck, J. (2020). Latent Ornstein-Uhlenbeck models for Bayesian analysis of multivariate longitudinal categorical responses. *Accepted in Biometrics.*

Wang, J., Luo, S. (2017). Multidimensional latent trait linear mixed model: an application in clinical studies with multivariate longitudinal outcomes. *Statistics in Medicine;* 36: 3244-3256

Wijesekera, L.C., Nigel Leigh, P. (2009). Amyotrophic lateral sclerosis. *Orphanet Journal of Rare Diseases*; 4: 3.

A Generalised Joint Count Data Regression Framework for Modelling Football Scores

Hendrik van der Wurp[1], Andreas Groll[1], Thomas Kneib[2], Giampiero Marra[3], Rosalba Radice[4]

[1] TU Dortmund University, Germany
[2] University of Göttingen, Germany
[3] University College London, United Kingdom
[4] City University of London, United Kingdom

E-mail for correspondence: `vanderwurp@statistik.tu-dortmund.de`

Abstract: We propose a versatile joint regression framework for count responses. The method is implemented in the R add-on package `GJRM` and allows for modelling linear and non-linear dependence through the use of several copulae. Motivated by a football application, an extension is proposed, which forces the regression coefficients of the marginal (linear) predictors to be equal via suitable penalisation. We investigate the method's empirical performance in both a simulation study and on FIFA World Cup data.

Keywords: Count data regression; FIFA World Cups; Football; Joint modelling; Regularisation.

1 Introduction

There are many data situations where bivariate (or even multivariate) counts are the end point of interest and a priori assuming independence between such variables may be questionable. In particular, in many team sports such as football, handball or ice hockey, one usually jointly observes the number of goals of both competing teams. These are certainly associated as the final scores are the outcome of many single game situations where the players of both teams are involved in.
In the present work, we present a flexible generalised joint regression framework for count responses. The dependence between the outcomes is modelled via means of copulae. Motivated by our case study, we also provide an extension of the method which enforces the linear regression coefficients of

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

the marginal predictors to be equal by introducing a penalty. This is particularly useful for modelling team sports data where the predictors of both competing teams are usually based on the same set of covariates whose effects are often assumed to be equal (e.g., Groll et al., 2018). The underlying method is incorporated in the `R` package `GJRM` (Generalized Joint Regression Modelling, Marra and Radice, 2019b).
For this purpose, we extend the modelling framework of Marra and Radice (2017). The proposed method's empirical performance will be evaluated in both a simulation study and an application to FIFA World Cup data.
For a more detailed version of this paper see van der Wurp et al. (2020).

2 Model structure and estimation approach

Let bivariate count data realisations $\boldsymbol{y}_i = (y_{i1}, y_{i2})^{\mathrm{T}}, i = 1, \ldots, n$, be given, together with certain covariates whose effects should be accounted for. We assume that the joint cumulative distribution function (cdf) $F(\cdot, \cdot)$ of the corresponding discrete outcome variables $Y_1, Y_2 \in \mathbb{N}_0$ can be expressed as

$$P(Y_1 \leq y_1, Y_2 \leq y_2) = C_\theta \left(P(Y_1 \leq y_1), P(Y_2 \leq y_2)\right) = C_\theta(F_1(y_1), F_2(y_2)) \,,$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are the marginal cdfs of Y_1 and Y_2, The bivariate copula function $C_\theta : (0,1)^2 \to (0,1)$ (which is independent from the marginals) with copula parameter θ determines the dependence structure. `GJRM` covers several different copulae, see Marra and Radice (2019a).
Next, we link the parameters of the two marginal distributions as well as of the copula parameter θ with sets of covariates of sizes p_1, p_2 and p_θ, respectively. Moreover, let the corresponding covariate vectors be denoted by $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$ and $\boldsymbol{x}^{(\theta)}$, including intercepts and/or dummy variables for categorical predictors. Exemplarily, for two Poisson margins with rate parameters λ_1 and λ_2 and a single-parameter copula function, we may have

$$\begin{aligned}
\log(\lambda_1) = \eta_1 \quad &= \beta_0^{(1)} + x_1^{(1)}\beta_1^{(1)} + \ldots + x_{p_1}^{(1)}\beta_{p_1}^{(1)} \quad &= (\boldsymbol{x}^{(1)})^{\mathrm{T}}\boldsymbol{\beta}^{(1)} \,, \\
\log(\lambda_2) = \eta_2 \quad &= \beta_0^{(2)} + x_1^{(2)}\beta_1^{(2)} + \ldots + x_{p_2}^{(2)}\beta_{p_2}^{(2)} \quad &= (\boldsymbol{x}^{(2)})^{\mathrm{T}}\boldsymbol{\beta}^{(2)} \,, \\
g(\theta) = \eta_\theta \quad &= \beta_0^{(\theta)} + x_{1\theta}^{(\theta)}\beta_{1\theta}^{(\theta)} + \ldots + x_{p_\theta}^{(\theta)}\beta_{p_\theta}^{(\theta)} \quad &= (\boldsymbol{x}^{(\theta)})^{\mathrm{T}}\boldsymbol{\beta}^{(\theta)} \,,
\end{aligned} \tag{1}$$

where $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}$ and $\boldsymbol{\beta}^{(\theta)}$ are p_1-, p_2- and p_θ-dimensional vectors of regression effects, respectively. Finally, $g(\cdot)$ is a link function whose choice depends on the employed copula (see Marra and Radice, 2019a). We would like to stress that the equations in (1) represent a substantial simplification of the possibilities allowed for in the proposed modelling framework. In particular, our implementation allows to include non-linear functions of continuous covariates, smooth interactions between continuous and/or discrete variables and spatial effects, to name but a few. For this purpose, the penalised regression spline approach was adopted and the reader is referred

to, e.g., Marra and Radice (2017) for some examples. Due to the specific type of penalisation employed in this paper (see the next section), in this work we focus on linear effects. Simultaneous estimation of all parameters is based on maximising the model's log-likelihood $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta} := ((\boldsymbol{\beta}^{(1)})^{\mathrm{T}}, (\boldsymbol{\beta}^{(2)})^{\mathrm{T}}, (\boldsymbol{\beta}^{(\theta)})^{\mathrm{T}})^{\mathrm{T}}$. The fitting algorithm is based on an iterative trust region algorithm. The `GJRM` infrastructure allows to incorporate any quadratic penalty of the form $\frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}$, where $\boldsymbol{S}$ is a penalty matrix.

3 A penalty approach for football data

The models used later on in Section 4 and 5 are based on $F(y_1, y_2|\boldsymbol{\vartheta}) = C\left(F_1(y_1|\lambda_1), F_2(y_2|\lambda_2); \theta\right)$, with marginals $Y_1 \sim \text{Poi}(\lambda_1)$, $Y_2 \sim \text{Poi}(\lambda_2)$ modelling the number of goals scored by team $j \in \{1, 2\}$. The expected number of goals for team j in a match i is given by

$$\lambda_{ij} = \exp\left(\beta_0^{(j)} + x_{i1}\beta_1^{(j)} + \ldots + x_{ip}\beta_p^{(j)}\right),$$

with $i = 1, \ldots, n, j = 1, 2$. Although inclusion of covariate information into θ is possible in `GJRM`, for simplicity, in the following the copula parameter θ is specified as function of an intercept $\beta_0^{(\theta)}$ only. This way, we additionally achieve explicit comparability of dependence strengths in terms of Kendall's τ among different copula functions.

In contrast to the setting of the equations in (1), in football it is sensible to consider the same set of covariates for both competing teams (i.e., $p_1 = p_2 =: p$). Specifically, assuming covariates that are ordered such that $x_{ir}^{(1)}$ and $x_{ir}^{(2)}, r = 1, \ldots, p$, correspond to the same regressors, we would like to achieve $\beta_r^{(1)} = \beta_r^{(2)}\ \forall r$. Without this restriction, being first- or second-named team could affect the estimation of $\beta_r^{(j)}$ and thus make the interpretation of the coefficients questionable, as stressed in Groll et al. (2018). To obtain (virtually) equal coefficients for both margins, we propose to use the following penalised version of the log-likelihood $\ell(\boldsymbol{\beta})$, i.e.

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\xi \sum_{r=0}^{p} w_j \left(\beta_r^{(1)} - \beta_r^{(2)}\right)^2, \tag{2}$$

where the ridge-type penalty acts on the differences of the pairs of coefficients corresponding to the same covariates, with suitably chosen weights w_j and penalty parameter ξ. This penalty can be easily incorporated in

`GJRM` via a suitably designed penalty matrix $\boldsymbol{S}$, which is equal to

$$\boldsymbol{S} = \xi \cdot \begin{pmatrix} \boldsymbol{w}^{\mathrm{T}} & \boldsymbol{w}^{\mathrm{T}} & 0 \\ \vdots & \vdots & \vdots \\ \boldsymbol{w}^{\mathrm{T}} & \boldsymbol{w}^{\mathrm{T}} & 0 \\ 0 & \dots & 0 \end{pmatrix} \circ \left(\begin{array}{cccc|cccc|c} 1 & 0 & \dots & 0 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 & -1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & -1 & 0 \\ \hline -1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & \dots & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -1 & 0 & 0 & \dots & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right), \tag{3}$$

where '$\circ$' denotes the Hadamard matrix product and ξ is a tuning parameter controlling the strength of the penalty. The weights $\boldsymbol{w} = (w_0, w_1, \dots, w_p)^{\mathrm{T}}$ depend on the current fit $\hat{\boldsymbol{\beta}}^{[k]}$ from iteration k of the algorithm. In order to shrink all paired differences jointly to zero, we use $w_j = \left|\hat{\beta}_j^{(1)} - \hat{\beta}_j^{(2)}\right|$, suppressing the iteration index for notational convenience.

4 Simulation Study

A first simulation study with covariates $x_1, \dots, x_6 \sim \mathcal{U}[0,1]$ for $n = 250$ bivariate observations was carried out. For each observation, λ_{i1} and λ_{i2} were specified via $\lambda_{ij} = \exp(\beta_0^{(j)} + \boldsymbol{x}_i^{(j)}(\boldsymbol{\beta}^{(j)})^{\mathrm{T}})$, where $\boldsymbol{x}_i^{(1)} = (x_{i1}, x_{i2}, x_{i3})^{\mathrm{T}}$ and $\boldsymbol{x}_i^{(2)} = (x_{i4}, x_{i5}, x_{i6})^{\mathrm{T}}$. Each pair of outcomes (y_{i1}, y_{i2}) is sampled from a given copula with marginal Poisson parameters λ_{i1} and λ_{i2}. For each copula, the respective θ is determined by fixed values of Kendall's τ. We define two different sets of coefficients, i.e. $\boldsymbol{\beta}^{(1)} \neq \boldsymbol{\beta}^{(2)}$, and create 100 datasets with aforementioned associations from different copula classes. The penalisation approach from Section 3 is not yet applied. Mean squared errors (MSEs) for the regression coefficients are used as goodness-of-fit measure. Setups with different dependency strengths were checked. As expected, the fits based on the correct copula class always yield best results (see Figure 1). In a second simulation study, we investigate our specific penalty structure. We therefore chose coefficients $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ that are equal in both margins. Only the true respective copula classes were used to fit the models. Figure 2 compares the perfomance of fits with our new penalty from Section 3 and without. It turns out that the proposed penalisation approach is essentially useful if the true coefficients can be assumed to be equal, which is specifically realistic in competitive setups such as sports competitions.

5 Application to FIFA World Cup data

We now apply our method to FIFA World Cup data from 2002 to 2018. The basic data set was described in detail in Groll et al. (2019). Using

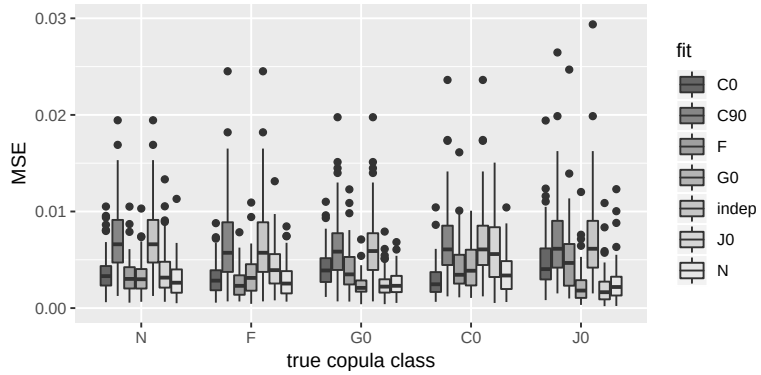


FIGURE 1. Results for MSE of the regression coefficients for different true copulae with each copula parameter θ derived from $\tau = 0.7$.

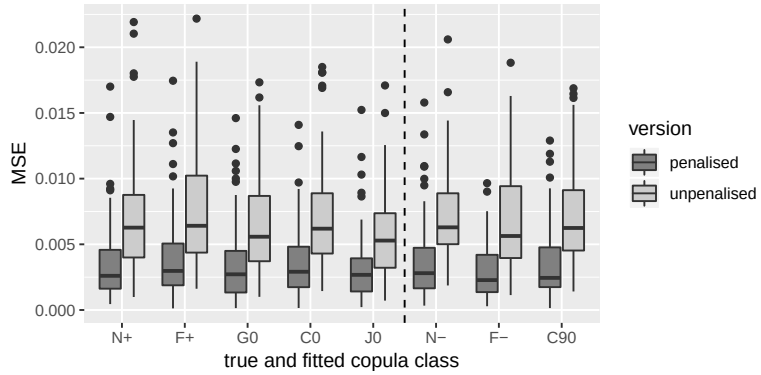


FIGURE 2. Results for the MSE of the regression coefficients obtained using the penalised (left boxes) and unpenalised (right boxes) estimation approaches for a set of different copulae and associations; $\tau = 0.25$ for copula N, F, G0, C0, J0 and $\tau = -0.25$ for copulae N, F, C90.

the fitted model we obtain probabilities for each result (y_1, y_2), from which probabilities for the three-way-results *win*, *draw*, *loss* can be deduced. A cross-validation-type strategy over the five tournaments is applied to compare the predictive performance of copula classes available in `GJRM` using different measures, i.e. MSE on the number of goals, the averaged likelihood, classification rate, and rank probability score (RPS), all on three-way-outcomes. An excerpt of the results can be found in Table 1. The F (Frank) and FGM (Farlie-Gumbel-Morgenstern) copula classes were found to deliver the best fits considering the ranked results. With copula parameters leading to Kendall's $\hat{\tau}$ values of about 0.1, both models indicate a weak positive correlation structure, which is compatible to current literature.
The penalised approach lead to substantially better results regarding all

TABLE 1. Results of selected measures for model fits based on different copulae obtained using the penalised and unpenalised approaches.

	RPS		likelihood		class. rate		MSE	
Copula	pen	unpen	pen	unpen	pen	unpen	pen	unpen
N	0.196	0.210	0.403	0.395	0.522	0.506	1.421	1.490
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
F	0.196	0.210	0.405	0.396	0.512	0.494	1.421	1.487
FGM	0.196	0.210	0.404	0.396	0.512	0.491	1.420	1.486
indep	0.198	0.211	0.398	0.390	0.531	0.509	1.419	1.486

chosen measures and all copula classes. Hence, the assumption of equal coefficients seems to be justified and our penalty useful in this context.
Future research will address several extensions. Firstly, the penalty discussed here could be extended to the context of more complex predictor structures (allowing, e.g., for non-linear effects via P-splines). Moreover, we believe that the method's predictive performance can be further improved by penalising covariate effects via LASSO-type penalties or via boosting.

References

Groll, A., C. Ley, H. Van Eetvelde, and G. Schauberger (2019). A hybrid random forest to predict soccer matches in international tournaments, Journal of Quantitative Analysis in Sports. *Journal of Quantitative Analysis in Sports*, **15**(4), 271-287.

Groll, A., T. Kneib, A. Mayr and G. Schauberger (2018). On the dependency of soccer scores - A sparse bivariate Poisson model for the UEFA European Football Championship 2016. *Journal of Quantitative Analysis in Sports*, **14**(2), 65-79.

Marra, G. and R. Radice (2017). Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis*, **112**, 99-113.

Marra, G. and R. Radice (2019a). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association.*

Marra, G. and R. Radice (2019b). *GJRM: generalised joint regression modelling.* R package version 0.2.

van der Wurp, H., A. Groll, T. Kneib, G. Marra and R. Radice (2020). Generalised Joint Regression for Count Data - A Penalty Extension for Competitive Settings. *Statistics and Computing*, to appear.

Bayesian modelling of treatment effects on panel outcomes

Helga Wagner[1], Sylvia Frühwirth-Schnatter[2], Liana Jacobi[3]

[1] Department of Applied Statistics, Johannes Kepler University Linz, Austria
[2] Department of Finance, Accounting and Statistics, Vienna University of Economics and Business, Austria
[3] Department of Economics, The University of Melbourne, Australia.

E-mail for correspondence: helga.wagner@jku.at

Abstract: We consider estimation of the effect of a binary treatment on a continuous outcome observed over subsequent time periods. We propose a new, flexible model that separates longitudinal association of the outcomes from association due to endogeneity of treatment selection and hence allows for unbiased estimation of dynamic treatment effects. We investigate the performance of the proposed method on simulated data and employ it to analyse the effects of a long maternity leave on earnings of Austrian mothers after their return to the labour market.

Keywords: endogeneity, bifactor model; switching regresson model; shared factor model, dynamic treatment effects

1 Introduction

Identification and estimation of treatment effects is an important issue in many fields, e.g. to evaluate the effectiveness of social programs, government policies or medical interventions. As each subject is observed either under control conditions or under treatment, the outcome difference which would allow straightforward estimation of treatment effects is not available for any particular subject. Additionally, for data from observational studies, endogeneity of treatment selection can cause unobserved confounding and bias of treatment effects estimates if not adequately accounted for.
Bayesian approaches to inference on treatment effects rely on specifying a joint model of treatment selection and the two potential outcomes (under control conditions and under treatment), of which only one is observed for each subject. To estimate the effect of a binary treatment on a continuous

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

outcome observed over subsequent time periods two models, the switching regression model (Chib and Jacobi, 2007) and the shared factor model (Carneiro et. al., 2003), have been suggested so far. Both approaches rely on a binary regression model for selection into treatment and two multivariate regression models for the outcome sequences under control and under treatment. They differ, however, with respect to the modeling of the dependence across these regression models: whereas Carneiro et al. (2003) model the association between treatment selection and both potential outcome sequences via shared latent factors, Chib and Jacobi (2007) specify only two marginal models for selection into treatment and one sequence of potential outcomes, under treatment and control respectively, but leave the joint distribution of the two potential outcomes sequences unspecified. Jacobi et.al. (2016) show for simulated data that both models can lead to biased treatment effects estimates if the assumptions on the correlation structure of treatment selection and the two potential outcomes sequences of the model used for data analysis are violated for the data generating process.

We propose a novel, flexible model that allows to separate longitudinal association of the outcomes sequences from association due to endogeneity of treatment selection and investigate its performance on simulated data. We employ the proposed model to re-analyse the effects of a long maternity leave on earnings of Austrian mothers previously analysed in Jacobi et.al. (2016).

2 Model specification

To specify a joint model for selection into treatment and the potential outcomes under control conditions and under treatment at time point $t = 1, \ldots, T$, we denote by x_i the binary treatment status of subject $i = 1, \ldots, n$ and by $y_{0,it}$ and $y_{1,it}$ the potential outcomes of this subject under control ($x_i = 0$) and under treatment ($x_i = 1$) respectively.

Treatment selection is allowed to differ with covariates via a probit model for x_i, which can be specified in terms of a latent Gaussian random variable x_i^* as

$$x_i^* = \mathbf{v}_i' \boldsymbol{\alpha} + \varepsilon_{xi}, \tag{1}$$

$$x_i = I_{\{x_i^* > 0\}}, \tag{2}$$

where ε_{xi} has a Normal distribution with mean 0, $\mathbf{v}_i$ denotes a vector of covariates and $\boldsymbol{\alpha}$ their effect on treatment selection.

The selection model given in equations (1) and (2) is combined with a model for the potential outcomes at time points $t = 1, \ldots, T$,

$$y_{0,it} = \mu_t + \mathbf{w}_{it}' \boldsymbol{\gamma} + \varepsilon_{0,it}, \tag{3}$$

$$y_{1,it} = (\mu_t + \kappa_t) + \mathbf{w}_{it}' (\boldsymbol{\gamma} + \boldsymbol{\theta}) + \varepsilon_{1,it}, \tag{4}$$

where the structural mean of the potential outcomes can depend on covariates $\mathbf{w}_{it}$ with effects $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma} + \boldsymbol{\theta}$ under control and under treatment conditions, respectively. For a subject with $\mathbf{w}_{it} = \mathbf{0}$ the expected outcome at time point t is μ_t and $\mu_t + \kappa_t$, respectively and $\varepsilon_{j,it}$ denotes the error term under control ($j = 0$) and treatment conditions ($j = 1$).
We are interested in the average longitudinal treatment effect for a subject with covariate values $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_T)'$, which is given as

$$\mathrm{ATE}_T(\mathbf{W}) = E(\mathbf{y}_{1,i} - \mathbf{y}_{0,i}|\mathbf{W}) = \boldsymbol{\kappa} + \mathbf{W}\boldsymbol{\theta},$$

where $\mathbf{y}_{j,i} = (y_{j,i1}, \ldots, y_{j,iT})$, $j = 0, 1$ are the potential outcome vectors and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_T)'$. Unbiased estimation of $\mathrm{ATE}_T(\mathbf{W})$ is straightforward based on unbiased estimates for $\boldsymbol{\kappa}$ and $\boldsymbol{\theta}$, which, however, require correct specification of the dependence structure of the error terms $\boldsymbol{\varepsilon}_i = (\varepsilon_{xi}, \boldsymbol{\varepsilon}_{0i}, \boldsymbol{\varepsilon}_{1i})$, where $\boldsymbol{\varepsilon}_{ji} = (\varepsilon_{j,i1}, \ldots, \varepsilon_{j,iT})$, $j = 0, 1$.
Hence we propose a flexible model for the association of treatment selection and the potential outcomes sequences. In the spirit of the bifactor model introduced by Holzinger and Swineford (1937) we assume that all dependencies in the error vector $\boldsymbol{\varepsilon}_i$ are captured by three subject specific latent factors. The common factor f_{ci} shared by the error terms of the latent utility x_i^* and both potential outcome vectors $\mathbf{y}_{0,i}$ and $\mathbf{y}_{1,i}$ accounts for unobserved confounding. Further two outcome specific factors $f_{0,i}$ and $f_{1,i}$ capture the additional longitudinal association in the outcome vectors that cannot be attributed to unobserved confounders. The joint model for the error terms is thus specified as

$$\varepsilon_{xi} = \lambda_x f_{ci} + \epsilon_{xi}, \qquad \epsilon_{xi} \sim \mathcal{N}(0, 1), \tag{5}$$

$$\boldsymbol{\varepsilon}_{0i} = \boldsymbol{\lambda}_0 f_{ci} + \boldsymbol{\zeta}_0 f_{0i} + \boldsymbol{\epsilon}_{0i}, \qquad \epsilon_{0,it} \sim \mathcal{N}(0, \sigma_{0t}^2), \tag{6}$$

$$\boldsymbol{\varepsilon}_{1i} = \boldsymbol{\lambda}_1 f_{ci} + \boldsymbol{\zeta}_1 f_{1i} + \boldsymbol{\epsilon}_{1i}, \qquad \epsilon_{1,it} \sim \mathcal{N}(0, \sigma_{1t}^2), \tag{7}$$

where the factors are assumed to be independent standard Normals. Hence, the factor loadings λ_x, $\boldsymbol{\lambda}_j$ and $\boldsymbol{\zeta}_j$, $j = 0, 1$ determine the joint variance-covariance matrix of all error terms.
This model avoids drawbacks of both the shared factor and the switching regression model. As correlation across panel outcomes is not attributed solely to the general factor it is more flexible than the shared factor model, which is recovered as the special case where $\boldsymbol{\zeta}_j = \mathbf{0}$ for $j = 0, 1$. Without an assumption on the joint distribution of the specific factors f_{0i} and f_{1i} the model is a switching regression model with the advantage that conditional on the latent factors the errors of latent utility and each potential outcome are independent.

3 Simulation Study

To illustrate the flexibility of the proposed bifactor model we analyse two data sets from Jacobi et.al (2016), which were simulated from the shared

factor model (SF) and the switching regression model (SR) respectively.

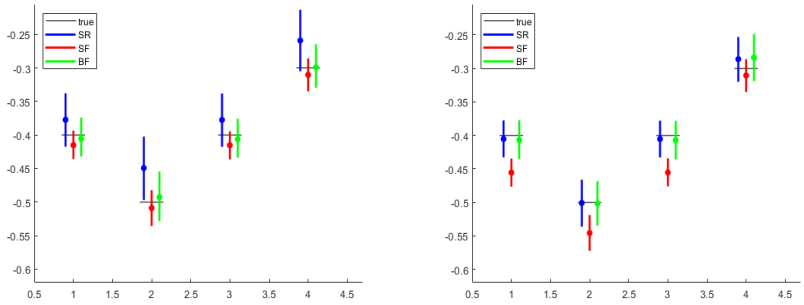


FIGURE 1. True and estimated insample average treatment effects with 95%-posterior intervals for data simulated from the shared factor model (left) and the switching regression model (right).

The average treatment effects ATE_t over all simulated subjects was estimated as

$$\widehat{\text{ATE}}_t = \hat{\kappa}_t + \frac{1}{n}\sum_{i=1}^{n} \mathbf{w}_{it}'\hat{\boldsymbol{\theta}}$$

via these two models as well as the bifactor model (BF).
Figure 1 shows for $t = 1, \ldots, 4$ the true average treatment effect and the estimated average treatment effect under each model with the 95%- posterior intervals. These intervals do not include the true average treatment effect at all time points if data generated from the SF model are analysed with the SR model or data generated from the SR model are analysed with the SF model. In contrast the proposed bifactor model always performs similar as the data generating model.

4 Analysing Earnings Effects of Maternity Leave

We apply the bifactor model to re-analyse the effects of a long maternity leave on earnings of Austrian mothers after their return to the labor market using the same data as Jacobi et. al. (2016). The analysis is based on data from the Austrian Social Security Data Base (ASSD), which is an administrative data set of the universe of Austrian employees providing detailed information on employment and maternity leave spells as well as demographic information on mothers (Zweimüller et. al., 2009), and a second data set collected as basis for wage taxes.
To exploit a change in the parental leave policy in Austria in July 2000 which extended the payment of parental leave benefits from 18 to 30 months Jacobi et.al. (2016) used data for mothers who gave birth to their last

child from June 1998 til July 2002. Figure 2 illustrates that the majority of mothers returned to the labour market within 18 months before this policy change whereas afterwards most mothers took a longer maternity leave of more than 18 months. Jacobi et.al. (2016) defined treatment as a maternity

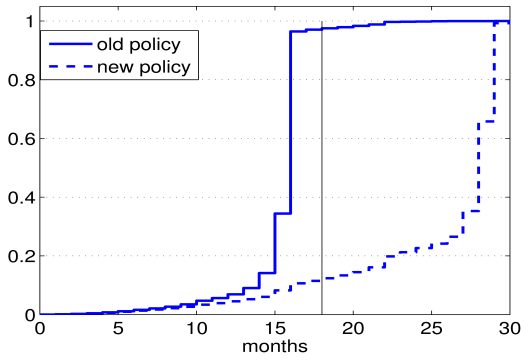


FIGURE 2. Empirical cdf of the duration of maternity leave after child birth

leave longer than 18 months and analysed the treatment effect on earnings for those mothers who returned to the labour market immediately after the maternity leave with the shared factor model as well as the switching regression model. We use the same specification for the mean of the latent utility and the potential outcomes, defined as the log income, as in their analysis, but model the joint distribution of the errors ε_i by the bifactor model.

Figure 3 shows the estimated average treatment effect over all mothers in the sample for the first 6 years after return to the labour market from the three models. In all models a long maternity leave results in considerably lower earnings in the first panel period with the gap decreasing over time. However, the evolvement of $\widehat{\text{ATE}}_t$ is slightly different for the three models: it is still negative in panel period 6 for the shared factor, positive for the switching regression model and practically zero for the bifactor model.

5 Conclusion

Inference on treatment effects for longitudinally observed outcomes can be biased when the model used for data analyses implies restrictions on the association between selection into treatment and the potential outcomes sequences as well as within the potential outcomes sequences which are violated for the data to be analysed. The proposed bifactor model explicitly models these associations by latent factors and is more flexible than the models used so far and hence allows better inference on dynamic treatment effects.

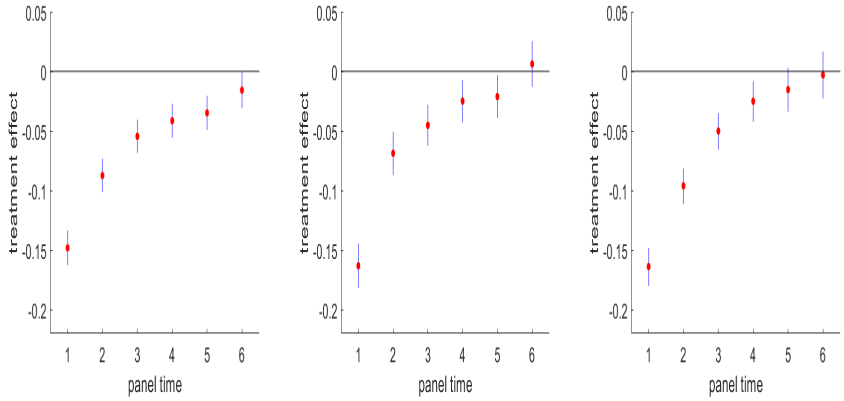


FIGURE 3. Estimated average treatment effects and 95% HPD - intervals of a long maternity leave. Analysis with the shared factor model (SF, left), the switching regression model (SR, middle) and the bifactor model (BF, right).

References

Carneiro P., Hansen K. T. and Heckman, J. J. (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty of college choice. *International Economic Review*, **44**, 361 – 422.

Chib, S. and Jacobi, L. (2007). Modeling and calculating the effect of treatment at baseline from panel outcome. *Journal of Econometrics*, **140**, 781 – 801.

Holzinger K. and Swineford F. (1937). The Bi-factor method. *Psychometrika*, **2**, 41 – 54.

Jacobi L., Wagner H. and Frühwirth-Schnatter S. (2016). Bayesian Treatment Effects Models with Variable Selection for Panel Outcomes with an Application to Earnings Effects of Maternity Leave. *Journal of Econometrics*, **193:1**, 234 – 250.

Zweimüller J., Winter-Ebmer R., Lalive R., Kuhn A., Wuellrich J.-P., Ruf O., and Büchi S. (2009). The Austrian Social Security Database (ASSD). Working paper 0903, *NRN: The Austrian center for labor economics and the analysis of the welfare state, Linz, Austria*

Multivariate spatial models for lattice data in complex surveys

Kevin Watjou[1], Christel Faes[1]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BIOSTAT), Data Science Institute, Hasselt University, Belgium

E-mail for correspondence: kevin.watjou@uhasselt.be

Abstract: When performing an analysis of a spatial health survey it often interesting to investigate multiple diseases. Furthermore, the correlation between these diseases could provide valuable information when analysing the data. We propose a joint spatial model which incorporates the survey weights in the modeling process while taking into account the correlation structure between the diseases.

Keywords: Joint Disease Mapping; Complex Survey Design; Hierarchical Spatial Modeling

1 Introduction

When performing small area estimation (SAE), the geographical distribution of diseases is the main subject of interest. These studies are often performed in a complex survey setting. Recently, efforts have been made to incorporate the survey design in the spatial estimation process (Mercer et al. (2014), Watjou et al. (2017)). However, this was usually done in a univariate framework. Since some diseases can share common risk factors, a multivariate spatial model can account for this correlation structure between the diseases. Many contributions have already been made in the context of bivariate spatial modelling. Dabney and Wakefield (2005) presented a comprehensive comparison between univariate and bivariate disease mapping models and considered the benefits and issues when performing the latter. Crainiceanu, Diggle and Rowlingson (2008) presented bivariate binomial geostatistical model in order to map the prevalence of Loa loa. Knorr-Held and Best (2001) proposed a spatial shared component model, separating the underlying risk surfaces for each disease using a shared and disease-specific

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

component. Several authors have proposed methodology which includes the complex survey design when working with general multivariate models. Asparouhov and Muthen (2005) demonstrated how the pseudo maximum likelihood could be extended to multistage stratified cluster sampling designs. In this study, we extend this approach taking into account geographical associations. We investigate the joint geographical distributions of asthma and chronic obstructive pulmonary disease (COPD) using the 2013 Florida Behavioral Risk Factor Surveillance System (BRFSS) health survey. The primary objective is to develop a joint bivariate spatial model, taking into account the effect of the complex design of the BRFSS survey.

2 Methodology

Let $Y_{ik,l}$ be the binary health outcome of disease l for an individual i in county k (i = 1,...,N_k, k = 1,...,K, l = 1,2), where N_k is the population size in county k. We want to estimate the true county-specific prevalence for disease l: $P_{k,l} = \sum_{i=1}^{N^k} Y_{ik,l}$. Each health outcome is accompanied by a survey weight w_{ik}, which is independent of disease l. The survey weights were recalibrated to account for post-stratification at the area level and normalized to sum up to the observed sample size.

2.1 Model 1: Univariate models

Congdon and Lloyd (2010) described a weighted likelihood which could be employed in a spatial setting. Binary health outcomes $y_{ik,l}$ are weighted by the normalized weights $\tilde{w}^*_{ik}$. This model is also called the pseudo-likelihood model. Mercer et al. (2014) remarked that this model could be written as a hierarchical model:

$$\begin{aligned}
\tilde{y}_{k,1}|P_{k,1} &\sim \text{Binomial}(m_k, P_{k,1}) \\
\tilde{y}_{k,2}|P_{k,2} &\sim \text{Binomial}(m_k, P_{k,2}) \\
\text{logit}(P_{k,1}) &= \beta_{0,1} + u_{k,1} + v_{k,1} \\
\text{logit}(P_{k,2}) &= \beta_{0,2} + u_{k,2} + v_{k,2},
\end{aligned}$$

$\tilde{y}_{k,l} = \sum_{i=1}^{m_k} y_{ik,l}\tilde{w}^*_{ik}$, $u_{k,l}$ is the spatially correlated random effect which follows an ICAR(0, $\sigma^2_{u,l}$)-distribution and $v_{k,l}$ is the spatially uncorrelated random effect which follows a N(0, $\sigma^2_{v,l}$)-distribution. For the precision parameters $\sigma^{-2}_{u,l}$ and $\sigma^{-2}_{v,l}$, a Gamma(0.5,0.008) prior distribution was assigned.

2.2 Model 2: Joint model

The univariate models do not account for the underlying correlation between the risk surfaces in the estimation process. We present a weighted

correlated random effects model which can take this correlation into account as follows:

$$\begin{aligned} \tilde{y}_{k,1}|P_{k,1} &\sim \text{Binomial}(m_k, P_{k,1}) \\ \tilde{y}_{k,2}|P_{k,2} &\sim \text{Binomial}(m_k, P_{k,2}) \\ \text{logit}(\mathbf{P_k}) &= \beta_\mathbf{0} + \mathbf{u_k} + \mathbf{v_k}, \end{aligned}$$

where $\mathbf{P_k} = \begin{pmatrix} P_{k,1} \\ P_{k,2} \end{pmatrix}, \beta_\mathbf{0} = \begin{pmatrix} \beta_{0,1} \\ \beta_{0,2} \end{pmatrix}$ and $\mathbf{u_k}$ and $\mathbf{v_k}$ are bivariately correlated ICAR and normal random effects respectively. The same prior distribution for the precision parameters was used as in section 2.2. The parameters ρ_u and ρ_v are the correlation coefficients for the spatially correlated and uncorrelated effect respectively.
The performance of both models was compared by means of the Deviance Information Criterion (DIC) (Spiegelhalter et al. (2002))

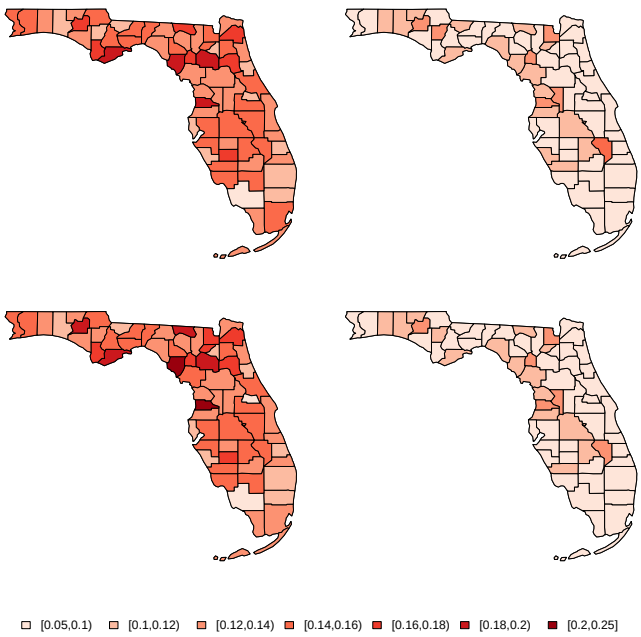


FIGURE 1. Map of estimated prevalences for asthma (left column) and COPD (right column) using the univariate models (top row) and the weighted correlated random effect model (bottom row).

3 Results

Figure 1 shows the geographic distribution for the prevalences of asthma and COPD for both the univariate and weighted correlated random effects model. While the results for both models seem similar, Model 2 performs slightly better in terms of DIC (1240.926) compared to Model 1 (1244.645).

4 Conclusion

We propose a joint spatial model which include the design weights of the BRFSS study in order to take the complex survey design into account in the estimation process. This weighted correlated random effects model performs better than the univariate models in terms of DIC. The former model has the added benefit that it can take the correlation between the two diseases into account.

References

Asparouhov, T. and Muthen, B. (2005). Multivariate statistical modeling with survey data. In: *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference*

Congdon, P., Lloyd, P., (2010). Estimating small area diabetes prevalence in the US using the behavioral risk factor surveillance system. *Journal of Data Science*, **8**, 235 - 252.

Crainiceanu, C.M., Diggle, P.J. and Rowlingson, B. (2008). Bivariate binomial spatial modeling of Loa loa prevalence in Tropical Africa. *Journal of the American Statistical Association*, **103** (481), 21 – 47.

Dabney, A.R. and Wakefield, J.C. (2005). Issues in the mapping of two diseases. *Statistical Methods in Medical Research*, **14**, 83 – 112.

Knorr-Held, L. and Best., N.G. (2001). A shared component model for joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, Series A*, **164** (1), 73 – 85.

Mercer, L., Wakefield, J.C., Chen, C. and Lumley, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics*, **8**, 69 – 85.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society.* **64** (4), 2002, 583 – 639.

Watjou, K., Faes, C., Lawson, A., Kirby, R.S., Aregay, M., Carroll, R. and Vandendijck, Y. (2017). Spatial small area smoothing models for handling survey data with nonresponse. *Statistics in Medicine*, **36** (23), 3708 – 3745.

A horseshoe based prior for shrinkage towards a predefined parametric subspace.

Paul Wiemann[1], Thomas Kneib[1]

[1] Göttingen University, Germany

E-mail for correspondence: pwiemann@uni-goettingen.de

Abstract: We introduce a new prior hierarchy that allows shrinking of smooth spline-based functional effects towards a predefined vector space of parametric functions. Instead of shrinking each spline coefficient towards zero, we adapt the horseshoe prior to control the deviation from the predefined vector space. Furthermore, the prior presented regularizes the wiggliness of the estimated effect. In this paper, we start with an application to energy consumption in Germany, then introduce the technical details and describe the prior's desirable shrinkage properties. We conclude with a simulation study to assess the validity of our approach.

Keywords: regression; spline; shrinkage; functional subspace; Bayesian;

1 Energy consumption in Germany over one day

Suppose the total energy consumption in Germany over one day needs to be estimated. The most naive approach might be a linear model based on a trigonometric polynomial of order Ω, i.e.

$$\mathrm{E}[y|x] = \beta_0 + \sum_{\omega=1}^{\Omega} \left[\beta_\omega \cos\left(\omega \frac{2\pi}{24} x\right) + \tilde{\beta}_\omega \sin\left(\omega \frac{2\pi}{24} x\right) \right], \tag{1}$$

where $x \in [0, 24)$ denotes the hour of the day. Another approach is the use of regression splines, as they can flexibly adapt to the data. The latter may have the disadvantage that the spline solution diverges from the parametric solution even though the parametric solution is preferred based on theoretical considerations, and the spline solution fits the data only negligibly better.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

In this paper, we present a prior for the spline coefficients that shrinks the implied function towards a predefined functional subspace, i.e., the trigonometric polynomial from above. Before we introduce the technical details, we demonstrate our method through a case study aiming at estimating the total energy consumption in Germany over one day.
For that, we use data freely available at `http://smard.de` and choose eight weekdays (only Mondays and Tuesdays) and eight weekend days from November 2018. The data is preprocess by rescaling and subtracting the daily mean. Then we specify the parameters in our prior such that it shrinks towards the polynomial from above with $\Omega = 4$. In addition, we fit the parametric approach from Equation (1) and a Bayesian P-Spline (Lang and Brezger, 2004).
A plot of the data together with the estimated functions is displayed in Figure 1. We observe that the proposed prior adopts to the data by shrinking the estimated effect to the parametric function for the weekend days and leave it basically untouched for the weekdays, i.e.; the estimated function of the new prior is similar to the P-Spline solution. This gets also reflected in the shrinkage coefficient κ that is almost one for the weekend, thus full shrinkage is observed. Weekend days can be modeled with the chosen trigonometric polygon while more flexibility is needed for weekdays. Our prior seems to negotiate well between both situations.

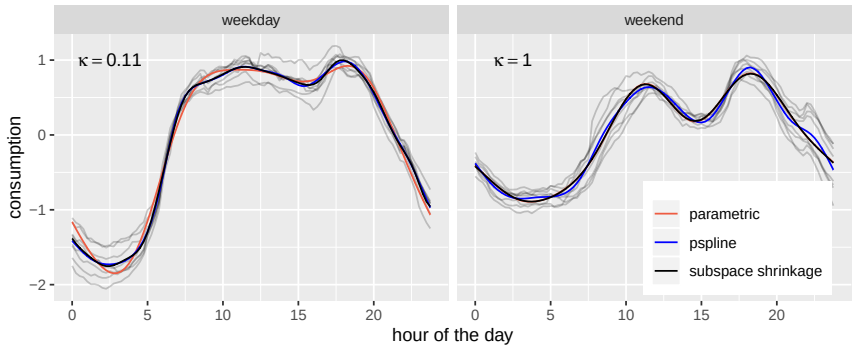


FIGURE 1. Plots of the total energy consumption in Germany for eight weekdays and eight weekend days in November 2018 (gray lines). Estimated posterior mean using the parametric approach (red), a Bayesian P-spline approach (blue) and the presented approach (black).

2 Introducing the prior

For $i = 1, \ldots, n$, consider the non-parametric regression problem

$$y_i = f(x_i) + \varepsilon_i$$

with iid error term $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, smooth function f, and continuous covariate x_i. A common approach for its estimation is to transform it into a semiparametric estimation problem by assuming that f can be approximated via a linear combination of k basis function evaluated for the covariate and denoted $\mathrm{B}_j(x)$. More precisely,

$$f(x) = \sum_{j=1}^{k} \mathrm{B}_j(x)\beta_j = \boldsymbol{B}(x)'\boldsymbol{\beta}$$

with $\boldsymbol{B}(x) = (\mathrm{B}_1(x), \ldots, \mathrm{B}_k(x))'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$. Thus, only a finite number of parameters need to be estimated that usually have no valid interpretation by themselves. In the following, let $\boldsymbol{Z}$ denote the $n \times k$ matrix of basis functions evaluated at the observed covariates $x_1, \ldots, x_n$. In this work, we employ equally spaced B-Splines of third order.
A constant or linear effect is usually the target of regularization or effect selection approaches in the context of regression splines (Lang and Brezger, 2004, Klein *et al.*, 2019).
With the proposed prior we pursue three goals:

- Instead of shrinking towards a constant or linear effect, a user-defined functional subspace is the target. Similar to Shin *et al.* (2019), this so-called null space $\mathcal{N}$ is spanned by the columns of $\boldsymbol{S}$, i.e., $[\mathbf{1}, \boldsymbol{x}, \boldsymbol{x}^2]$.
- The shrinkage is adaptive, i.e., a strong signal is left untouched.
- To ensure high flexibility the prior allows for a large number of basis functions but prevents highly oscillating and overfitting estimation.

We achieve these objectives by reusing the half Cauchy distribution (denoted C_+) from the horseshoe prior (Carvalho *et al.*, 2010) for the scale parameters to compile the following hierarchy:

$$\begin{aligned}
&\boldsymbol{\beta}|\lambda, \tau^2, \sigma^2 \sim \mathrm{N}_{\mathrm{prec}}(\mathbf{0}, \boldsymbol{Q}) && \text{with } \boldsymbol{Q} = \sigma^{-2}\lambda^{-2}\boldsymbol{Z}'\boldsymbol{P}_1\boldsymbol{Z} + \tau^{-2}\boldsymbol{K} \\
&\lambda|\xi \sim \mathrm{C}_+(0, \xi) && \xi \sim \mathrm{C}_+(0, \xi_0) \\
&\tau|\nu \sim \mathrm{C}_+(0, \nu) && \text{with } \nu \text{ s.t. } \lim_{\lambda \to \infty} \Pr\left(\max_{x \in \mathcal{D}} |f''(x)| < c \,\middle|\, \lambda \right) = 1 - \alpha
\end{aligned}$$

where $\boldsymbol{P}_1$ is the projection matrix with kernel equal to $\mathcal{N}$ and $\boldsymbol{K}$ is an appropriate penalty matrix based on a second-order random walk (for details visit Lang and Brezger, 2004). Furthermore, the probability statement in the hierarchy is used to determine the prior on the variance of the random walk. Over the domain of the covariates, the second order derivative should be, in absolute terms, smaller than the prespecified threshold c with probability $1 - \alpha$.
$\mathrm{N}_{\mathrm{prec}}(\mathbf{0}, \boldsymbol{Q})$ denotes a potentially improper multivariate Normal distribution with mean $\mathbf{0}$ and precision matrix $\boldsymbol{Q}$. Within the precision matrix $\boldsymbol{Q}$, the first summand ensures that small effects are shrunk towards the null

space while the second term regulates the wiggliness of the estimate, which is especially important for otherwise unpenalized estimates.
In the following, we assume that the hyper-parameter ν or its identifying factors c and α are chosen such that they give the estimated function enough flexibility to imitate the parametric estimate within the null space, and thus do not affect the shrinkage properties. This gets reflected in taking τ^2 to the infinite limit when investigating the shrinkage properties.
Even though the prior is not proper, one can show that the posterior is proper and that the estimated functional effect is, in expectation, a weighted average between the spline solution and the parametric solution with weight or shrinkage parameter $\kappa = (1 + \lambda^2)^{-1}$

$$\lim_{\tau^2 \to \infty} E(\boldsymbol{Z}\boldsymbol{\beta}|\boldsymbol{y}, \lambda, \tau^2) = ((1 - \kappa)\boldsymbol{P}_{\boldsymbol{Z}} + \kappa \boldsymbol{P}_0)\boldsymbol{y}$$

where $\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ and $\boldsymbol{P}_0 = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{S})^{-1}\boldsymbol{S}'$.
As the literature suggests (Klein *et al.*, 2019), we study the marginal prior density of the spline coefficients. We find an infinite peak for $\boldsymbol{\beta}$ such that $\boldsymbol{Z}\boldsymbol{\beta}$ is in the null space and we follow from a numerical approximation that the prior has heavy tails. Both properties are in particular beneficial for a shrinkage prior as they imply strong shrinkage for effects close to the null space and Bayesian robustness for strong signals.
To allow for more user-control over the shrinkage, the parameter ξ can be fixed. Furthermore, the prior can easily be extended to cover multiple additive functional effects with the common global shrinkage parameter ξ.

3 Simulation study

We assess the validity of our approach in a simulation study. For that, we generate 100 observations of one covariate x_i equally spaced in the interval $[-2\pi, 2\pi]$ and the response given by

$$y_i = (1 + 10\sin(x_i) + x + .64x^2)/20 + \varepsilon_i$$

with independent normal errors ε_i with variance σ^2. To feature different signal-to-noise ratios (SNRs), we repeat the study with σ set to 0.75 and 2.5. We fit models comprising the proposed prior with null spaces spanned by $[\mathbf{1}, \boldsymbol{x}]$ and $[\mathbf{1}, \boldsymbol{x}, \boldsymbol{x}^2]$, $[\mathbf{1}, \sin(\boldsymbol{x})]$ and $[\mathbf{1}, \boldsymbol{x}, \boldsymbol{x}^2, \sin(\boldsymbol{x})]$. The operations on the covariates are defined element-wise and the last null space is referred to as complex in the following. Each scenario is replicated 100 times and the main results are summarized in Figure 2.
From the histogram of the shrinkage parameter, we can deduce that the prior is mostly able to decide between signal and no signal as most values of κ are either close to 0 or 1. Furthermore, we can see that in the high noise scenario the prior forces the estimate to be on average very close to the parametric solution and thus to be in or close to the null space.

This is especially true in both SNR scenarios for the *complex* null space which includes the data generating function. In the low-noise scenario, the estimate is mainly able to capture the true function with some difficulties with the *quadratic* null space.

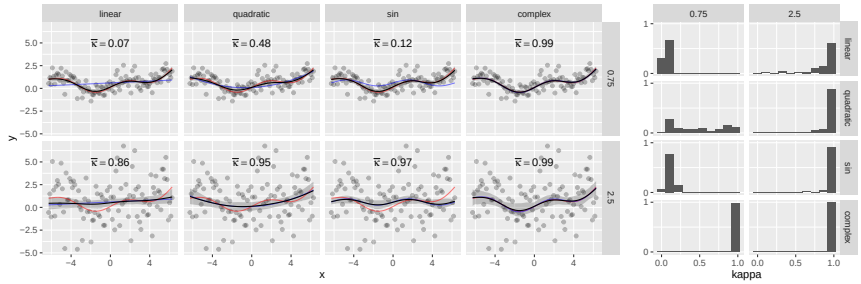


FIGURE 2. Summarized results of the simulation study. The left plot shows the data generating function (red line), data of one iteration (gray points), the mean of the parametric estimates (blue line), and the mean of the posterior mean of the estimated effect (black line) together with its 90% point-wise quantiles. Histograms of the shrinkage parameter κ are displayed on the right. Both plots breakdown the different variances and various null spaces.

References

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). Functional Horseshoe Priors for Subspace Shrinkage. *Biometrika*, **97(2)**, 465 – 480.

Klein, N., Carlan, M., Kneib, T., Lang, S., and Wagner, H. (2019). Bayesian Effect Selection in Structured Additive Distributional Regression Models. *arXiv:1902.10446 [stat.ME]*.

Lang, S., and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, **13(1)**, 183 – 212.

Shin, M., Bhattachrya, A., and Johnson, V. E. (2019). Functional Horseshoe Priors for Subspace Shrinkage. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2019.1654875.

Part II

Statistical Modelling of Habitat Selection

Shaykhah Aldossari[1], Jason Matthiopoulos[2] and Dirk Husmeier[1]

[1] School of Mathematics & Statistics, University of Glasgow, Scotland, UK
[2] Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Scotland, UK

E-mail for correspondence: s.aldossari.1@research.gla.ac.uk

Abstract: To understand the impact of habitat destruction or modification on biodiversity there is increasing demand on predictive models that reliably forecast future changes in species distributions. In the present paper, we build on an existing model, the Generalized Functional Response, whose predictions about habitat preferences and species distribution are robust to changes in habitat availability. We improve upon this model in two distinct ways by using Gaussian mixtures to approximate habitat availability and Gaussian basis functions to describe habitat preferences. The proposed model is found to improve descriptive and predictive performance when applied to realistic simulated data and real species abundance data.

Keywords: Biodiversity, habitat selection function, basis functions, Gaussian mixture model

1 Introduction

The need to understand the ecological impact of land management, building construction and urban expansion on biodiversity is driving demand for new statistical models that can reliably forecast future changes in animal population distribution. Conventional approaches in ecological modelling aim to draw inferences about the importance and direction of the relationship between habitat preference $h(\mathbf{x})$ and environmental covariates $\mathbf{x} = (x_1, \ldots, x_I)$:

$$h(\mathbf{x}) \;=\; exp\left(\sum_{i=1}^{I} \beta_i x_i\right) \tag{1}$$

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

for fixed coefficients $\beta_i \in \mathbb{R}$. This can works well if the habitat availability does not change. However, Matthiopoulos et al. (2011) discuss the limitations of this approach, and argue that it is essential to model the change in animal's habitat selection as well, by allowing the coefficients to vary as functions of habitat availability. Their derivation leads to the following expression for the habitat selection coefficients:

$$\beta_i \;=\; \mathbb{E}\Big(\gamma_i(\mathbf{x})\Big) + \varepsilon_i \;=\; \int \gamma_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} + \varepsilon_i \tag{2}$$

where $f(\mathbf{x})$ is a probability density function for habitat availability where each point $\mathbf{x}$ in environmental space represents a habitat, ε_i represents measurement or observation noise, and $\gamma_i(\mathbf{x})$ is a polynomial function in environmental covariate $\mathbf{x}$ which describes how the selection coefficient β adapts to changes in habitat availability $f(\mathbf{x})$ (introducing an integer order parameter M_j):

$$\gamma_i(\mathbf{x}) \;=\; \sum_{j=1}^{I} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} x_j^m \tag{3}$$

2 Methological Innovation

Matthiopoulos et al. (2011) demonstrate that modelling habitat selection in this way leads to a significant improvement in models of species distributions over the conventional model based on (1). However, the model still suffers from the following limitations: (1) The degree of nonlinear complexity and smoothness is restricted in advance: the functions have only M_j non-zero derivatives. (2) A complex function with a high degree of nontrivial differentiability requires a large number of parameters. (3) While the degree of smoothness is allowed to vary with respect to the choice of environmental variable, it is assumed to be global with respect to its entire range. (4) The expectation value in (2) is approximated by an empirical observed frequency, as the habitat availability is not explicitly modelled. The objective of the present paper is to propose a new statistical model that addresses these limitations. We start by replacing the polynomial with the following basis function approach:

$$\gamma_i(\mathbf{x}) \;=\; \sum_{j} \sum_{m=0}^{M_j} \delta_{i,j}^{(m)} \phi(x_j, \boldsymbol{\theta}_{j,m}) \;=\; \sum_{j} \sum_{m=0}^{M} \delta_{i,j}^{(m)} \phi(x_j, \boldsymbol{\theta}_{j,m}) \tag{4}$$

where ϕ is a basis function (e.g. splines, wavelets, basis functions of a reproducing kernel Hilbert space etc.) with parameters $\boldsymbol{\theta}_{j,m}$, chosen to represent known functional characteristics. Note that on the right-hand side we have simplified the notation by defining $M = \max\{M_j\}$, given that we have the freedom to set $\delta_{i,j} = 0$. Next, we follow Matthiopoulos

et al. (2015) and model the probability distribution $f(\mathbf{x})$ with a Gaussian mixture model:

$$f(\mathbf{x}) \;=\; \sum_k \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) \tag{5}$$

Inserting this into (2) and making use of (4) gives:

$$\begin{aligned}
\beta_i &= \gamma_{i,0} + \int \gamma_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} + \varepsilon_i \\
&= \gamma_{i,0} + \int \left[\sum_j \sum_m \delta_{i,j}^{(m)} \phi(x_j, \boldsymbol{\theta}_{j,m}) \right] \left[\sum_k \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) \right] d\mathbf{x} + \varepsilon_i \\
&= \gamma_{i,0} + \sum_j \sum_m \sum_k \delta_{i,j}^{(m)} \pi_k \left[\int \phi(x_j, \boldsymbol{\theta}_{j,m}) N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) d\mathbf{x} \right] + \varepsilon_i
\end{aligned} \tag{6}$$

If we choose an RBF basis function for $\phi(x_j, \boldsymbol{\theta}_{j,m})$:

$$\phi(x_j, \boldsymbol{\theta}_{j,m}) \;=\; \exp\left(-\frac{1}{2} \frac{(x_j - \xi_{j,m})^2}{\sigma_{j,m}^2} \right) \tag{7}$$

with parameter vector $\boldsymbol{\theta}_{j,m} = (\xi_{j,m}, \sigma_{j,m})$, then the integral

$$\psi(\boldsymbol{\theta}_{j,m}, \boldsymbol{\mu}_k, \mathbf{C}_k) \;=\; \int \phi(x_j, \boldsymbol{\theta}_{j,m}) N(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{C}_k) d\mathbf{x} \tag{8}$$

has a closed-form solution (see e.g. Bishop, Section 2.3) and we get:

$$\beta_i \;=\; \gamma_{i,0} + \sum_j \sum_m \sum_k \delta_{i,j}^{(m)} \pi_k \psi(\boldsymbol{\theta}_{j,m}, \boldsymbol{\mu}_k, \mathbf{C}_k) + \varepsilon_i \tag{9}$$

TABLE 1. Comparison of the proposed method with the approach of Matthiopoulos et al. (2011) on simulated habitat data. Smaller values indicate better performance.

Method	AIC	BIC	RMSE
Matthiopoulos et al. (2011)	907217.4	907860.4	3.97
Method proposed here	907206	907818.2	3.94

3 Empirical Evaluation

We evaluate the performance of the proposed model on the simulated data described in Matthiopoulos et al. (2011). The simulation was an individual-based model of the dependence of species abundance on two habitat variables: food and cover (the converse of predation risk).

TABLE 2. Comparison of the proposed method with the approach of Matthiopoulos et al. (2011) on the sparrow population data from Matthiopoulos et al. (2018).

Method	AIC	BIC	RMSE
Matthiopoulos (2011)	1696.024	1902.209	23.52
Method proposed here	1683.656	1889.841	11.44

We set the polynomial order for the model in Matthiopoulos et al. (2011) and the number of basis functions in the model proposed here equal to 10 based on model selection scores. We evaluate the predictive performance in terms of root mean square error (RMSE) on out-of sample test data that have not been used for parameter estimation, and compare the accuracy of the model proposed in Matthiopoulos et al. (2011) with our proposed model. The results are shown in Table 1 and suggest that a noticable improvement in terms of model selection scores (AIC, BIC) and out-of-sample (RMSE) over the state-of-the-art Generalized Function Response (GFR) model can be achieved.

4 Real-World Application

We have applied our model to the sparrow population data described in Matthiopoulos et al. (2018). This habitat use model consists of three habitat variables (the percentage of grass, bush and roof in each cell). The best polynomial order for the model in Matthiopoulos et al. (2011) and the number of basis functions in the model proposed here is 3 based on the model selection scores. The results of model selection scores and the evaluation of predictive performance on out-of-sample test data are shown in Table 2. Our model outperforms the GFR in Matthiopoulos et al. (2011), with an improvement of the model selection scores and out-of-sample RMSE.

5 Conclusions

We have modelled habitat preference with a flexible approach that extends the model proposed in Matthiopoulos et al. (2011) in two distinct ways, by using Gaussian mixtures to approximate habitat availability and Gaussian basis functions to describe habitat preferences. We have tested the new model on both simulated data and real survey data, using the sparrow population data from Matthiopoulos et al. (2018). Our results suggest that a noticeable improvement can be obtained in terms of AIC, BIC and RMSE.

References

Bishop (2006). Pattern Recognition and Machine Learning. *Springer*.

Matthiopoulos, Hebblewhite, Aarts and Fieberg (2011). Generalized functional responses for species distributions. *Ecology*, **92**, 583–589.

Matthiopoulos et al. (2015). Establishing the link between habitat selection and animal population dynamics *Ecol. Monographs*, **85**, 413–436.

Matthiopoulos, Field and MacLeod (2018). Predicting population change from models based on habitat availability and utilization. *Proceedings of the Royal Society B*, **286**, 20182911.

Stochastic Profiling of mRNA Counts Using HMC

Lisa Amrhein[1,2], Christiane Fuchs[1,2,3]

[1] Helmholtz Zentrum München, Germany
[2] Technical University Munich, Germany
[3] Bielefeld University, Germany

E-mail for correspondence: christiane.fuchs@helmholtz-muenchen.de

Abstract: We dissect transcriptional heterogenity from RNA sequencing counts taken from small pools of cells when single-cell data is disadvantageous. For this purpose, we extend the stochastic profiling algorithm (Amrhein & Fuchs 2020) to discrete data. In addition, we perform Bayesian inference using Hamiltonian Monte Carlo. Our implementation uses Stan to optimize computational efficiency.

Keywords: Stochastic profiling, HMC, Stan, Heterogeneity, Deconvolution.

1 Stochastic Profiling

Gene expression of cells is determined by their amount of mRNA. With the rapid development of technologies, the widely used microarray measurements (determining relative, *continuous* amounts of mRNA, Kurimoto 2006) are being replaced by sequencing methods (counting *discrete* numbers of mRNA molecules, Sandberg 2014). While bulk measurements assess the overall gene expression of millions of cells, single-cell measurements inform on a much finer resolution. Since gene expression is not only heterogeneous between individuals and cell types, but also within tissues, single-cell data appears to be best-suited to fully identify heterogeneity. However, single-cell data is more cost-intensive and prone to technical noise than measurements of pools. The joint measurement of small pools of cells is a suitable trade-off between the bulk and single-cell approach. Stochastic profiling deconvolutes joint measurements of small pools of cells into parametric mixtures of single-cell distributions. Tailored to continuous microarray data, Bajikar et al. (2014) develop and apply the method using lognormal and exponential distribution models. Amrhein & Fuchs

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

(2020) describe modelling, inference and R implementation in detail. Recent technological advances make small-pool sequencing possible, resulting in discrete small-pool mRNA counts (Singh et al., 2019). It is necessary to develop the stochastic profiling algorithm further to apply it to novel data.

2 Statistical Convolution Model

We aim to deconvolute gene expression measurements $Y_1, \ldots, Y_K$ of K pools of n cells with independent latent single-cell expression $X_{i1}, \ldots, X_{in}$ such that $Y_i = \sum_{j=1}^{n} X_{ij}$. This corresponds to the model in Amrhein & Fuchs (2020) but with discrete single-cell distributions. We employ a common choice for single-cell mRNA counts, the negative binomial (NB) distribution (Kharchenko et al. 2014; Amrhein et al. 2019). The probability mass function (PMF) of $X \sim \mathrm{NB}(\alpha, p)$ with $\alpha \in \mathbb{R}_+$ and $p \in [0, 1]$ is

$$f_{\mathrm{NB}}(x; \alpha, p) = \frac{\Gamma(\alpha + x)}{\Gamma(\alpha) x!} p^{\alpha} (1 - p)^{x} \quad \text{for } x \in \mathbb{N}_0. \tag{1}$$

Assume that each single cell stems from one of T populations with probabilities $\theta_1, \ldots, \theta_T$ (which sum up to one), and within each population j the gene expression is modelled by $\mathrm{NB}(\alpha_j, p_j)$. Then, the PMF $f_n(y; \vec{\theta}, \vec{\alpha}, \vec{p})$ of the cumulative measurement $Y = y$ of a pool of n cells reads

$$\sum_{\ell_1=0}^{n} \sum_{\ell_2=0}^{n-\ell_1} \cdots \sum_{\ell_{T-1}=0}^{n-\ell_1-\cdots-\ell_{T-2}} \binom{n}{\ell_1, \ldots, \ell_T} \theta_1^{\ell_1} \ldots \theta_T^{\ell_T} f_{(\ell_1,\ldots,\ell_T)}(y; \vec{\alpha}, \vec{p}), \tag{2}$$

where $f_{(\ell_1,\ldots,\ell_T)}$ is the PMF of a deterministic mixture of $0 \leq \ell_j \leq n$ cells of type j for each $j = 1, \ldots, T$ and $\ell_T = n - \ell_1 - \ldots - \ell_{T-1}$ (Amrhein & Fuchs 2020). Hence, $f_{(\ell_1,\ldots,\ell_T)}$ describes the convolution $Y^* = X_1^* + \ldots + X_n^*$ of n independent random variables $X_i^* \sim \mathrm{NB}(\alpha_i, p_i)$ with PMF (Furman, 2007)

$$f_{n\text{-NB}}(y^*; \vec{\alpha}, \vec{p}) = R \sum_{k=0}^{\infty} \delta_k f_{\mathrm{NB}}(y^*; \alpha + k, p_{\max}) \quad \text{for } y^* \in \mathbb{N}_0, \tag{3}$$

where $\alpha = \alpha_1 + \cdots + \alpha_n$, $p_{\max} = \max_j\{p_j\}$, $R = \prod_{j=1}^{n} \left(\frac{(1-p_j)p_{\max}}{(1-p_{\max})p_j} \right)^{\alpha_j}$ and $\delta_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} \sum_{j=1}^{n} \alpha_j \left(1 - \frac{(1-p_{\max})p_j}{(1-p_j)p_{\max}}\right)^{i} \delta_{k+1-i}$. We cut the infinite sum in (3) where the following summand equals zero. Computation of the PMF can be simplified further: Within each population j, the distribution parameters α_j and p_j are identical such that the sum of the ℓ_j random variables follows the $\mathrm{NB}(\ell_j \alpha_j, p_j)$ distribution. Consequently, $f_{(\ell_1,\ldots,\ell_T)}$ is the convolution of at maximum T NBs (exactly T-fold if all $\ell_j > 0$).

3 Implementation of Bayesian parameter estimation

We infer the parameters $\vec{\theta}$, $\vec{\alpha}$ and $\vec{p}$ from n-cell measurements using the Hamiltonian Monte Carlo (HMC)-based No-U-Turn sampler (NUTS, Hoffman & Gelman 2014), implemented in the programming language Stan through its interface RStan (Stan Development Team 2019). All code is provided at `https://github.com/fuchslab/mcmcNB_Stan`. In contrast to the original HMC, NUTS does not require the specification of the number of steps L. In addition, the RStan implementation tunes the step size ϵ in an automated manner. There remained the implementation of our model-specific likelihood function based on Equation (2).

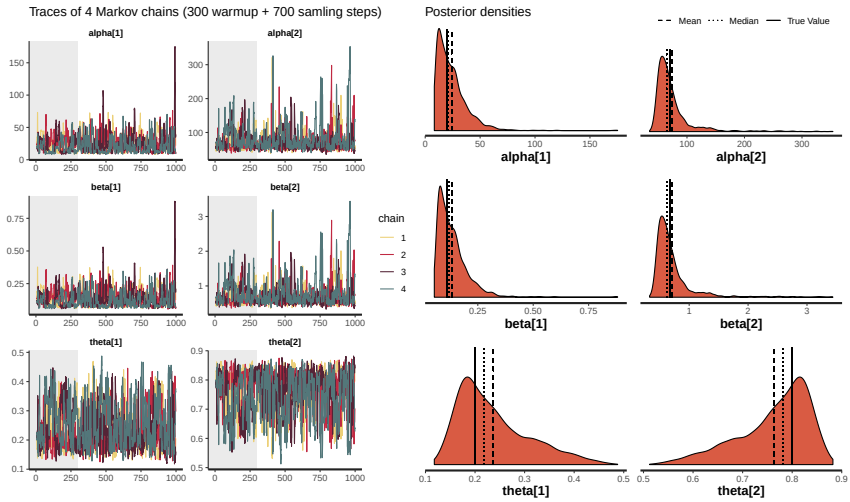


FIGURE 1. Parameter traces and densities of the posteriors of the NUTS chains. The algorithm uses parameters $\beta = p/(1-p)$.

NUTS requires calculating the gradient of the log-posterior density. For this purpose, Stan uses auto-differentiation and creates a so-called expression tree to evaluate all required gradients of the likelihood. As described above, Equation (3) contains an infinite sum that must be approximated. Since the expression graph is built only once at the beginning, the number of summands cannot vary for each iteration; this would affect the size of the expression tree. One way out is to always approximate the sum by a constant very high number of summands, e. g. 10,000. An alternative solution is to implement different versions with different constant numbers of summands (e. g. 1, 5, 10, 50, 100, 500, 1,000, 5,000 and 10,000). Then one expression tree can be built with several subtrees (in this case: nine), and in each iteration it is checked whether more summands are needed and thus which subtree to use. Figure 1 shows results from this alternative version 2.

4 Results and Conclusion

We apply the algorithm to a synthetic dataset with 1,000 2-cell samples of two populations and frequencies $\vec{\theta} = (20\%, 80\%)$ and negative binomial parameters $\vec{\alpha} = (20, 70)$ and $\vec{p} = (0.1, 0.4)$. Figure 1 indicates that our algorithm is able to capture the true parameter values. However, more excessive confirmatory simulation studies are required. These are time-consuming as the expression tree is huge and thus each gradient calculation extensive. Modifying the implementation is ongoing work with the aim to approximate formula (3) more efficiently and decrease evaluation time.

Acknowledgments: We thank Mara Santarelli and Susanne Pieschner for Stan support. We were funded by the SFB 1243 (A17), the Helmholtz Association (Uncertainty Quantification) and the NIH (U01-CA215794).

References

Amrhein, L. and Fuchs, C. (2020). stochprofML: Stochastic Profiling Using Maximum Likelihood Estimation in R. *arXiv* 2004.08809.

Amrhein, L., Harsha, K., and Fuchs, C. (2019). A mechanistic model for the negative binomial distribution of single-cell mRNA counts. *BioRxiv*, 657619.

Bajikar, S.S., Fuchs, C., Roller, A., Theis, F.J. and Janes, K.,A. (2014). Parameterizing Cell-to-Cell Regulatory Heterogeneities via Stochastic Transcriptional Profiles. *PNAS*, **111(5)**, 626 – 635.

Furman, E. (2007). On the convolution of the negative binomial random variables. *Statistics & Probability Letters*, **77 (12)**, 169 – 172.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**, 1593 – 1623.

Kharchenko, P. V., Silberstein, L. and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat Methods*, **11**, 740 - 742.

Singh, S., Wang, L., Schaff,D. L., Sutcliffe, M. D., Koeppel, A. F., Kim, J., Onengut-Gumuscu, S., Park, K., Zong, H. and Janes, K. A. (2019). In situ 10-cell RNA sequencing in tissue and tumor biopsy samples. *Sci Rep*, **9**, 4836.

Stan Development Team (2019). RStan: the R interface to Stan. *R package version 2.19.1, http://mc-stan.org/*

A Bayesian naïve Bayes classifier for dating archaeological sites

Carmen Armero[1], Gonzalo García-Donato[2], Joaquín Jiménez-Puerto[1], Salvador Pardo-Gordó[1], Joan Bernabeu[1]

[1] Universitat de València, Spain
[2] Universidad de Castilla-La Mancha, Spain

E-mail for correspondence: carmen.armero@uv.es

Abstract: Dating is a key element for archaeologists. We propose a Bayesian approach to provide chronology to sites that have neither radiocarbon dating nor clear stratigraphy and whose only information comes from bifacial flint arrowheads. This classifier is based on the Dirichlet-multinomial inferential process and posterior predictive distributions. The procedure is applied to predict the period of a set of undated sites located in the east of the Iberian Peninsula during the IVth and IIIrd millennium cal. BC

Keywords: Bifacial flint arrowheads; Dirichlet-multinomial process; Posterior predictive distribution

1 Introduction

Dating is a key element for archaeologists because they need a time scale to locate the information collected from the excavations and field work in order to build, albeit with uncertainty, our most remote past. Archaeological scientists generally use stratigraphic expert information and dating techniques for examining the age of the relevant artifacts. Bayesian inference is commonly used in archaeology as a tool to construct robust chronological models based on information from scientific data as well as expert knowledge (e.g. stratigraphy) (Buck *et al.*, 1996).
Radiocarbon dating is one of the most popular techniques for obtaining data due to its presence in any being that has lived on Earth. However, it is not always possible in all studies to collect organic material and obtain that type of data or to have good stratigraphic references. In this context, we propose a Bayesian approach to provide chronology to some archaeological

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

sites that do not have radiocarbon dates and show unprecise stratigraphic relationships.

We propose an automatic Bayesian procedure, very popular in text classification (Wang *et al.*, 2003), based on predictive probability distributions, for classifying the period to which an undated site belongs based on the type and number of arrows found in it. This proposal takes into account on the Dirichlet-multinomial inferential process for learning about the proportion of different types of arrowheads in each chronological period and the concept of posterior predictive distribution for a new undated site. This procedure is applied to date a set of sites located in the east of the Iberian Peninsula during the IVth and IIIrd millennium cal. BC. During this time, bifacial flint arrowheads appear and spread. Archaeological research suggests that the shape of these arrowheads could be related with specific period and/or geographical social units spatially defined.

2 Bayes classifier

The prediction of the period to which an undated site belongs based on information about the number and type of arrows that have been collected at this site includes two different phases.

2.1 Dirichlet-multinomial inferential process

Let Y_{ij} be the random variable that describes the number of type j arrowheads, of the total n_i collected, in the sites belonging to period i, $i = 1, \ldots, I$, and consider $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$.

A probabilistic model for $\boldsymbol{Y}_i \mid \boldsymbol{\theta}_i$ is the multinomial distribution, $\text{Mn}(\boldsymbol{\theta}_i, n_i)$, where $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iJ})'$ is a probability vector and θ_{ij} is the probability that an arrowheads of period i is of type j.

We assume a Perks prior distribution (Armero *et al.*, 2018) for $\boldsymbol{\theta}_i$. The subsequent posterior posterior distribution is the Dirichlet (Dir) distribution

$$\pi(\boldsymbol{\theta}_i \mid \mathcal{D}_i) = \text{Dir}(\alpha_{i1} = y_{i1} + (1/J), \ldots, \alpha_{iJ} = y_{iJ} + (1/J))$$

where y_{ij} is number of arrowheads of type j in the period i and $\mathcal{D}_i = \{y_{i1}, \ldots, y_{iJ}\}$. The marginal posterior distribution for each probability θ_{ij} ia beta distribution $\text{Be}(\alpha_{ij}, \alpha_{i+} - \alpha_{ij})$, with $\alpha_{i+} = \sum_{j=1}^{J} \alpha_{ij}$.

2.2 Classification process

After learning about the distribution of the number of arrowheads types in each site, we have to assign a period m^* to a new site with a given number and type of arrowsheads recorded. We consider a new undated site s^* in which we found a total of n^* arrowheads distributed by type

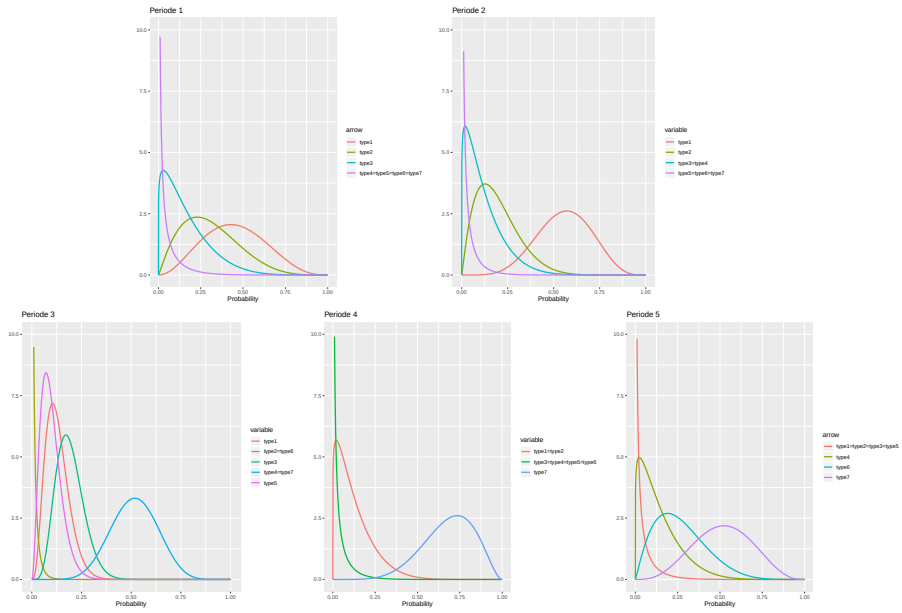
according to $\boldsymbol{y}^* = (y_1^*, \ldots, y_J^*)$. The relevant scientific question is now about the probability that this site belongs to each of the different time periods considered. Following Bayes' theorem:

$$P(m^* = m_i \mid \boldsymbol{y}^*, \mathcal{D}) \propto P(\boldsymbol{y}^* \mid m^* = m_i, \mathcal{D})\, P(m^* = m_i \mid \mathcal{D})$$

where $\mathcal{D} = \cup \mathcal{D}_i$, $(\boldsymbol{y}^* \mid m^* = m_i, \mathcal{D})$ follows a Dirichlet-multinomial distribution $\mathrm{DiMn}(n^*, \boldsymbol{\alpha}_i)$ with $n^* = \sum y_j^*$, and $P(m^* = m_i \mid \mathcal{D})$ can be estimated as the proportion of sites in the sample for each of the periods under consideration.

3 East of the Iberian Peninsula sites during the IVth and IIIrd millennium cal. BC.

Five chronological periods in the east of the Iberian Peninsula sites during the IVth and IIIrd millennium cal. BC. were studied. They include arrowheads data from several archaeological contexts, *Niuet*, *Jovades 1* and *Jovades 2* from period 1, *Quintaret*, *Jovades 3*, *Jovades 4*, and *Niuet 2* from period 2, *Migdia 1*, *Beniteixir* , *La Vital 1* , *Randero 1*, *Niuet 3*, *Niuet 4*, and *Diablets* from period 3, *Migdia 2*, *Missena 1*, and *La Vital 2* from period 4, and *Arenal costa*, *Missena 2*, and *La Vital 3* from period 5.



The Figure above shows the posterior marginal distribution of the abundance of the different types of arrowheads in each of the five chronological

periods considered. Type 1 and 2 arrowheads are most abundant in periods 1 and 2, with an increase in type 1 compared to type 2 arrowheads in the second period. During period 3, type 4 and type 7 arrowheads are more abundant. The latter are clearly the most used in period 4, which become less used in period 5 when type 6 arrowheads appears with more probability.

The posterior probability that a new site belongs to each of the periods considered was estimated as 0.15 for periods 1, 4 and 5, 0.20 for period 2, and 0.35 for period 3.

The following table presents the posterior predictive distribution of the period to which a series of new undated sites belong, whose only available information is based on the number and type of arrows found collected.

Site	Period 1	Period 2	Period 3	Period 4	Period 5
Rambla C.	0.0001	0.0000	0.0004	0.9339	0.0654
Ereta I	0.7804	0.2196	0.0000	0.0000	0.0000
Ereta II	0.5019	0.4901	0.0000	0.0075	0.0005
Ereta III	0.0694	0.0912	0.8330	0.0060	0.0004
Ereta IV	0.0021	0.0098	0.6358	0.3504	0.0019

The results obtained present a great agreement with the expert information of the archaeologists of the project, so it is a proposal that can be very useful in archaeological research.

References

Alvares, D., Armero, C., and Forte, A. (2018). What Does Objective Mean in a Dirichlet Multinomial Process? *International Statistical Review*, **86**, 106 – 118.

Buck, I. C. E., Cavanagh, W. G., and Litton, C. D. (1996). *Bayesian Approach to Interpreting Archaeological Data.* Chiscester: Wiley.

Wang, Y., Hodges, J. and Tang, B. (2003). Classification of Web Documents Using a Naive Bayes Method. *15th IEEE International Conference on Tools with Artificial Intelligence*, **124**, 560 – 564.

A Sensitivity Analysis and Error Bounds for the Adaptive Lasso

Tathagata Basu[1], Jochen Einbeck[1], Matthias C. M. Troffaes[1]

[1] Durham University, United Kingdom

E-mail for correspondence: tathagata.basu@durham.ac.uk

Abstract: Sparse regression is an efficient statistical modelling technique which is of major relevance for high dimensional problems. There are several ways of achieving sparse regression, the well-known lasso being one of them. However, lasso variable selection may not be consistent in selecting the true sparse model. Zou (2006) proposed an adaptive form of the lasso which overcomes this issue, and showed that data driven weights on the penalty term will result in a consistent variable selection procedure. Weights can be informed by a prior execution of least squares or ridge regression. Using a power parameter on the weights, we carry out a sensitivity analysis for this parameter, and derive novel error bounds for the Adaptive lasso.

Keywords: Adaptive lasso; Sensitivity analysis; Variable selection.

1 Introduction

Let $\mathbf{X} = (X_1, \cdots, X_p)$ with $X_j = (X_{1j}, \dots X_{nj})^{\mathrm{T}}$ for $1 \leq j \leq p$, and $Y = (Y_1, \dots, Y_n)^{\mathrm{T}}$. We can characterise their relation in the linear regression setting

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ is a vector of regression coefficients and $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, with $\mathbf{I}_n$ denoting the n-dimensional identity matrix. We assume $\mathbf{X}$ and Y to be scaled to mean 0.

The least squares method is the conventional way to estimate these regression coefficients. However, in high dimension (i.e $p > n$), the least squares method, which involves inversion of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, cannot be used. Several estimators have been proposed which solve the issue by introducing bias in the estimation process. Tikhonov (1963) introduced ℓ_2 penalised regression or Ridge regression. The ℓ_2 penalty achieves a stable solution through the

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

eigen value decay method, which, however, fails to be sparse which is a desirable property in high dimensional statistics. Tibshirani (1996) introduced the lasso or least absolute shrinkage and selection operator, which attains sparsity through a ℓ_1 penalty. Zou (2006) proposed an adaptive form of lasso based on data-driven weights in the penalty term that satisfies desired asymptotic properties for high-dimensional problems as suggested by Fan and Li (2001). We exploit the framework given by Zou (2006) to investigate and understand the sensitivity of the adaptive lasso. For this we apply a two-step approach. We employ least squares or ridge estimates, say $\hat{\beta}_j$, and a parameter γ to initialise the weights of type $1/|\hat{\beta}_j|^\gamma$ which are then embedded in the penalty term. The effect of the parameter γ is then investigated, theoretically, through error bounds, and experimentally, through a sensitivity analysis.

2 Adaptive Lasso

Let us consider the linear model (1) which can be written in alternative form as

$$\mathbb{E}[Y \mid \mathbf{X}] = \mathbf{X}\boldsymbol{\beta} = \beta_1 X_1 + \cdots + \beta_p X_p. \tag{2}$$

Note that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is guaranteed to be positive semi-definite but not necessarily positive definite, even for $p < n$. We make the following two assumptions on the design $\mathbf{X}$:

(A1) $\mathbb{E}[\mathbf{X}^{\mathrm{T}}\epsilon \mid \mathbf{X}] = 0$

(A2) $\lim_{n\to\infty} \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \Sigma$ exists, where Σ is positive definite.

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \cdots, \hat{\boldsymbol{\beta}}_p)^{\mathrm{T}}$ be any root-n consistent estimator of $\boldsymbol{\beta}$. Then the adaptive lasso estimates are given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{alasso}}(\lambda, \gamma) = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2}\|Y - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j(\gamma)|\beta_j| \right) \tag{3}$$

where

$$w_j(\gamma) = |\hat{\beta}_j|^{-\gamma}, \quad \text{for} \quad \gamma > 0. \tag{4}$$

We generally use least squares estimates or ridge estimates as weights since these are root-n consistent.

3 Main Result

Let $\hat{\boldsymbol{\beta}}_{\mathrm{alasso}}(\lambda, \gamma)$ be the adaptive lasso estimates with respect to the parameters λ and γ and $\Sigma_n = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Let $\boldsymbol{\beta}^*$ be the true regression coefficients.

Theorem: For any root-n consistent estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)$, we have the following error bounds:

$$\left\| \hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda, \gamma) - \boldsymbol{\beta}^* \right\|_2^2 \leq \frac{\sigma^2}{n} \left\| \Sigma_n^{-1} \right\| + \frac{\lambda^2 p}{n^2} \left\| \Sigma_n^{-1} \right\|^2 \min_{1 \leq j \leq p} |\hat{\beta}_j|^{-2\gamma} \quad (5)$$

$$\left\| Y - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda, \gamma) \right\|_2^2 \leq \frac{\lambda^2 p}{n} \left\| \Sigma_n^{-1} \right\| \min_{1 \leq j \leq p} |\hat{\beta}_j|^{-2\gamma} \quad (6)$$

We see that the error bounds increase with increasing λ (increased bias from regularisation) but tend to decrease with increasing γ.

4 Simulation Study

We simulate the predictors from a standard normal distribution such that, $X_{ij} \sim N(0, 1)$ for $j = 1, \cdots, 20$ and $i = 1, \cdots, n$. We assign the regression coefficients to be $(\beta_1, \cdots, \beta_6) = (5, 3, 1, -1, -3, -5)$ and $\beta_j = 0$ for $j > 6$. We consider standard normal noise to construct the response vector $y_i = \sum_{j=1}^{6} X_{ij}\beta_j + \epsilon_i$ where, $\epsilon_i \sim N(0, 1)$ for $i = 1, \cdots, n$. The experiment is repeated for $n = 100, 500, 1000$.

We analyse the sensitivity of the model for $0 \leq \gamma \leq 1$ ($\gamma = 0$ yields regular lasso estimates). We use least squares estimates for the choice of weights. In Table 1, we compare prediction accuracy of different lasso variants, and also display the number of active co-variates, p^*. In the first row we give the results of the adaptive lasso for $\gamma = 1$. We specify λ through cross-validation. In the next three rows we show results for varying γ and fixed λ. In Figure 1, we show the coefficient path and RMSE curve evaluated over γ for 100 observations. From Figure 1 we see that as the value of γ increases, the bias and RMSE decrease which is plausible in the light of Theorem 1. However, we also notice that it overfits and selects six extra variables as important. In the last row we show results from the lasso.

5 Conclusion

We have presented a sensitivity analysis for the adaptive lasso with respect to γ, and obtained novel bounds for the lasso estimates. We have shown through simulation that the bias due to regularisation with λ can be reduced for larger values of γ, however, especially for small sample sizes, at the potential expense of overfitting and selection of some non-important variables in the model.

Acknowledgments: This work is funded by the European Commissions H2020 programme, through the UTOPIAE Marie Curie Innovative Training Network, H2020-MSCA-ITN-2016, Grant Agreement number 722734.

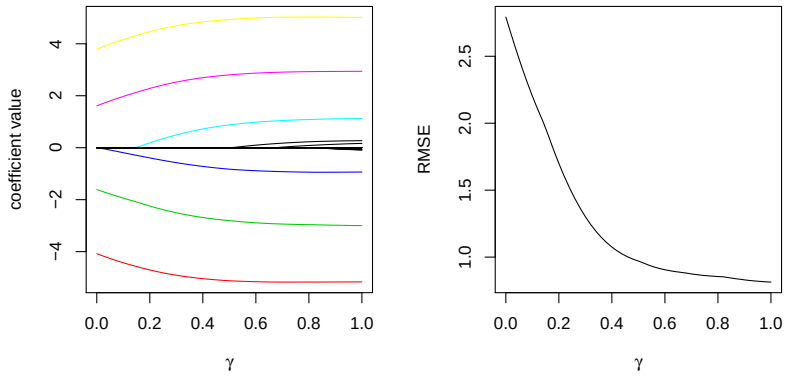


FIGURE 1. Coefficient path and fitting accuracy w.r.t. γ ($\lambda = 1$) for $n = 100$.

TABLE 1. Comparison of prediction accuracy (RMSE) between different methods.

	$n = 100$		$n = 500$		$n = 1000$	
	RMSE	p^*	RMSE	p^*	RMSE	p^*
Adaptive Lasso						
$\gamma = 1$, λ by CV	0.94	6	1.02	6	0.99	6
$\gamma = 0.1$, $\lambda = 1$	2.20	5	2.02	6	1.94	6
$\gamma = 0.5$, $\lambda = 1$	0.97	6	1.05	6	1.02	6
$\gamma = 1$, $\lambda = 1$	0.81	12	0.99	6	0.97	6
Lasso, λ by CV	0.93	10	1.03	6	1.00	6

References

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**(456), 1348 – 1360.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **58**(1), 267 – 288.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR*, **151**(3), 501 – 504.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**(476), 1418 – 1429.

Penalised Complexity priors for copula estimation

Diego Battagliese[1], Clara Grazian[2], Brunero Liseo[1], Cristiano Villa[3]

[1] MEMOTEF Department, Sapienza University of Rome, Rome, Italy
[2] University of New South Wales, Sydney, Australia
[3] SMSAS, University of Kent, Canterbury, United Kingdom

E-mail for correspondence: diego.battagliese@uniroma1.it

Abstract: We consider a multivariate model with independent marginals as a benchmark for a generic multivariate model where the marginals are not independent. The Penalised Complexity (PC) prior takes natural place in such a context, as we can include in the simpler model an extra-component taking into account for dependence. In this paper, the additional component is represented by the parameter of the Gaussian copula density function. We show that the PC prior for a generic copula parameter can be derived regardless of the parameters of the marginal densities. Then, we propose a hierarchical PC prior for the Gaussian copula model.

Keywords: PC prior; Gaussian copula; Intrinsic prior; Hierarchical PC prior.

1 Introduction

In many statistical models it is natural to have a nested structure. Consider a model of a given complexity, one way to obtain a richer and more flexible model is to include an extra-component so that the simpler model would be nested in the more complex one. We may think, for instance, of a situation where we want to model the joint distribution of several random variables through a copula function. In the case of dependence among variables, the joint density can be expressed as the product of the marginal distributions times a copula function, on the contrary, the joint density boils down to the only product of the marginals when the latter are independent. We derive the PC prior for the correlation parameter of the Gaussian copula by exploiting the following result on the Kullback-Leibler divergence (KLD).

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

Notice that the KLD is used to measure the distance between the two models. For a review of the principles behind the construction of a Penalised Complexity prior, see Simpson *et al.* (2017).

2 Method

We consider as a base model a certain multivariate density where the marginal densities are independent. Then, we could render this model more flexible by allowing a copula function to account for dependence, on the basis of the Sklar's representation. The flexible model is

$$M_1 = \{f_{\mathbf{X};\phi}(\mathbf{x};\phi), \mathbf{x} \in \mathbb{R}^k, \phi \in \mathbb{R}^q\}, \tag{1}$$

where, according to the Sklar's theorem, the joint density can be written

$$f_{\mathbf{X};\phi}(\mathbf{x};\phi) = \prod_{j=1}^{k} f_j(x_j;\underline{\theta}_j) c_\psi(F_1(x_1;\underline{\theta}_1), \ldots, F_k(x_k;\underline{\theta}_k);\psi), \tag{2}$$

and $\phi = \{\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_k, \psi\}$.
Furthermore, let

$$M_0 : \qquad \mathbf{X} \sim f_0(\mathbf{x};\varphi) = f_{\mathbf{X};\varphi}(\mathbf{x};\varphi) = \prod_{j=1}^{k} f_j(x_j;\underline{\theta}_j) \tag{3}$$

be the base model, where f_0 is the density of $\mathbf{X}$ in the case in which there is independence among the marginals, namely, when the value of ψ returns the independence copula. Here, $\varphi = \{\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_k, \psi = \psi_0\}$. Then, the theorem below follows

Theorem 1 (Invariance wrt marginals) *Let* $\mathbf{X} \sim f_{\mathbf{X}}(x_1, \ldots, x_k)$ *be a random vector with density* $f_{\mathbf{X}}$ *(we assume it is absolutely continuous with respect to the Lebesgue measure). Furthermore, let* $\mathbf{Y}$ *be a random vector with distribution* $f_{\mathbf{Y}}(y_1, \ldots, y_k) = \prod_{j=1}^{k} f_j(y_j)$ *where* f_j *is the marginal density of* X_j *and* Y_j*, then*

$$\mathrm{KLD}(f_{\mathbf{X}} \| f_{\mathbf{Y}}) = \int_{[0,1]^k} c(u_1, \ldots, u_k;\psi) \log c(u_1, \ldots, u_k;\psi) du_1 \ldots du_k, \tag{4}$$

where $c(u_1, \ldots, u_k)$ *represents the copula function associated with the density of* $\mathbf{X}$ *and* $U_j \sim \mathrm{Unif}(0,1),\ j = 1, \ldots, k$.

The theorem above states that the distance between a generic multivariate density and the one with independent marginals can be expressed as the distance between the copula density function and the independence copula. This result allows us to derive the PC prior for the copula parameter

regardless of the parameters of the marginals. Notice also that it applies to any copula function and any dimension. Nevertheless, apart from the case of equicorrelation, for multidimensional elliptical copulas we need to define a multivariate PC prior. Suppose now to have only two marginal distributions, on the basis of Theorem 1 we can write

$$\mathrm{KLD}(f_{\mathbf{X};\phi} \| f_{\mathbf{X};\varphi}) = \int_{\mathcal{U}} \int_{\mathcal{V}} c(u, v; \rho) \log c(u, v; \rho) du dv, \tag{5}$$

where c is the density function of a bivariate Gaussian copula with parameter ρ. Then, $\mathrm{KLD}(\rho) = -\frac{1}{2}\log(1-\rho^2)$, and the prior is easily obtained

$$\pi^{PC}(\rho) = \frac{\theta}{2} \exp\Big(-\theta\sqrt{-\log(1-\rho^2)}\Big) \frac{|\rho|}{(1-\rho^2)\sqrt{-\log(1-\rho^2)}}. \tag{6}$$

The latter prior is proper, clearly symmetric as it depends on ρ only through the square and the absolute value, and has any odd moment equal to zero. Simpson *et al.* (2017) proposed to use a probability statement on a tail event to select the parameter θ. This latter plays a key role as it regulates the shrinkage of the prior towards the base model, so a wrong choice of this parameter may be misleading, especially in Bayesian hypothesis testing.
From an objective point of view, we calculate the intrinsic prior for the rate parameter θ and then we specify the hyperparameter of such an intrinsic prior distribution by maximizing the variance of the hierarchical PC prior for ρ where the intrinsic prior is put on θ. The procedure to derive the intrinsic prior is borrowed from Pérez and Berger (2002) as it coincides with the expected-posterior prior.
We use $\pi^N(\theta) = \frac{1}{\theta}$ as an improper starting distribution, then the intrinsic prior is given by

$$\pi^I(\theta) = \int_{-1}^{1} \pi(\theta|\rho_\ell) f(\rho_\ell|H_0) d\rho_\ell, \tag{7}$$

where $f(\rho_\ell|\theta_0)$ is the PC prior in (6) calculated in θ_0, say the null hypothesis, and $\pi(\theta|\rho_\ell) = \frac{\pi^N(\theta) f(\rho_\ell|\theta)}{m^N(\rho_\ell)}$, where in turn ρ_ℓ represents the training sample. If there is no subset of ρ_ℓ for which $0 < m^N(\rho_\ell) < \infty$, then ρ_ℓ is called *minimal training sample*. Berger and Pericchi (1996) showed that often it will simply be a sample of size $\max(\dim(\theta))$. So, we need just an observation to convert the improper starting distribution into a proper prior. Therefore

$$\pi^I(\theta) = \frac{\theta_0}{(\theta + \theta_0)^2} \tag{8}$$

will be proper. We set the hyperparameter θ_0 in an objective manner. In particular, we numerically maximize the variance with respect to θ_0, i.e.

$$\max_{\theta_0} \int_{-1}^{1} \int_{0}^{\infty} \rho^2 \pi^{PC}(\rho|\theta) \pi^I(\theta|\theta_0) d\theta d\rho. \tag{9}$$

The maximizer is $\theta_0 = 0.491525$ as it renders the prior as flat as possible.

3 Simulation study and real data

We check out the frequentist performance of our hierarchical PC prior via a simulation study. For each true ρ^* $(-0.95, -0.5, 0, 0.05, 0.5, 0.95, 0.999)$ and for each fixed sample size $(n = 5, 30, 100, 1000)$ we have generated 200 independent samples from the Gaussian copula and for each of them we have calculated the posterior mean, the 95% credible interval and the Bayes factor. We use the Jeffreys' prior as competitor for inference, while for Bayesian hypothesis test we use the Arc-sine prior, since it is proper.
As one can expect, for $\rho = 0$, our hierarchical PC prior is superior to the Jeffreys' prior in terms of MSE; this is because of the little spike at the base model induced by the hierarchical approach. However, for intermediate correlations, there seems to be a bias-variance trade-off; the Jeffreys' prior looks less biased but less efficient, whilst the PC prior seems to be more biased but more efficient. To compare overall values of ρ^*, we also compute an overall MSE, and the latter is basically in favour of our prior.
We use Bayes factor to select among models. Theorem 1 allows us to write

$$B_{01} = \frac{c_\rho(u, v; \rho)|_{\rho=0}}{\int c_\rho(u, v; \rho)\pi^{PC}(\rho|\theta_0 = 0.491525)d\rho}. \tag{10}$$

We compute the frequency of times that $B_{01} \leq 0.5$. It turns out to be basically smaller for the PC prior compared to the Arc-sine prior when the true model is the base model, whilst it is larger when the true ρ deviates far away from the independence model, especially for small sample sizes.
Finally, we analyze the `danube` data set which contains ranks of base flow observations for two stations situated at Scharding (Austria) on the Inn river and at Nagymaros (Hungary) on the Danube. The data have been pre-processed to remove any time trend. Specifically, a linear time series model with 12 seasonal components is fitted. Then residuals are extracted. The correlation between time series is computed over the residuals, otherwise we would carry back correlation within the series. The results of the Bayesian test are in line with the ones of the frequentist test, providing strong evidence for $\rho \neq 0$.

References

Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109 – 122.

Pérez, J.M. and Berger, J.O. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, **89**, 491 – 511.

Simpson, D., Rue, H., Riebler, A., Martins, T.G. and Sørbye, S.H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with Discussion). *Statistical Science*, **32**, 1 – 28.

Practical consistent estimation of the structural parameters of true fixed-effects stochastic frontier models

Ruggero Bellio[1], Luca Grassetti[1]

[1] Dept. of Economics and Statistics – University of Udine, Italy

E-mail for correspondence: `ruggero.bellio@uniud.it`

Abstract: True fixed-effects stochastic frontier models are employed in panel data settings to separate time-invariant heterogeneity from efficiency effects. These models have some desirable properties, but the estimation of their structural parameters is hindered by the incidental parameter problem, which may be severe for settings with a large number of short panels. Some consistent estimators have been recently proposed in the econometric literature, but they are rather involved and overly complex to implement. Here we propose an alternative estimator, which has optimality properties while being computationally simple. The proposal results from the application of the equivariance property of maximum likelihood estimation in group families, and it provides a consistent estimator regardless of the size of the panel. The solution covers a broad range of stochastic terms, and it does not require any simulation. The `TMB R` package for automatic differentiation is employed to obtain a scalable implementation.

Keywords: Integrated Likelihood; Panel Data; Stochastic Frontier Models; True Fixed-Effects Approach.

1 Introduction

Stochastic frontier models are widely used for the study of economic efficiency, and the generalisation of stochastic frontier models to panel data framework had a central role in the econometric literature of the last decades. In particular, the specification of the individual effects entering the model characterised the discussion, and both random and fixed effects approaches have been largely studied. In the last 15 years the literature focused on the fixed-effect approach, mostly for reasons of specification robustness. The true fixed-effect model, first proposed by Greene (2005),

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

provides a flexible fixed-effects resolution for panel data stochastic frontiers. The model is

$$y_{it} = \alpha_i + x_{it}^{\top}\beta - u_{it} + \varepsilon_{it}\,, \tag{1}$$

where α_i are panel-specific fixed parameters and u_{it} time-varying inefficiency terms. Here $i = 1, \ldots, n$ is the index for panel and $t = 1, \ldots, T$ is the index for time. Note that in the specification (1) efficiency is separated from panel heterogeneity, and specification robustness derives from the fixed-effects formulation for panel-specific terms. The two random components u_{it} and ε_{it} can assume a variety of specifications, including heteroscedasticity and dynamic models for u_{it} for the case of longer panels.

The fixed-effects approach is generally affected by the incidental parameter problem; see Bartolucci *et al.* (2016) for a review. This concerns the joint maximum likelihood estimation (MLE) of the model parameters (θ, α), where $\theta = (\beta, \sigma)$ are the structural parameters and $\alpha = (\alpha_1, \ldots, \alpha_n)$ the panel-specific intercepts. Here σ collects the parameters of the distribution of u_{it} and ε_{it}, which are crucial for efficiency evaluation. In short, the MLE of α is consistent only for $T \to \infty$, and for finite T the resulting bias is transmitted to the MLE of σ. The problem gets worse when n increases for fixed panel size T, and the efficiency evaluation can be heavily affected for settings with many short panels.

To overcome the nuisance parameters issue, Chen *et al.* (2014) proposed a Marginal Maximum Likelihood Estimator (MMLE) for θ. The MMLE is $\sqrt{n}$-consistent as the α are removed by considering a marginal likelihood, akin to the classical within-group estimation approach for fixed-effects panel data models. The MMLE solves the incidental parameter problem entirely, but it is feasible only for normal ε_{it} and truncated normal inefficiencies u_{it}. More general proposals can be found in Belotti and Ilardi (2018). They defined the Marginal Maximum Simulated Likelihood Estimator (MMSLE) and the Pairwise Difference Estimator (PDE). Both these two approaches are defined for exponential or half-normal distributed u_{it}. The MMSLE overcomes some of the computational difficulties of the MMLE by the use of simulation. It can handle some form of heteroscedasticity, but not dynamic models for the inefficiency. The PDE is an ad-hoc $\sqrt{n}$-consistent estimator. It can handle heteroscedasticity and dynamic models for the inefficiency in the truncated normal case.

2 Our proposal

The MMLE/MMSLE and PDE estimators are useful solutions for many settings. Yet the structure of the true fixed-effects model allows for a simpler, fully efficient and more general estimation method. The theory of inference in composite group families (Pace and Salvan, 1997, Chapter 7) readily gives that the integrated likelihood with flat weight function for α gives a marginal likelihood for $\theta = (\beta, \sigma)$.

A general marginal likelihood for θ can be defined as

$$L_M(\beta, \sigma) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} L_i(\alpha_i, \beta, \sigma) d\alpha_i \tag{2}$$

where $L_i(\beta, \sigma, \alpha_i)$ is the likelihood function for the i-th panel corresponding to the model assumed for the data $y_{i1}, \ldots, y_{iT}$ of the i-th panel. Formally, the results reported in Pace and Salvan (1997) are valid for independent observations, but the extension to dynamic efficiency is straightforward.
The fact that the incidental panel-specific parameters α_i are location parameters implies that $L_M(\beta, \sigma)$ has much better properties than the usual integrated likelihood with uniform prior. In fact, Arellano and Bonhomme (2009) showed that in general such integrated likelihood leads to a finite sample bias for the resulting estimator of θ of order $O(T^{-1})$. Here the asymptotic bias disappears for $n \to \infty$ for fixed T, since the maximiser $\widehat{\theta}_M$ of $L_M(\beta, \sigma)$ is $\sqrt{n}$-consistent. Furthermore, being based on a marginal likelihood, the estimator is fully efficient.
The proposed method overcomes all the limitations of both MMLE/MMSLE and PDE, as it is both fully general and fully efficient. Obtaining the estimator requires the approximation of one-dimensional integrals, a standard task for which many methods exist. It should be noted that as both the distribution of u_{it} and of ε_{it} are continuous, the integrand function is always smooth and the required integration simpler than that needed for random-intercepts models for discrete panel data, a rather standard task.
A possible method for obtaining $\widehat{\theta}_M$ is the approximation of the one-dimensional integrals with respect to α_i employing the first-order Laplace's approximation. This is rather simple, but it would introduce a $O(T^{-1})$ error which accumulates across panels. A better resolution is obtaining by the approximation of the integrals in (2) by an accurate adaptive Gauss-Hermite approach (Liu and Pierce, 1994), reducing the approximation error for the computation of the integrals in (2) to a negligible size.

3 Preliminary results

On the computational side, a streamlined approach has been obtained via the `R` package `Template Model Builder (TMB)` for automatic differentiation (Kristensen et al., 2016), which operates via `C++` templates.
Some simulation studies are ongoing for studying the performances of $\widehat{\theta}_M$. The preliminary results seem to confirm the good properties of the methods. To give a flavour about the results, we include here some illustrative results for one of the settings considered by Belotti and Ilardi (2018). In particular, we take half-normally distributed u_{it} and normal errors. This is an important benchmark, since the estimator from the integrated likelihood should agree with the MMLE of Chen *et al.* (2014). Indeed, we implemented both

TABLE 1. Some simulation results with half-normal inefficiencies.

θ	$\widehat{\theta}_L$		$\widehat{\theta}_M$	
	Bias (MSE) $n = 100$	**Bias (MSE)** $n = 250$	**Bias (MSE)** $n = 100$	**Bias (MSE)** $n = 250$
β	0.004 (0.008)	0.001 (0.004)	0.004 (0.008)	0.001 (0.004)
σ_u	-0.184 (0.198)	-0.148 (0.114)	-0.133 (0.191)	-0.089 (0.101)
σ_ε	0.015 (0.011)	0.024 (0.006)	-0.002 (0.012)	0.006 (0.006)

these two methods, and the two sets of estimates agree with very high accuracy. Here we take a single covariate x_{it} and generate its values and those of all the model parameters following Belotti and Ilardi (2018, case $\sigma_u = \sigma_\varepsilon$), to which we refer for a more comprehensive description. Table 3 summarizes the results for $\widehat{\theta}_M$ from (2) for 1000 simulated samples, with $n = 100$ and $n = 250$, with T fixed at 5. For a comparison, we also report the estimator $\widehat{\theta}_L$ that approximates (2) using the first-order Laplace approximation. The superiority of $\widehat{\theta}_M$ seems apparent.

References

Arellano, M. and Bonhomme, S. (2009). Robust priors in nonlinear panel data models. *Econometrica*, 77, 489–536.

Bartolucci, F., Bellio, R., Salvan, A. and Sartori, N. (2016). Modified profile likelihood for fixed-effects panel data models. *Econometric Reviews*, 35, 1271–1289.

Belotti, F. and Ilardi, G. (2018). Consistent inference in fixed-effects stochastic frontier models. *Journal of Econometrics*, 202, 161–177.

Chen, Y.Y., Schmidt, P. and Wang, H.J. (2014). Consistent estimation of the fixed effects stochastic frontier model. *Journal of Econometrics*, 181, 65–76.

Greene, W. (2005). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics*, 126, 269–303.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H.J. and Bell, B. (2016). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, 70.

Liu, Q. and Pierce, D.A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 90, 624–629.

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

Correction for the shrinkage effect in Gaussian graphical models

Victor Bernal[1 2], Victor Guryev[3], Rainer Bischoff[2], Peter Horvatovich[2], Marco Grzegorczyk[1]

[1] Bernoulli Institute, University of Groningen, Groningen, NL.
[2] Department of Pharmacy, Analytical Biochemistry, University of Groningen, Groningen, NL.
[3] Universitair Medisch Centrum Groningen (UMCG), ERIBA, University of Groningen, Groningen, NL.

E-mail for correspondence: v.a.bernal.arzola@rug.nl

Abstract: Gaussian graphical models (GGMs) are probabilistic graphical models based on partial correlation. A GGM consists of a network of nodes (representing the random variables) connected by edges (their partial correlation). To infer a GGM, the inverse of the covariance matrix (the precision matrix) is required. The main challenge is that when the number of variables is larger than the sample size, the (sample) covariance is ill conditioned (or not invertible). Shrinkage methods consist in regularizing the estimator of the covariance matrix to make it invertible (and well conditioned); however, the effect of the shrinkage on the final network topology has not been studied so far.

Keywords: Gaussian graphical models; Shrinkage; Genetic Networks; Partial correlation.

1 Introduction

Gaussian graphical models (GGMs) are un-directed graphical models represented by a matrix of partial correlations. They are a popular tool to infer regulatory networks from quantitative molecular profiles (e.g. gene-expression data). In a GGM each random variable is a node and an edge is placed between node-pairs according to their partial correlation. Partial correlations measure the linear dependence between a pair of random variables after adjusting for the contribution from all the others. The resulting GGM structure encodes full conditional correlations in

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

the form of a network (i.e. a matrix of pair-wise partial correlations). The inference of the partial correlations demands the estimation of the inverse of the covariance matrix. Therefore the covariance estimator needs to be invertible and well-conditioned. The sample covariance estimator $\hat{\mathbf{C}}_{\mathrm{sm}}$ with p variables and n samples is not invertible if $n \ll p$; a common scenario in systems biology; usually referred to as a "high dimensional problem", or "small n, large p". Shrinkage is a method to regularize the covariance estimator, where the resulting "shrunk" estimator is well conditioned, and invertible. This "shrunk" covariance estimator can be further used to compute the "shrunk" partial correlations. In this work, we study the role of the shrinkage on the network topology. We will show that is possible to de-regularize or "un-shrink" the partial correlation.

2 Shrinkage based Gaussian graphical models

Partial correlations are a measure of full-conditional linear dependence between two variables (where the effects coming from all other variables are adjusted). Gaussian graphical models (GGMs)are represented with a matrix $\mathbf{P}$ where the entry ij is the partial correlation between the variables i and j. $\mathbf{P}$ can be found from the covariance matrix $\mathbf{C}$, and/or from its inverse $\mathbf{C}^{-1}$ via

$$\mathbf{P}_{\mathrm{ij}} = -\frac{\mathbf{C}^{-1}_{\mathrm{ij}}}{\sqrt{\mathbf{C}^{-1}_{\mathrm{ii}}\mathbf{C}^{-1}_{\mathrm{jj}}}} \tag{1}$$

In principle, $\mathbf{C}$ can be estimated from the data; however, the sample covariance estimator $\hat{\mathbf{C}}_{\mathrm{sm}}$ is ill-conditioned (or non-invertible) when $n \leq p$. In particular, the LW- covariance $\hat{\mathbf{C}}^{[\lambda]}$ is a (convex) linear combination of $\hat{\mathbf{C}}_{sm}$ with a target estimator $\mathbf{T}$ (e.g. a diagonal matrix)(Ledoit and Wolf(2004)). The "shrunk" estimator takes the form $\hat{\mathbf{C}}^{[\lambda]} = \lambda\mathbf{T} + (1-\lambda)\hat{\mathbf{C}}_{\mathrm{sm}}$. The shrinkage λ is in the interval $(0, 1)$, and is usually fixed according to an optimization criteria. In this way, the inverse of $\hat{\mathbf{C}}^{[\lambda]}$ is used in Equation 1 to estimate "shrunk" partial correlations $\mathbf{P}^{[\lambda]}$(Schäfer and Strimmer (2005.)).
However, from Equation 1 we see that λ has a non-linear effect on the resulting partial correlation (through the matrix inversion and square roots). Consequently, GGMs inferred with different shrinkages e.g. from different experimental conditions can not be compared. Therefore, the shrinkage effect needs to be first removed from the partial correlation.

In this way, we define the "un-shrunk" partial correlation $\mathbf{P}^{[0]}$ as the limit of $\mathbf{P}^{[\lambda]}$ when λ approaches to zero. Symbolically,

$$\lim_{\lambda \to 0} \mathbf{P}^{[\lambda]}_{\mathrm{ij}} = \mathbf{P}^{[0]}_{\mathrm{ij}} \tag{2}$$

To ensure that $\mathbf{P}^{[\lambda]}$ is a continuous function of λ we recall the following theorems. First, the inverse of a matrix can be written in term of determinants. Second, that determinants are polynomial (in this case are polynomial of $1-\lambda$). Third, that determinants, quotients and square roots are continuous functions (for positive arguments and provided that the denominator is not zero). Therefore, $\mathbf{P}^{[\lambda]}$ is a continuous function of λ. We show that it is a bounded function as well; any term $1/\lambda$ will cancel in the quotient. The limit when $\lambda \to 0$ is well-defined, and the shrinkage distortion can be removed obtaining the "un-shrunk" partial correlation. In the following sections, we approximate $\mathbf{P}^{[0]}$ via a polynomial model fitted in the range $0.1 \leq \lambda \leq 1$, and extrapolating the model to $\lambda = 0$.

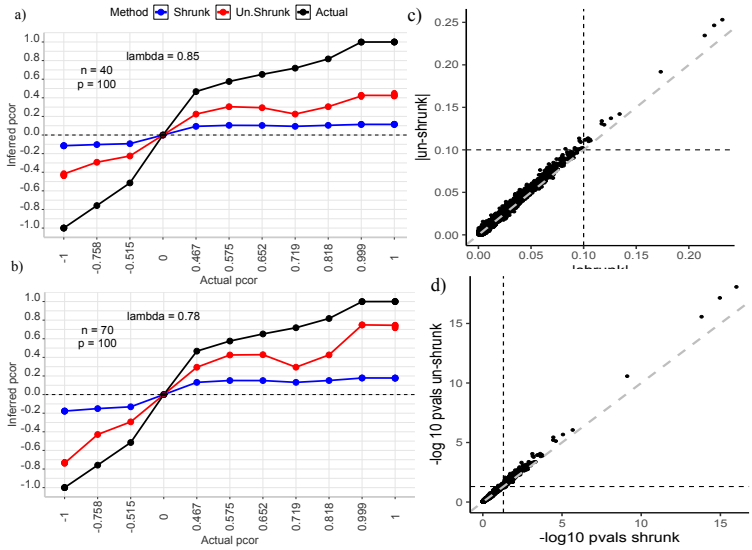


FIGURE 1. The "shrunk" and "un-shrunk" partial correlation. In black: the positives edges (non-zero partial correlations). In blue: the average of the "shrunk" method. In red: the average of the new "un-shrunk" method. Error bars represent 2 standard errors. Panels c, d) Edge-wise and $\log_{10}$ p-value comparison for E. coli dataset. The panels are segmented into four regions using a threshold of $|$ pcorr $|= 0.1$ or p-values $= 0.05$.

3 Results

In this section, we simulate a network structure and Gaussian random data ($p = 100$, and $n = 40,\ 70$). Figure 1 shows the average partial correlations (over 25 simulations) obtained from simulated data with each

model. We observe that the "un-shrunk" model leads to better results. Subsequently, we use both methods to model gene interactions from microarray expression data. The data comes from a study of Escherichia coli after IPTG induction of the recombinant protein SOD ($p = 102$, $n = 9$). As the ground truth is unknown, we compare the results with scatter plots for the partial correlations and for their p-values (Bernal *et al.*(2019)). Points scattering away from the diagonal line reflect that the edges have different order. In particular, the upper left and lower right quadrants display the edges that are significant exclusively in one of the methods. We observe different network structure with additional connection from the new method.

4 Discussion and Conclusions

We show that the partial correlation can be de-regularized (or "un-shrunk"), and that the numerical instabilities (that originally required the shrinkage) can be avoided. Our "un-shrunk" method is consistently closer to the population value compared to "shrunk" result (see Figure 1). This makes the "un-shrunk" estimator superior in terms of interpretability and cross comparison of networks. For the E. coli dataset, the strongest partial correlations (in both networks) were lacA-lacZ, lacY-lacZ, and lacA-lacY, all related to the lac operon (that was induced by IPTG in the experiment). Additionally, the new "un-shrunk" model retrieves 34 significant partial correlations (p-values ≤ 0.05) that were not found with the traditional approach.

Acknowledgments: We want to acknowledge the Data Science and System Complexity Center (DSSC) of the University of Groningen.

References

Ledoit,O. and Wolf,M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, **88**, 365 – 411.

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**, 1 – 30.

Bernal,V. et al. (2019). Exact hypothesis testing for shrinkage-based Gaussian graphical models. *Bioinformatics*, **4**, 1 – 70.

Schmidt-Heck, W. et al. (2004). Reverse Engineering of the Stress Response during Expression of a Recombinant Protein. *Proceedings of the EUNITE symposium*, Aachen, 10 – 12.

Uncertainty propagation in shrinkage-based partial correlations

Victor Bernal[1 2], Victor Guryev[3], Rainer Bischoff[2], Peter Horvatovich[2], Marco Grzegorczyk[1]

[1] Bernoulli Institute, University of Groningen, Groningen, NL.
[2] Department of Pharmacy, Analytical Biochemistry, University of Groningen, Groningen, NL.
[3] Universitair Medisch Centrum Groningen (UMCG), ERIBA, University of Groningen, Groningen, NL.

E-mail for correspondence: v.a.bernal.arzola@rug.nl

Abstract: Gaussian graphical models (GGMs) are network models where random variables are represented by nodes and their pair-wise partial correlation by edges. The inference of a GGM demands the estimation of the precision matrix (i.e. the inverse of the covariance matrix); however, this becomes problematic when the number of variables is larger than the sample size. Covariance estimators based on shrinkage (a type of regularization) overcome these pitfalls and result in a 'shrunk' version of the GGM. Traditionally, shrinkage is justified at model level (as a regularized covariance). In this work, we re-interpret the shrinkage from a data level perspective (as a regularized data). Our result allows the propagation of uncertainty from the data into the GGM structure.

Keywords: Gaussian Graphical Models; Shrinkage; Genetic Networks; Partial correlation.

1 Introduction

Gaussian graphical models (GGMs) are network models consisting of nodes (the random variables) inter-wired by edges (their pair-wise partial correlations). Partial correlations measure the correlation between pairs of full conditional variables. The estimation (inference) of a GGM structure requires the inverse of the covariance matrix (i.e. the precision matrix), however, when the number of variables is larger than the sample size the sample covariance estimator is ill conditioned (or not invertible). Other covariance

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

estimators based on shrinkage have proven to be useful in this case. They consist of a regularized estimator that is invertible at the expenses of introducing some bias (due to the bias-variance trade-off). Typically, the interpretation of the shrinkage has been in terms of the covariance matrix, however, this justification is half way divorced from the data (the main subject of study). In this work, we present the shrinkage from a data level perspective and study how the data uncertainty propagates through the analysis.

2 Methods

Gaussian graphical models (GGMs) are represented with a matrix of partial correlations $\mathbf{P}$. The element ij of $\mathbf{P}$ is the partial correlation between the variables i and j, and it is given by

$$\mathbf{P}_{\mathrm{ij}} = -\frac{\mathbf{\Omega}_{\mathrm{ij}}}{\sqrt{\mathbf{\Omega}_{\mathrm{ii}}\mathbf{\Omega}_{\mathrm{jj}}}} \tag{1}$$

where $\mathbf{\Omega} = \mathbf{C^{-1}}$. In principle, $\mathbf{C}$ is unknown but can be estimated from the data e.g. by means of the sample covariance $\hat{\mathbf{C}}_{\mathrm{sm}}$,

$$\hat{\mathbf{C}}_{\mathrm{sm}} = \frac{1}{n-1}\mathrm{D}^{\mathrm{t}}\,\mathrm{D} \tag{2}$$

with D being the (centered) data matrix with p variables (columns) and n samples (rows). $\mathbf{P}$ can be estimated *indirectly* using the inverse $\hat{\mathbf{C}}_{\mathrm{sm}}$ in Equation 1, however, when n is less or equal p, $\hat{\mathbf{C}}_{\mathrm{sm}}$ becomes ill-conditioned or non-invertible. A well-conditioned alternative is the LW-covariance $\hat{\mathbf{C}}^{[\lambda]}$ (Ledoit and Wolf (2004)) $\hat{\mathbf{C}}^{[\lambda]}$ which consists of a (convex) linear combination of $\hat{\mathbf{C}}_{\mathrm{sm}}$ and a target $\mathbf{T}$ as,

$$\hat{\mathbf{C}}^{[\lambda]} = (1-\lambda)\hat{\mathbf{C}}_{\mathrm{sm}} + \lambda\mathbf{T} \tag{3}$$

where the shrinkage $\lambda \in (0,1)$ is fixed according to an optimization criterion (Ledoit and Wolf (2004)). The inverse of $\hat{\mathbf{C}}^{[\lambda]}$ can replace $\mathbf{\Omega}$ in Equation 1 to obtain a 'shrunk' partial correlation $\mathbf{P}^{[\lambda]}$ (Schäfer and Strimmer (2004)). In this study we will assume that all variables in D were centered (substracting their sample averages) and that they have the same variance σ^2.

3 Results

3.1 'Shrunk' data

To re-interpret the shrinkage at data level we turn our attention to C and its eigenvalues α, and to their estimates $\hat{\mathbf{C}}_{\mathrm{sm}}$ and $\hat{\alpha}$. Using Equation 3 we get that

$$\hat{\alpha_i}^{[\lambda]} = (1-\lambda)\hat{\alpha_i} + \lambda\hat{\sigma^2} \tag{4}$$

where $\hat{\alpha_i}^{[\lambda]}$ is i-th eigenvalue of $\hat{\mathbf{C}}^{[\lambda]}$, and $\hat{\sigma}^2$ are the variances of $\hat{\mathbf{C}}_{\mathrm{sm}}$. Using the Singular value decomposition (SVD),

$$\hat{\mathbf{C}}^{[\lambda]} = (\mathrm{U}\ \mathrm{diag}(\pm\sqrt{\hat{\alpha}^{[\lambda]}})\ \mathrm{V}^{\mathbf{t}})^{\mathbf{t}}(\mathrm{U}\ \mathrm{diag}(\pm\sqrt{\hat{\alpha}^{[\lambda]}})\ \mathrm{V}^{\mathbf{t}}) \tag{5}$$

with the matrices U and V are the (left and right) singular vectors of $\hat{\mathbf{C}}_{\mathrm{sm}}$, and $\mathrm{diag}(\pm\sqrt{\hat{\alpha}^{[\lambda]}})$ is the diagonal matrix of singular values (the square root of the eigenvalues). Comparing Equation 2 to Equation 5 we find the 'shrunk' data $\mathrm{D}^{[\lambda]}$ as,

$$\mathrm{D}^{[\lambda]} = \mathrm{U}\ \mathrm{diag}(\pm\sqrt{(n-1)\hat{\alpha}^{[\lambda]}})\ \mathrm{V}^{\mathbf{t}} \tag{6}$$

where n is known while U, V, and the $\hat{\alpha_i}^{[\lambda]}$ can be computed from the SVD of D (without $\pm$ sign ambiguity).

3.2 'Shrunk' residuals

Correlations result from the standardized covariance matrix $\mathbf{C}$, while partial correlation from the standardized $\mathbf{\Omega} = \mathbf{C^{-1}}$. In this sense, $\mathbf{\Omega}$ encodes the covariances between full conditioned random variables. Resembling Equation 2 we can write that,

$$\hat{\mathbf{\Omega}}^{[\lambda]} = \frac{1}{\mathrm{k}}\left(\mathrm{Res}^{[\lambda]}\right)^{\mathbf{t}}\left(\mathrm{Res}^{[\lambda]}\right) \tag{7}$$

where $\mathrm{Res}^{[\lambda]}$ is the (centered) matrix of residuals, and k their degrees of freedom. To find $\mathrm{Res}^{[\lambda]}$ we can use the SVD of $\hat{\mathbf{\Omega}}^{[\lambda]}$. For this, we recall two facts between a matrix and its inverse: (i) that they share the same set of eigenvectors, and (ii) that their eigenvalues are reciprocals. Then,

$$\hat{\mathbf{\Omega}}^{[\lambda]} = (\mathrm{U}\ \mathrm{diag}(\pm\sqrt{\frac{1}{\hat{\alpha}^{[\lambda]}}})\ \mathrm{V}^{\mathbf{t}})^{\mathbf{t}}\ (\mathrm{U}\ \mathrm{diag}(\pm\sqrt{\frac{1}{\hat{\alpha}^{[\lambda]}}})\ \mathrm{V}^{\mathbf{t}}) \tag{8}$$

comparing Equation 7 to Equation 8 we find the 'shrunk' residuals,

$$\mathrm{Res}^{[\lambda]} = \mathrm{U}\ \mathrm{diag}(\pm\sqrt{\frac{\mathrm{k}}{\hat{\alpha}^{[\lambda]}}})\ \mathrm{V}^{\mathbf{t}} \tag{9}$$

3.3 Propagation of uncertainty

Often the measurement of data would include some level of technical (or external) variablity ϵ which is usually modeled as an additive *iid* $N(0, {\sigma_\epsilon}^2\mathbb{I})$. In contrast to sampling variability, ϵ does not decreases with larger sample

sizes. Linearity of the covariance implies that $\mathrm{cov}[\mathrm{X_i} + \epsilon_\mathrm{i}, \mathrm{X_j} + \epsilon_\mathrm{j}]$ is equal to $\mathrm{cov}[\mathrm{X_i}, \mathrm{X_j}]$ for $i \neq j$, and equal to $\mathrm{cov}[\mathrm{X_i}, \mathrm{X_j}] + {\sigma_\epsilon}^2\mathbb{I}$ otherwise. Therefore,

$$\mathbf{C}_\epsilon = \mathbf{C} + {\sigma_\epsilon}^{\mathbf{2}}\mathbb{I} \qquad \Rightarrow \mathbf{C}_\epsilon \vec{u} \qquad = (\alpha + {\sigma_\epsilon}^2)\vec{u} \tag{10}$$

In principle, $\hat{\mathbf{C}}_{\mathrm{sm}}$ is an un-biased estimator of $\mathbf{C}$. Now, the presence of ϵ turns $\hat{\mathbf{C}}_{\mathrm{sm}}$ into an estimator of $\mathbf{C}_\epsilon$. The uncertainty coming from the collected measurements in D propagates to the (partial) correlation by replacing $\hat{\alpha}$ with $\hat{\alpha} + \hat{\sigma_\epsilon^2}$ in Equations 5-9.

4 Discussion and Conclusions

In real application data is limited which can make the sample covariance ill-conditioned. Shrinkage (regularization) approaches overcomes this, however, it's interpretation has been limited at the model level (as a modification of the covariance). While valid, it is half way divorced from the data (the original subject of study). In this work, we used SVD to show that shrinking the covariance is equivalent to transforming the data into a 'shrunk' data. With this result, shrinkage based (partial) correlations can be interpreted as classical (partial) correlations estimated from a 'shrunk' data. Equations 6 and 9 illustrate the shrinkage role from a data-driven perspective. Additionally, we showed how uncertainty in the data measurements propagate through the analysis. Among the limitations in our study we have that (i) the assumption of equal variances in the data, that (ii) the SVD is not unique, e.g. for degenerate (repeated) singular values their singular vectors can be permuted, and that (iii) many data sets can produce the same sample covariance matrix.

Acknowledgments: We want to acknowledge the Data Science and System Complexity Center (DSSC) of the University of Groningen.

References

Ledoit,O. and Wolf,M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, **88**, 365 – 411.

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**, 1 – 30.

Modelling the impact of spatial proximity on scientific collaboration networks

Hannah Busen[12], Christiane Fuchs[123]

[1] Faculty of Business Administration and Economics, Bielefeld University, Germany
[2] Institute of Computational Biology, Helmholtz Zentrum Mnchen, Germany
[3] Faculty of Mathematics, Technical University Munich, Germany

E-mail for correspondence: christiane.fuchs@uni-bielefeld.de

Abstract: Spatial proximity between researchers may lead to more frequent or more intense collaboration than between scientists who work at large distance from each other. We aim to investigate the impact of proximity within research institutions on the implementation of interdisciplinary collaborations. Our data contains publications and building distances from two research institutions, the Helmholtz Zentrum Mnchen and Bielefeld University. Defining collaboration as the number of joint publications, we use exploratory and network analyses to answer the question if researchers are more likely to work together if they are at small distance to each other. The methodological focus lies on accounting for the dependency structure in network data. Outcomes of this study may inform about how to target the promotion of interdisciplinary research.

Keywords: Collaboration networks; Spatial proximity; QAP; Hurdle model.

1 Introduction

Research at Bielefeld University has been self-characterized by its guiding principle of interdisciplinarity. The special structure of its main building allows researchers from different faculties to meet each other without going outside. The spatial design clearly differs at the Hemholtz Zentrum Mnchen, a research center for environment and health: Although many institutes are united on one main campus, they are distributed over individual buildings. Here, too, interdisciplinary research plays an important role. Based on this observation and inspired by the work of Claudel *et al.* (2017), we aim to assess the role of building and campus structures for interdisciplinary research: How does spatial proximity between two scientists

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

influence their collaboration? We measure the strength of interdisciplinary collaboration by the number of joint publications between two researchers from two different faculties or departments.

2 Data Collection and Preprocessing

We consider open data from the publication databases *Web of Science*, *Scopus* and *Pubmed* for Helmholtz Zentrum M´unchen and Bielefeld University. This source enables us to easily extend our analyses to other research institutions, in contrast to data from local libraries which use non-uniform data formats. We use the R package `bibliometrix` (Aria and Cuccurullo, 2017) to merge the data from the three databases. The final data set contains the title, abstract, authors, detailed affiliations, publication year, document type and journal of each publication. Because of ambiguous writings of the affiliations (including departments and working groups), we use cluster analysis and string matching to assign the authors to their main faculties and institutes. Since GPS-based measurement of spatial distances was impossible inside the Bielefeld university building, we assessed the distance between institutions as the number of steps on foot and also considered the effort to use stairs and lifts. To be consistent, we applied the same procedure at the campus of the Helmholtz Zentrum Mnchen.

3 Methods

We analyze publication and distance data descriptively and by forming collaboration networks of author pairs. Here, each author is one node, and two authors are connected if they have at least one joint publication. In addition, we use regression models to describe the effect of distance on the number of publications. This poses some challenges: Firstly, there is row- and columnwise auto-correlation due to the network structure, and the observations cannot be seen as independent. Second, there is a natural excess of zeros in the publication data since every author can only collaborate with a small fraction of all other authors.

We account for these issues as follows: The quadratic assignment procedure (QAP; Krackhardt, 1987) accounts for autocorrelation and adjusts error terms in binary networks where the interest lies in the existence or non-existence of an edge. QAP is typically combined with logistic regression (e. g. Broekel and Hartog, 2011). Here, we link it to a hurdle model (Mullahy, 1986; Zeileis *et al.*, 2008) to analyse whether two authors publish together and if so, how many joint publications they have. Our hurdle model mixes a binomial distribution for the zeros with a Poisson distribution (restricted to positive numbers) for the non-zero counts.

4 Results

For space restrictions, we exemplarily present results for publication data from the year 2015. Figure 1 shows for both considered institutions empirical cumulative distribution functions of all possible distances between each pair of researchers, independently of whether they published together or not (grey), and the distances of author pairs which actually published together (blue). We would expect both curves to describe the same distribution if there was no effect of spatial distance on the number of joint publications. The Kolmogorov-Smirnov test yields for both Munich and Bielefeld a p-value of less than $2 \cdot 10^{-16}$, implying that the null hypothesis that the samples are drawn from the same distribution can be rejected at 5% level.

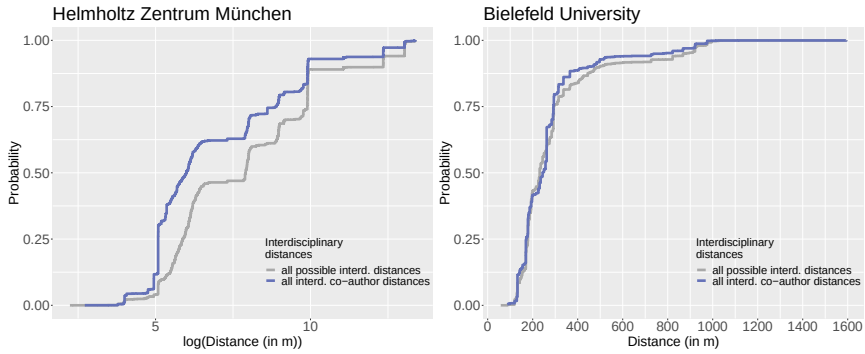


FIGURE 1. Empirical cumulative distribution functions of spatial distances. Left: Helmholtz Zentrum Mnchen, right: Bielefeld University

Estimates of the hurdle model for the year 2015 show a small negative effect of distance on the number of joint publications for Helmholtz Zentrum Mnchen when considering between-institute distances smaller than 1000m (zero part: -0.0012, count part: -0.0005). Thus, both the probability of active collaboration and the number of interdisciplinary publications decrease with increasing distance. According to the QAP results, both coefficients are significant at 5% level with a p-value close to zero (zero part) and 0.023 (count part). For distances smaller than 6000m, the coefficients are -0.0002 for the zero and -0.0001 for the count component (QAP p-values: close to zero and 0.002). For Bielefeld University, we obtain coefficients 0.0010 for the zero and 0.0004 for the count part. The QAP p-values are 0.032 and 0.220, i. e. there is a small positive significant effect at 5% significance level for the zero part, indicating that more distant authors are more likely to collaborate, independently of the number of publications.

5 Discussion

We suspect that spatial distance may generally have an effect on interdisciplinary collaboration. This is supported by our results for Helmholtz Zentrum M´unchen. At Bielefeld University, interdisciplinarity is facilitated by, amongst others, the special structure of the main building, where all faculties are located under one roof. Given these conditions, spatial distance matters less. Interpretation of associations has to consider the possibility of spurious correlations: While it may be that close spatial proximity leads to more intensive collaboration, it may also well be that office space has been assigned in close proximity *because of* well-known scientific links between two research fields. The risk of wrong conclusions can be reduced by including additional data on work across research fields. Having examined the interdisciplinary publication behaviour at one university and one research center, there remain comparisons to other institutions as well as the inclusion of several other influencing factors such as scientific focus, teaching assignment, budget and internationality, to mention just a few.

Acknowledgments: We thank Irina Janzen, Nina Langius and Minh Viet Tran for help in data collection and the Faculty of Business Administration and Economics at Bielefeld University for project funding.

References

Aria, M. and Cuccurullo, C. (2017) *bibliometrix: An R-tool for comprehensive science mapping analysis.* Journal of Informetrics, 11(4), 959-975.

Broekel, T., Hartog. M. (2011) *Explaining the structure of inter-organizational networks using exponential random graph models: does proximity matter?* Industry and Innovation 20(3).

Claudel, M., Massaro, E., Santi, P., Murray, F. and Ratti, C. (2017) *An exploration of collaborative scientific production at MIT through spatial organization and institutional affiliation.* PLoS ONE 12(6): e0179334.

Mullahy, J. (1986) *Specification and testing of some modified count data Models.* Journal of Econometrics, 33, 341365.

Krackhardt, D. (1987) *QAP Partialling as a test of spuriousness.* Social Networks 9, pp. 171-186.

Zeileis, A., Kleiber, C., Jackmann, S. (2008) *Regression models for count data in R.* Journal of statistical software 27(8): 1-25

Bayesian shared-parameter models for analysing sardine fishing in the Mediterranean Sea

Gabriel Calvo[1], Carmen Armero[1], Maria Grazia Pennino[2], Luigi Spezia[3]

[1] Universitat de València, Spain
[2] Instituto Español de Oceanografía, Spain
[3] Biomathematics & Statistics Scotland, Aberdeen, UK

E-mail for correspondence: gabriel.calvo@uv.es

Abstract: European sardine is experiencing an overfishing around the world. The dynamics of the industrial and artisanal fishing in the Mediterranean Sea from 1970 to 2014 by country was assessed by means of Bayesian joint longitudinal modelling that uses the random effects to generate an association structure between both longitudinal measures. Model selection was based on Bayes factors approximated through the harmonic mean.

Keywords: Joint modelling; Longitudinal data; Model comparison.

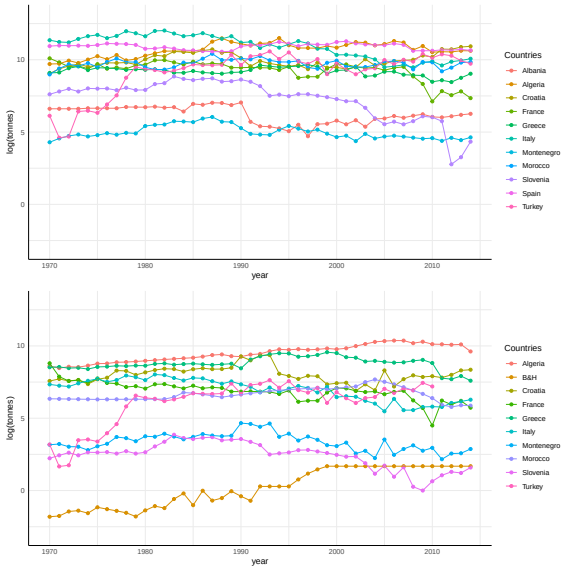
1 Introduction

European sardine (*Sardina pilchardus*) is one of the most commercial species showing high over-exploitation rates over the last years in the Mediterranean Sea. Mediterranean fisheries are highly diverse and geographically varied due not only to the existence of different marine environments, but also because of different socio-economic situations and fisheries status.

We consider data of European sardine landings from 1970 to 2014 from both the artisanal and the industrial fisheries which are defined in terms of small-scale and large-scale commercial fisheries, respectively. Data are recorded by country (Albania, Algeria, Bosnia and Herzegovina, Croatia, France, Greece, Italy, Montenegro, Morocco, Slovenia, Spain and Turkey) and come from *Sea Around Us* (www.seaaroundus.org), a research initiative at the University of British Columbia.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

Top plot of the next figure shows the dynamics of the industrial amount of fish caught, in logarithmic scale, from 1970 to 2014 in all the Mediterranean countries included in the study. Bottom plot presents the dynamics of artisanal fishing.



2 Bayesian joint modelling

Let $Y_{it}^{(I)}$ and $Y_{it}^{(A)}$ be the amount of sardine caught by industrial and artisanal methods in the country i during year t, respectively. Calendar time is the time scale and $t = 0$ is the first year of the study (i.e. 1970).
We assume a Bayesian shared-parameter approach to jointly model both processes that uses the random effects to generate an association structure between both longitudinal measures. The joint distribution of the longitudinal fishing vectors, $\boldsymbol{Y}_i^{(I)} = (Y_{i0}^{(I)}, \ldots, Y_{iT}^{(I)})$ and $\boldsymbol{Y}_i^{(A)} = (Y_{i0}^{(A)}, \ldots, Y_{iT}^{(A)})$, parameters and hyperparameters $\boldsymbol{\theta}$, and random effects $\boldsymbol{b}_i$ for country i is:

$$\begin{aligned} f(\boldsymbol{y}_i^{(I)}, \boldsymbol{y}_i^{(A)}, \boldsymbol{\theta}, \boldsymbol{b}_i) &= f(\boldsymbol{y}_i^{(I)}, \boldsymbol{y}_i^{(A)} | \boldsymbol{\theta}, \boldsymbol{b}_i) f(\boldsymbol{b}_i | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \\ &= f(\boldsymbol{y}_i^{(I)} | \boldsymbol{\theta}, \boldsymbol{b}_i^{(I)}) f(\boldsymbol{y}_i^{(A)} | \boldsymbol{\theta}, \boldsymbol{b}_i^{(A)}) f(\boldsymbol{b}_i | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}). \end{aligned}$$

We propose two specific models for $f(\boldsymbol{y}_i^{(I)} | \boldsymbol{\theta}, \boldsymbol{b}_i^{(I)})$ and $f(\boldsymbol{y}_i^{(A)} | \boldsymbol{\theta}, \boldsymbol{b}_i^{(A)})$ within the framework of mixed linear models. The random effects vector for country i can be divided in two subvectors, $\boldsymbol{b}_i = (\boldsymbol{b}_i^{(I)}, \boldsymbol{b}_i^{(A)})$ corresponding to industrial and artisanal fishing, respectively. In addition, we impose a structure of association between the random effects associated to the industrial and artisanal fishing, $f(b_{0i}^{(I)}, b_{0i}^{(A)} | \Sigma_0) = \mathrm{N}(0, \Sigma_0)$ and

$f(b_{1i}^{(I)}, b_{1i}^{(A)}|\Sigma_1) = \mathrm{N}(0, \Sigma_1)$, with variance - covariance matrices given by

$$\Sigma_0 = \begin{pmatrix} {\sigma_0^{(I)}}^2 & \rho_0\sigma_0^{(I)}\sigma_0^{(A)} \\ \rho_0\sigma_0^{(I)}\sigma_0^{(A)} & {\sigma_0^{(A)}}^2 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} {\sigma_1^{(I)}}^2 & \rho_1\sigma_1^{(I)}\sigma_1^{(A)} \\ \rho_1\sigma_1^{(I)}\sigma_1^{(A)} & {\sigma_1^{(A)}}^2 \end{pmatrix}.$$

Both models can be expressed in terms of a conditional normal density of the subsequent longitudinal model given parameters, hyperparameters and random effects, $f(\boldsymbol{y}_i^{(I)}|\boldsymbol{\theta}, \boldsymbol{b}_i) = \mathrm{N}(\boldsymbol{\mu}_i^{(I)}, \sigma^2 I)$, $f(\boldsymbol{y}_i^{(A)}|\boldsymbol{\theta}, \boldsymbol{b}_i) = \mathrm{N}(\boldsymbol{\mu}_i^{(A)}, \sigma^2 I)$. The two models differ in the conditional mean by the inclusion in one of them of an autoregressive term as it can be seen in Table 1.

TABLE 1. Conditional mean of the amount of sardine caught by industrial and artisanal methods in the country i during year t specified by each of the proposed models.

Model	$\mu_{it}^{(I)}$	$\mu_{it}^{(A)}$
$M1$	$\beta_0^{(I)} + b_{0i}^{(I)} + b_{1i}^{(I)}t$	$\beta_0^{(A)} + b_{0i}^{(A)} + b_{1i}^{(A)}t$
$M2$	$\beta_0^{(I)} + b_{0i}^{(I)} + b_{1i}^{(I)}t + \rho^{(I)}w_{i,t-1}^{(I)}$	$\beta_0^{(A)} + b_{0i}^{(A)} + b_{1i}^{(A)}t + \rho^{(A)}w_{i,t-1}^{(A)}$

Coefficients $\beta_0^{(I)}$ and $\beta_0^{(A)}$ are the regression industrial and artisanal fishing intercept, respectively. The autoregressive term in model $M2$ for industrial and artisanal fishing in country i is $w_{i,t-1}^{(I)} = y_{i,t-1}^{(I)} - (\beta_0^{(I)} + b_{0i}^{(I)} + b_{1i}^{(I)}(t-1))$ and $w_{i,t-1}^{(A)} = y_{i,t-1}^{(A)} - (\beta_0^{(A)} + b_{0i}^{(A)} + b_{1i}^{(A)}(t-1))$ (Weiss, 2005).
To fully specify the Bayesian model we elicit a prior distribution for all the uncertainties in the model. We assume a noninformative prior scenario with prior independence: normal distributions for the regression coefficients and uniform distributions for all standard deviation parameters. The prior for the autoregressive parameters is $\mathrm{U}(-1, 1)$ to induce the stationarity of $w_{it}^{(I)}$ and $w_{it}^{(A)}$, as well as for the correlation parameters ρ_0 and ρ_1.

3 Posterior inferences

The posterior distribution for both models was approximated via JAGS software (Plummer, 2003) through Markov chain Monte Carlo simulation. Table 2 summarizes the approximate posterior distribution for the models of our study.
Results indicate that the random effects associated with each country play an important role in every model. Although the deviation of the random trends $\sigma_1^{(*)}$ has a small value, little variations on the trend produce big changes over time. Since the response variable is the logarithm of the tonnes, the random effects on the trend associated with each country play an important role in these models. On the other hand, the inclusion of the autoregressive term seems to absorb a large part of the variability explained

TABLE 2. Summary of the approximate sample from the posterior distribution for models $M1$ and $M2$.

	$M1$		$M2$	
	mean	sd	mean	sd
$\beta_0^{(I)}$	8.731	0.875	8.243	0.793
$\beta_0^{(A)}$	5.651	1.208	5.707	1.155
ρ_0	0.673	0.258	0.695	0.259
ρ_1	0.900	0.097	0.763	0.253
$\rho^{(I)}$	-	-	0.806	0.042
$\rho^{(A)}$	-	-	0.916	0.035
$\sigma_0^{(I)}$	2.648	0.786	2.526	0.745
$\sigma_0^{(A)}$	3.823	1.014	3.632	1.046
$\sigma_1^{(I)}$	0.051	0.013	0.045	0.013
$\sigma_1^{(A)}$	0.052	0.012	0.037	0.015
σ	0.565	0.014	0.349	0.008

by the rest of the random effects and consequently, random effects become less correlated.

As a first approach to model comparison, we have computed the marginal likelihood for each model by means of the harmonic mean (Newton and Raftery, 1994). The values obtained for models $M1$ and $M2$ in logarithmic scale are -817.74 and -523.27, respectively. The subsequent Bayes factors provide a decisive evidence in favour of the joint autoregressive model.

Acknowledgments: Calvos research was funded by the ONCE Foundation and the Spanish Ministry of Education and Professional Training, grant FPU18/03101. Spezias research was funded by the Scottish Governments Rural and Environment Science and Analytical Services Division.

References

Armero, C., Forte, A., Perpiñán, H., Sanahuja, M. J. and Agustí, S. (2018). Bayesian joint modeling for assessing the progression of chronic kidney disease in children. *Statistical Methods in Medical Research*, **27**, 298 – 311.

Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B*, **56**, 3 – 48.

Plummer, M. (2003). Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*, **124**, 1 – 106.

Weiss, R.E. (2005). *Modeling longitudinal data*. Springer Science & Business Media.

Comparison of Experimental Designs for Normal and Gamma distributions

Víctor Casero–Alonso[1], Sergio Pozuelo–Campos[1], Mariano Amo–Salas[1]

[1] Deparment of Mathematics, Castilla–La Mancha University, Spain

E-mail for correspondence: victormanuel.casero@uclm.es

Abstract: The aim of this work is to show the effect of misspecification in the probability distribution on optimal design. A generalized Fisher information matrix is obtained using the elemental information matrix, which includes information on the probability distribution of the response variable. Relevant theoretical results were obtained, for different regression models, comparing heteroscedastic gamma and normal distributions. Finally a practical case which considers a 4-parameter Hill dose-response model is used.

Keywords: Elemental Information Matrix; Approximate design; D–optimatilty; Hill dose-response model.

1 Introduction

It is a common assumption in the context of Optimal Experimental Design that the response variable follows a homoscedastic normal distribution. There are, however, other studies that assume different probability distributions based on prior experience or additional information. Nonetheless, the available references that set out a general framework for this theory for any probability distribution of the response variable are very few (Atkinson et al., 2014), and the effect on the optimal design of the probability distribution under consideration has not been previously studied. This work analyzes that effect.
The model of interest for the experimenter can be expressed generally as

$$y = g^{-1}(\eta(x;\theta)) + \varepsilon, \tag{1}$$

where y is the response variable that is assumed to follow a probability distribution with density function $d(y;\rho)$, where ρ is the vector of param-

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

eters of the distribution, $\eta(x;\theta)$ is the regression function (not necessarily linear in the parameters), x is the explicative variable, and θ the vector of unknown parameters that must be estimated. Finally, g is the link function relating the mathematical expectation of the probability distribution to the regression function, and ε is the error term.

2 Optimal experimental design

An approximate design is a probability measure ξ over the design space $\mathcal{X}$:

$$\xi = \begin{Bmatrix} x_1 & \dots & x_q \\ w_1 & \dots & w_q \end{Bmatrix} \in \Xi,\ x_i \in \mathcal{X},\ \sum_{i=1}^{q} w_i = 1,$$

where $\xi(x_i) = w_i$ and Ξ represents the set of all approximate designs.
The elemental information matrix (EIM), introduced by Atkinson et al. (2014), is defined as

$$\nu(\eta(x;\theta)) = -E\left[\frac{\partial^2 \log d(y;\eta(x;\theta))}{\partial \eta(x;\theta)^2}\right]. \tag{2}$$

It gathers information about the probability distribution given by $d(y;\rho)$. The relation between the parameters, ρ, of the probability distribution and the model $\eta(x;\theta)$ is determined by the linking function, g, seen in (1).
The single-point information matrix in $x \in \mathcal{X}$ is given by

$$I(x;\theta) = -E\left[\frac{\partial^2 \log d(y;\eta(x;\theta))}{\partial \theta_i \partial \theta_j}\right] = \nu(\eta(x;\theta)) f^T(x;\theta) f(x;\theta),$$

where $f^T(x;\theta) = \partial\eta(x;\theta)/\partial\theta$. And the Fisher information matrix (FIM) is defined for the approximate design with probability measure ξ as

$$M(\xi;\theta) = \int_{\mathcal{X}} I(x;\theta)\xi(x)dx.$$

By definition, the inverse of the FIM is asymptotically proportional to the variance and co-variance matrix of estimators of the parameters of the model θ to be estimated.
Optimization criteria play an important role in the theory of Optimal Experimental Design, as they allow functions of the FIM to be determined that optimize this matrix in different ways. This study uses the D-optimization criterion, whose aim is to minimize the volume of the confidence ellipsoid of the estimators of the model parameters, $\hat{\theta}$. This criterion is given by

$$\Phi_D(M(\xi;\theta)) = \log|M^{-1}(\xi;\theta)|.$$

The D-efficiency of a design allows the goodness of any design ξ to be compared to the D-optimal ξ^* design,

$$\text{eff}_D(\xi|\xi^*) = \left(\frac{|M(\xi;\theta)|}{|M(\xi^*;\theta)|}\right)^{1/m}. \tag{3}$$

3 Theoretical results

This paper looks at two probability distributions of the response variable: Gamma and Normal. For the Gamma distribution, $\mathrm{Var}[y] = \mathrm{E}[y]^2/\alpha$. Therefore a heteroscedastic Normal distribution with a variance structure which allows it to be compared to the Gamma distributions is considered:

$$\mathrm{Var}[y] = k\mathrm{E}[y]^{2r}, \tag{4}$$

where $k \in \mathbb{R}^+$ and $r \in \mathbb{R}$ are constants and $\mathrm{E}[y] = \eta(x;\theta)$. Thus, if the parameter α of the Gamma distribution, $\Gamma(\alpha, \beta)$, is constant, a similar variance structure to the heteroscedastic Normal distribution is achieved with $k = 1/\alpha$ and $r = 1$. On the other hand, the case $r = 0$ corresponds to the homoscedastic Normal distribution. Using (2) the EIM of the heteroscedastic Normal distribution with variance given by (4) is

$$\nu(\eta(x;\theta); r, k) = \frac{2r^2}{\eta(x;\theta)^2} + \frac{1}{k\eta(x;\theta)^{2r}}.$$

Theorem 1. Let $\eta(x;\theta) > 0$ be the function of some regression model, for some optimization criterion Φ based on the FIM, then the Φ-optimal designs for the heteroscedastic Normal distribution with $r = 1$ in the variance defined in (4) and for the Gamma distribution with constant α coincide. Also, the Φ-optimal design obtained is independent of α and k.

Theorem 2. Let $\eta(x;\theta) = \theta_0 + \theta_1 x + \theta_2 x^2 > 0$ be the function of a quadratic linear regression model, where x is defined as a design space $\mathcal{X} = [x_l, x_u]$. If the response variable follows a Gamma distribution with constant parameter α, the D-optimal design is $\xi_\Gamma^* = \begin{Bmatrix} x_l & x_2 & x_u \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}$, where $x_2 \in (x_l, x_u)$ is a solution of the quadratic equation

$$(\theta_1 + \theta_2(x_l + x_u))x_2^2 - (2\theta_2 x_l x_u - 2\theta_0)x_2 - (\theta_0(x_l + x_u) + \theta_1 x_l x_u) = 0. \tag{5}$$

4 Practical application of the 4-parameter Hill model

The well-known Hill model widely used in the literature to describe dependence between the concentration of a substance and a variety of biochemical, physiological or pharmacological responses. In the context of Optimal Experimental Design, this model was studied by several authors. Khinkis et al. (2003) look at the 4-parameter Hill model, where the response variable is the effect of a number of drugs which inhibit the growth of tumor cells, without completely eliminating them. These authors calculate D-optimal designs for the different types of drug, assuming that the response variable follows a normal heteroscedastic distribution with the variance structure given by (4). For brevity we omit here detailed results. But our analysis shows a different behavior of the efficiency of the designs obtained by

assuming the heteroscedastic normal distribution, when the relationship between the mean and the variance changes, that is when r varies, with respect to the designs obtained for the Gamma distribution. These results allow to identify those cases in which it is important to give special attention to the assumed probability distribution.

References

Atkinson, A.C., Fedorov, V.V., Herzberg, A.M. and Zang R. (2014). Elemental information matrices and optimal experimental design for generalized regression models. *Journal of Statistical Planning and Inference*, **144**, 81 – 91.

Khinkis LA, Levasseur L, Faessel H, Greco WR. (2003) Optimal Design for Estimating Parameters of the 4-Parameter Hill Model. *Nonlinearity in Biology, Toxicology and Medicine* **1**: 363 – 377.

Sensitivity analysis approaches to investigate uncertainty in process-based models, with application to aquaculture

Michael Currie[1], Claire Miller[1], Marian Scott[1], Alan Hills[2]

[1] School of Mathematics & Statistics, University of Glasgow, Scotland
[2] Scottish Environment Protection Agency, Scotland

E-mail for correspondence: m.currie.1@research.gla.ac.uk

Abstract: Sensitivity and uncertainty analyses are effective tools for assessing uncertainty around parameter estimation for complex computer models, and hence increase confidence in model predictions. NewDEPOMOD is a particle tracking model used for monitoring the environmental impacts of aquaculture, and will be used as an application to extend sensitivity analysis methods to consider models with a multivariate response.

Keywords: Sensitivity Analysis; Shape Analysis; Aquaculture.

1 Introduction & Background

There remain environmental challenges which can only acccurately be assessed by process-based modelling. An example of this is monitoring the environmental impacts of aquaculture, where the difficulty and cost of collecting data over large areas make modelling the more effective approach. Such modelling approaches are computationally intensive and do not account for uncertainty. Therefore, sensitivity and uncertainty analyses of these models provide approaches to quantify uncertainty in model responses and attribute them to variations in the model parameters. NewDEPOMOD is a development of DEPOMOD (Cromey *et al.* 2002) that was created to estimate and predict the transportation of waste particles from fish farm cages to the seabed. NewDEPOMOD produces a number of scalar outputs as well as a map of the impacted area. Scalar outputs such as the Total Area Impacted, 99th Percentile and Mass balance will be considered as a starting point for the sensitivity analysis, before extending this to consider the shape of the impact as the response.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Sensitivity Analysis

A sensitivity analysis can be used to address uncertainties within a model to increase confidence in the predictions (Saltelli *et al.* 2000). Given a process-based model that maps the n inputs, $\mathbf{x} = [x_1, \ldots, x_n]$, to the output, $\mathbf{y}$, using the function, f,

$$\mathbf{y} = f(\mathbf{x}) = f(x_1, \ldots, x_n).$$

A sensitivity analysis considers how variations in the output, $\mathbf{y}$ can be associated with variations in the inputs, $\mathbf{x} = [x_1, \ldots, x_n]$. Saltelli *et al.* (2000) described a typical sensitivity analysis workflow: 1) Determine the questions relating to the model that should be answered and identify the inputs required, 2) Establish suitable ranges of variation for each input, 3) Identify an appropriate design to generate the input matrix, 4) Complete model runs to create the required outputs, and 5) Analyse the effect of each input on the output. Traditionally, a sensitivity analysis is applied to a model with a scalar output, but in this work we extend this to develop a sensitvity analysis approach that uses area and shape as the response which will be illustrated using output from NewDEPOMOD.

2.1 Sensitivity Analysis for Scalar Outputs

The scalar outputs for NewDEPOMOD, mentioned previously, were considered as the outputs of the sensitivity analysis. In collaboriation with the Scottish Environment Agency (SEPA), a set of inputs were identified as being of most importance, which included Critical Shear Stress for Erosion and Settling Velocity of Faeces. Suitable ranges for the inputs were established using the literature, where possible, and the experience and knowledge of SEPA in cases where no literature was available. A correlated Latin Hypercube Sampling (LHS) was used to account for the relationships between inputs and capture the sample space effectively. It relies on a restricted pairing procedure (Iman & Conover 1982), where a target correlation matrix, $\mathbf{C}^*$, is identified at the outset. Following this, an initial LHS, $\mathbf{L}$, is calculated with sample correlation, $\mathbf{T}$. A Cholesky Decomposition of $\mathbf{T}$ and using other variance reduction techniques, allows a matrix, $\mathbf{S}$, to be found such that the correlated LHS is given as follows:

$$\mathbf{L}^*_{\mathbf{B}} = \mathbf{L}\mathbf{S}^{\mathrm{T}}$$

where the correlated LHS, $\mathbf{L}^*_{\mathbf{B}}$ has a sample correlation matrix $\mathbf{M}_{\mathbf{B}}$, approximately equal to $\mathbf{C}^*$. The input matrix allowed 10,000 model runs to be completed in order to calculate the scalar summaries and determine the impact of uncertainties in the inputs. Random forests were used as a ranking method as they are able to deal with non-linear relationships, interactions

TABLE 1. Top 3 ranked inputs using Total Area Impacted as the output.

Inputs	Importance Value
Settling Velocity of Faeces	127.24
Critical Shear Stress for Erosion	75.99
Rate of Erosion	50.81

between inputs and also the interpretability of the results. Table 1 shows the 3 highest ranked inputs for the scalar output, Total Area Impacted. The random forest model identified Settling Velocity of Faeces as having the biggest impact on the Total Area Impacted, which was expected as this determines the time particles spend settling from the cages to the seabed.

2.2 Sensitivity Analysis of the Shape of the Impact

As NewDEPOMOD produces a map of the predicted shape and size of the impacted area on the seabed, it was important to extend the traditional sensitivity analysis of scalar outputs and consider the influence of uncertainty in the inputs on the predicted shape and size. A landmark approach (Dryden & Mardia 2016) was used to identify the main shape of the impact (example seen in Figure 1), by considering transects from the farm.

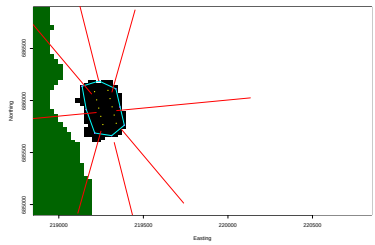


FIGURE 1. Plot illustrating the landmarks calculated to analyse shape.

Landmarks were calculated for each map of the predicted impact using this approach, before using a Generalised Procrustes Analysis (GPA) approach to identify variations in the shapes. GPA is defined as the translation, rescaling and rotation of the shape configurations $(X_1, X_2, \ldots, X_n)$ relative to each other, to minimize a total sum of squares (Dryden & Mardia 2016):

$$G(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} \parallel (\beta_i X_i \Gamma_i + 1_k \gamma_i^{\mathrm{T}}) - \mu \parallel^2$$

with respect to $\beta_i, \Gamma_i, \gamma_i$, for $i = 1, \ldots, n$ and μ, subject to an overall size constraint that is chosen. $\beta_i > 0$ refers to a scale parameter, Γ_i is a

rotation matrix, γ_i is a location vector and μ is the population mean shape. A Principal Components Analysis was then applied to the landmarks data to identify the areas of variability in the shapes.

TABLE 2. Table of the Principal Component percentages.

Principal Component	% of Variability Captured
PC 1	59.0%
PC 2	22.0%
PC 3	7.5%
Total	88.5%

Table 2 shows the variability described by the first 3 principal components (PCs). Settling Velocity of Sediment and Release Height were identified, with the 3 inputs from Table 1, as having an influence on the variations described by the first 3 PCs.

3 Conclusion

Traditional sensitivity analysis methods were extended to multivariate response models and applied to NewDEPOMOD to identify parameters that influenced the variation in the shape of the impacted area on the seabed. Further work aims to develop a spatio-temporal emulator of NewDEPOMOD to allow predictions to be made without the computational cost.

Acknowledgments: Special thanks to EPSRC (Award Ref: 1953182) and SEPA for funding the PhD and Andrew Berkeley (formerly SEPA) for his support in the completion of this work.

References

Cromey, C.J., Nickell, T.D., and Black, K.D. (2002). DEPOMOD - modelling the deposition and biological effects of waste solids from marine cage farms. *Aquaculture*, **214**, 211 – 239.

Dryden, I. and Mardia, K. (2016). *Statistical shape analysis: with applications in R, 2nd Edition* John Wiley & Sons Ltd.

Iman, R. and Conover, W. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, **11**, 311 – 334.

Saltelli, A. and Chan, K. and Scott, M. (2000). *Sensitivity Analysis*. John Wiley & Sons Ltd.

Spatial seismic point pattern analysis with Integrated Nested Laplace Approximation

Nicoletta D'Angelo[1], Antonino Abbruzzo[1], Giada Adelfio[1]

[1] Department of Economics, Business and Statistics, University of Palermo, Italy

E-mail for correspondence: nicoletta.dangelo@unipa.it

Abstract: This paper proposes the use of Integrated Nested Laplace Approximation (Rue et al., 2009) to describe the spatial displacement of earthquake data. Specifying a hiechical structure of the data and parameters, an inhomogeneuos Log-Gaussian Cox Processes model is applied for describing seismic events occurred in Greece, an area of seismic hazard. In this way, the dependence of the spatial point process on external covariates can be taken into account, as well as the interaction among points, through the estimation of the parameters of the covariance of the Gaussian Random Field, with a computationally efficient approach.

Keywords: Integrated Nested Laplace Approximation; Stochastic Partial Differential Equation; Spatial Point Process; Cox process; Seismology.

1 Introduction

Usually, Bayesian inference for spatial and spatio-temporal data refers to Markov Chain Monte Carlo algorithm. Unfortunately, for such models, this can be computationally demanding, given the complexity of the spatio-temporal models, the dataset and the parametric space dimensions. The Integrated Nested Laplace Approximation (INLA) (Rue *et al.*, 2009) approach has been developed as a computationally efficient alternative to MCMC. Furthermore, INLA can be combined with the Stochastic Partial Differential Equation (SPDE) approach proposed by Lindgren *et al.* (2011) in order to implement spatial and spatio-temporal models for point-reference data. In this paper, after a brief overview of spatial point pattern analysis with INLA in section 2, we provide an application on spatial seismic data, fitting an inhomogeneous Log-Gaussian Cox Process in section 3, followed by the Conclusions.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Spatial point pattern analysis with INLA

INLA is designed for *Latent Gaussian Models*, a very wide and flexible class of models ranging from (generalized) linear mixed to spatial and spatio-temporal models. In these models the distribution of the response variable $\mathbf{y}=(y_1,\ldots,y_n)$ is assumed to belong to the exponential family, in which the linear predictor η_i can include terms on covariates and different types of random effects in an additive way. The vector of all these latent effects is denoted by $\boldsymbol{\theta}$ and, its distribution follows a Gaussian Markov Random Field (GMRF), with zero mean and precision matrix $\mathrm{Q}(\boldsymbol{\psi_2})$, where $\boldsymbol{\psi_2}$ is the vector of hyperparameters. In addition, the distribution of $\mathbf{y}$ depends on some vector of hyperparameters $\boldsymbol{\psi_1}$, to which are assigned priors, not necessarily Gaussian. The vector of the hyperparameters is denoted by $\boldsymbol{\psi}=(\boldsymbol{\psi_1},\boldsymbol{\psi_2})$.
The main goal is to obtain the marginal distributions for the elements of the latent fields $p(\boldsymbol{\theta}|\boldsymbol{y})$ and the hyperparameters $p(\boldsymbol{\psi}|\boldsymbol{y})$, and use these to compute posterior summary statistics. This is achieved by exploiting the computational properties of the GMRF and the Laplace approximation for multidimensional integration, assuming the observed variables $\mathbf{y}$ to be independent given $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. More details can be found in Rue *et al.* (2009) and Blangiardo *et al.* (2013).
In the case of geostatistical data, $\boldsymbol{\theta}$ is assumed to follow a multivariate Normal distribution with mean μ and spatially structured covariance matrix Σ, whose generic elements is $\Sigma_{ij} = Cov(\theta_i,\theta_j) = \sigma_C^2 C(\Delta_{ij})$ where $C(\Delta_{ij}) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\kappa\Delta_{ij})^{\nu}K_{\nu}(\kappa\Delta_{ij})$ is the isotropic Matérn spatial covariance function (Cressie, 1992) depending on the Euclidean distance between the locations $\Delta_{ij} = \|s_i - s_j\|$. The parameter K_ν denotes the modified Bessel function of second kind and order $\nu > 0$, which measures the degree of smoothness of the process and is usually kept fixed. Conversely, $\kappa > 0$ is a scaling parameter related to the range r, through $r = \frac{\sqrt{8\nu}}{\kappa}$, with r corresponding to the distance at which the spatial correlation is close to 0.1 for each ν. Other models are possible for the spatial covariance function but in this paper we only focus on the Matrn model, since it is required for the SPDE approach that we consider in the application. Indeed, when dealing with point-reference data, this approach is particularly computationally effective, as it consists in representing a continuous spatial process (e.g. a GF) with the Matérn covariance as a discretely indexed spatial random process (e.g. a GMRF). We refer to Lindgren *et al.* (2011) for a complete description.

3 Spatial Log-Guassian Cox Process for seismic data

In this section, an application to a spatial dataset is provided, specifying a Log-Gaussian Cox Process with inhomogeneous intensity $\lambda(\boldsymbol{s})$. We consider

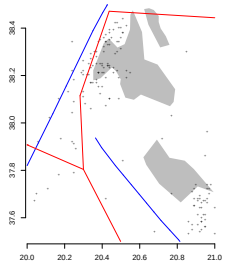


FIGURE 1. Earthquakes occurred in Ithaki, Kefalonia and Zakynthos between 2005 and 2014. The plate boundary is in red while the faults are blue.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
r	0.61	0.15	0.38	0.59	0.97	0.55
σ^2	1.60	0.30	1.11	1.57	2.28	1.50
β_0	2.60	1.09	0.23	2.66	4.58	2.79
β_1	0.09	3.38	-6.52	0.06	6.81	0.03
β_2	5.88	2.85	0.52	5.79	11.72	5.62
β_3	-9.56	11.65	-32.76	-9.47	13.06	-9.27

TABLE 1. Summary statistics of the distributions of the parameters and hyperparameters of the model (1)

the SPDE model, as developed in Simpson *et al.* (2016). The 149 analysed seismic events are mainly clustered in the Western area of Kefalonia island and South to the Zakynthos island (Figure 1). The spatial covariates considered in the analysis are the Distance from the faults (D_f) and the Distance from the plate boundary (D_{pb}). The proposed model is specified as follows:

$$\begin{gathered} y_i|\lambda \sim \exp^{|D|-\Lambda} \prod \lambda(s_i) \\ \log \lambda(s_i) = \beta_0 + \beta_1 D_f + \beta_2 D_{pb} + \beta_3 D_f D_{pb} + u(s_i) \\ u|\sigma^2, r \sim GRF(0, \Sigma(\sigma^2, r)) \end{gathered} \tag{1}$$

where latent field is represented by $\boldsymbol{\theta} = [\boldsymbol{\beta}, \boldsymbol{u}]^T$ and the hyperparameters are $\boldsymbol{\psi} = [\sigma^2, r]^T$, as ν is set equal to 1. The SPDE approach for point pattern analysis defines the model at the nodes of the mesh, so this is built on the entire domain extent, with the largest allowed triangle edge length equal to 0.1 for the interior edges and 0.4 for the exterior. PC-priors, derived as in Fuglstad *et al.* (2018), are considered for r and σ^2. Summary statistics of the parameters and hyperparameters of the model (1) are reported in table 1. We notice how the negative sign of the interaction term parameter β_3 suggests what we expected, that is, moving away from a seismic source the probability of the occurrence of an earthquake decreases.

Furthermore, such values of r and σ prove that $\lambda(\boldsymbol{s})$ changes rapidly over the study window, around its mean, that is to say, the clustered structure of the point pattern is correctly identified by the model parameters. These results are close to those obtained fitting a LGCP model with the same inhomogeneous intensity, through the local Palm likelihood maximization, in D'Angelo *et al.* (2020). Nevertheless, in the latter approach, as the parameters are allowed to vary spatially, it is also possible to identify the most inhomogeneous areas.

4 Conclusions

In this paper we briefly explored the potentialities of INLA, in modelling spatial seismic data, proposing an inhomogeneous LGCP model to describe a Greek seismic spatial point pattern, and obtaining similar results as the local frequentist approach. As combining the INLA and SPDE approaches it is possible to implement both spatial and spatio-temporal models for point-reference data, future analyses might take into account the temporal component of the analysed process, and therefore the spatio-temporal structure that typically characterizes the complex seismic phenomenon.

References

Blangiardo, M., Cameletti, M., Baio, G. and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology*, 4:33-49.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, 4(5):613-617

D'Angelo, N., Siino, M., D'Alessandro, A. and Adelfio, G. (2020). Local spatial point processes for seismic data. *Submitted*.

Fuglstad, G. A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114(525), 445-452

Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423-498.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319-392.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H. and Rue, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1):49:70.

Nonstationary, Nonparametric, Nonseparable Bayesian Spatio-Temporal Modeling Using Kernel Convolution of Order Based Dependent Dirichlet Process

Moumita Das[1], Sourabh Bhattacharya[2]

[1] Basque Center for Applied Mathematics,Bilbao, Spain
[2] Indian Statistical Institute, Kolkata, India

E-mail for correspondence: mdas@bcamath.org

Abstract: In this work, using kernel convolution of order based dependent Dirichlet process (Griffin and Steel (2006)) we construct a nonstationary, nonseparable, nonparametric space-time process, which, as we show, satisfies desirable properties, and includes the stationary, separable, parametric processes as special cases.

Keywords: Nonstationary; Nonseparable; Order Based Dependent Dirichlet Process.

1 Introduction

Recent years have witnessed considerable amount of research on spatial and spatio-temporal modeling. It is common practice to assume that the underlying spatial or spatio-temporal process is stationary and isotropic Gaussian process, as it facilitates prediction. In particular, the geostatistical method of kriging assumes a Gaussian process structure for the unknown spatial or spatio-temporal field and focuses on calculating the optimal linear predictor of the field. When performing kriging, researchers generally assume a stationary, often isotropic, covariance function. The covariance of responses at any two locations is assumed to be a function of the separation vector or of the distance between locations, but not a function of the actual locations. The standard kriging approach allows one to flexibly estimate a smooth spatial field, with no pre-specified parametric form. However, these approaches have several drawbacks. The most important is

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

that the true covariance structure may not be stationary. This is because there may be local influences affecting the correlation structure of the random process. For instance, orographic effects influence the atmospheric transport of pollutants, and result in a correlation structure that depends on different spatial locations. Griffin and Steel, 2006 (henceforth, GS) proposed the novel order-based dependent Dirichlet processes (ODDP). They introduced a framework for nonparametric modeling with dependence on continuous covariates. Dependence is induced through relevant weights utilizing similarities in the covariate information.a simple analytical expression for the correlation function of the distributions, which ensures that if two points are similar in the covariate space they will get higher correlation compared to the points that are not. Furthermore when the distance between two points is large enough in the covariate space, the correlation approaches zero. In spatial/spatio-temporal context, it translates into the fact that when two observations are widely separated in space/space-time, the model based correlations tend to zero. But the ODDP process suffers from the limitation of being stationary. Preserving all the desirable properties of the correlation function of ODDP, we attempt to incorporate further flexibility in our spatial/temporal/spatio-temporal model in terms of nonstationarity and nonseparability through our proposed kernel convolution based methodology.

2 Kernel Convolution of ODDP

We consider the following model for the data $\boldsymbol{Y} = \{y_1, \ldots, y_n\}$ at locations/times $\{\boldsymbol{x}_i = (\boldsymbol{s}_i', t_i)';\ i = 1, \ldots, n\}$:

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \tag{1}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, for unknown σ^2. We represent the spatio-temporal process $f(\boldsymbol{x})$ as a convolution of ODDP $G_{\boldsymbol{x}}$ with a smoothing kernel $K(\boldsymbol{x}, \cdot)$:

$$f(\boldsymbol{x}) = \int K(\boldsymbol{x}, \boldsymbol{\theta}) dG_{\boldsymbol{x}}(\boldsymbol{\theta}) = \sum_{i=1}^{\infty} K(\boldsymbol{x}, \boldsymbol{\theta}_{\pi_i(\boldsymbol{x})}) p_i(\boldsymbol{x})\ \forall \boldsymbol{x} \in D \subseteq \mathbb{R}^d, \tag{2}$$

$d\ (\geq 1)$ being the dimension of $\boldsymbol{x}$.

Before deriving the covariance structure of $f(\cdot)$, we define the necessary notation following GS. Let

$$T(\boldsymbol{x}_1, \boldsymbol{x}_2) = \{k :\ \text{there exists } i, j \text{ such that } \pi_i(\boldsymbol{x}_1 = \pi_j(\boldsymbol{x}_2) = k\}.$$

For $k \in T(\boldsymbol{x}_1, \boldsymbol{x}_2)$, we further define $A_{lk} = \{\pi_j(\boldsymbol{x}_1) : j < i \text{ where } \pi_i(\boldsymbol{x}_1) = k\}$, $S_k = A_{1k} \cap A_{2k}$ and $S_k' = A_{1k} \cup A_{2k} - S_k$. Then, the following theorem, provides an expression for the covariance structure of $f(\cdot)$.

Theorem 1 *If* $\int |K(\boldsymbol{x}, \boldsymbol{\theta})| dG_0(\boldsymbol{\theta}) < \infty$ *and* $\int |K(\boldsymbol{x}_1, \boldsymbol{\theta}) K(\boldsymbol{x}_2, \boldsymbol{\theta})| dG_0(\boldsymbol{\theta}) < \infty$, *then for a fixed ordering at* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$,

$$Cov(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)) = Cov_{G_0}(K(\boldsymbol{x}_1, \boldsymbol{\theta}), K(\boldsymbol{x}_1, \boldsymbol{\theta})) \times \frac{2}{(\alpha+1)(\alpha+2)} \sum_{k \in T(\boldsymbol{x}_1, \boldsymbol{x}_2)} \left(\frac{\alpha}{\alpha+2}\right)^{\#S_k} \left(\frac{\alpha}{\alpha+1}\right)^{\#S'_k}, \tag{3}$$

where

$$Cov_{G_0}(K(\boldsymbol{x}_1, \boldsymbol{\theta}), K(\boldsymbol{x}_2, \boldsymbol{\theta})) = \int K(\boldsymbol{x}_1, \boldsymbol{\theta}) K(\boldsymbol{x}_2, \boldsymbol{\theta}) dG_0(\boldsymbol{\theta}) - E_{G_0}(K(\boldsymbol{x}_1, \boldsymbol{\theta}) E_{G_0}(K(\boldsymbol{x}_2, \boldsymbol{\theta}). \tag{4}$$

3 Computation of the Model

Since our model entails an infinite random series, for Bayesian model fitting purpose we must either truncate the series or more appropriately consider a random number of summands, which renders the model dimension a random variable. We attack the variable dimensionality problem using Transdimensional Transformation based Markov Chain Monte Carlo algorithm.

4 Application

We have used our method, to analyse two separate real data sets, one spatial real data and one spatio-temporal data. In spatial real data we consider annual indexes of ozone values for 76 locations in the US. For spatio-temporal data analysis, airborne particulate matter (PM) for 180 time points and 50 locations. We have ensured the nonstationarity property of the data using a newly developed test for detecting stationarity in spatial/spatio-temporal data.

5 Conclusion

Although for the current work we restricted ourselves to spatio-temporal applications only, our model is readily applicable in the functional data context. In fact, in the context of nonparametric function estimation, a new class of prior distributions can be introduced through our proposed model.

References

Griffin J. E and Steel M.F.J (1994). *Order Based Dependent Dirichlet Processes*. Journal of American Statistical Association, 101, 179 – 194.

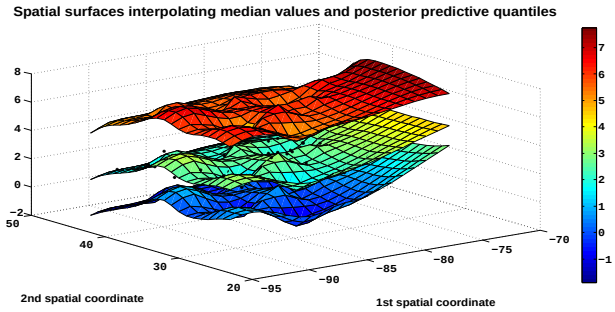


FIGURE 1. Real spatial data analysis: Surface plot of the posterior medians (middle) along with the lower and the upper 95% credible intervals associated with the leave-one-out posterior predictive densities The observed data points are indicated by *.

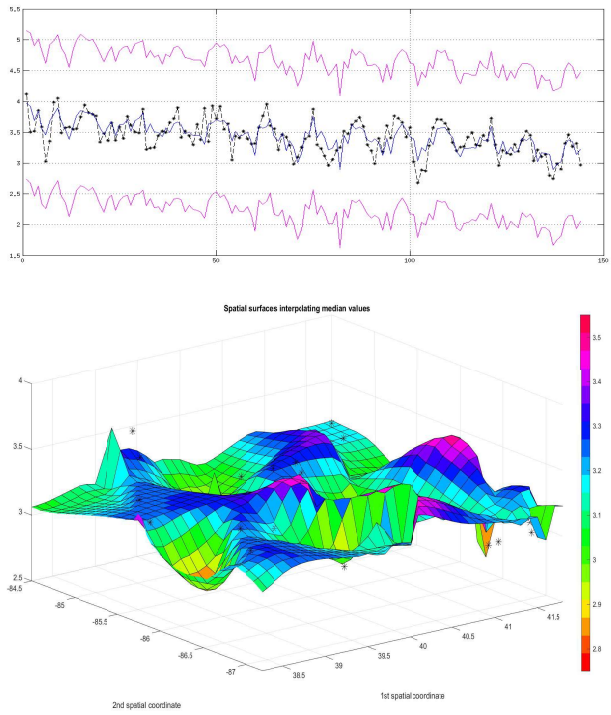


FIGURE 2. Real spatio-temporal data analysis: The top panel shows Posterior predictive distributions summarized by the median (middle line) and the 95% credible intervals as a function of t for one randomly chosen spatial locations. The bottom panel exhibits the surface plot of posterior median values at 50 locations, averaged over all month specific predictions from 1988-2002. The observed data points are denoted by *.

I-optimal designs for Antoines equation: A genetic algorithm approach

Carlos de la Calle-Arroyo[1], Miguel Ángel González Fernández[2], Jesús López-Fidalgo[3], Licesio J. Rodríguez-Aragón[1]

[1] Instituto de Matemática Aplicada a la Ciencia y la Ingeniería, E. I. Industrial y Aeroespacial de Toledo,Universidad de Castilla-La Mancha, Spain
[2] Departamento de Informtica, Universidad de Oviedo, Spain
[3] Instituto de Cultura y Sociedad, Universidad de Navarra, Spain

E-mail for correspondence: carlos.callearroyo@uclm.es

Abstract: In the distillation processes it is very important to know precisely the relationship between temperature and vapor pressure. The vapor pressures not only depend on the temperature but vary enormously for different substances.
The study of optimal designs for the estimation of the parameters of Antoine equation, according to the I-optimality criterion is shown. It is particularly interesting for this model due to the importance of prediction on boundary regions of the space of the design, which usually correspond to the proximity of state change points.
Genetic algorithms are one of the several nature-inspired algorithms, mainly used for the calculation of optimal solutions to problems that are hard to solve through direct algorithms. A genetic algorithm that find optimizes the designs presented in this work has been developed.

Keywords: Optimal Design; I-optimality; Heuristics; Genetic Algorithms

1 Introduction

The Antoine's equation is a hyperbolic equation, from a class of semi-empirical correlations describing the relation between vapor pressure, P, and temperature, T, for pure components (Wisniak, 2001). The statistical model can be written as:

$$P = \eta(T, \theta) + \varepsilon = 10^{A - \frac{B}{C+T}} + \varepsilon; \quad var(P) = \sigma^2; \quad T \in [T_{min}, T_{max}].$$

Along this work, optimal designs will be obtained for the particular case of water in the range of temperatures $\mathcal{X} = (1°C, 100°C)$. Ini-

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

tial best guesses of the parameters A, B and C, will be needed: $\theta_0 = (8.07131, 1730.63, 233.426)^t$ (Dortmund Databank, 2020).
For this work approximate designs will be used. Approximate designs consist on a set of points, T_i, and the proportion of observations, ω_i, that should be taken on each of the points which don't have to be necessarily a ratio of an integer. It is obvious that the weights must be positive, and verify $\sum_i \omega_i = 1$. An approximate design, ξ, can also be seen as a probability measure over the design space,

$$\xi = \begin{pmatrix} T_1 & T_2 & ... & T_m \\ \omega_1 & \omega_2 & ... & \omega_m \end{pmatrix}.$$

The Information Matrix for this design is then calculated as:

$$M(\xi) = \sum_{i=1}^{m} \omega_i \cdot f(T_i) f^t(T_i); \quad \text{where } f(T_i) = \frac{\partial \eta(T, \theta)}{\partial \theta}$$

and, if not singular, its inverse is proportional to the variance-covariance matrix of the unknown parameters (Fedorov and Leonov, 2014).

2 I-optimization

There is a wide array of criteria on which designs can be evaluated, depending on the desired properties of the design. In this section the criterion of I-optimality, in which this work focuses, is explained.
I-optimality seeks designs that minimize the average variance of prediction over a region of interest, $\mathcal{R}$, which could be outside the space design. It takes the first expression, which can be then rewritten as:

$$I(\xi) = \frac{\int_{\mathcal{R}} f^t(T) M^{-1} f(T) dT}{\int_{\mathcal{R}} dT} = \frac{Tr[M^{-1}B]}{\int_{\mathcal{R}} dT}, \text{ with } B = \int_{\mathcal{R}} f(T) f^t(T) dT.$$

Here, only $M(\xi)$ depends on the design, with B being the moment matrix of the model over the interest region, $\mathcal{R}$.
I-optimality is a linear function of the elements of $M^{-1}(\xi)$ and the Equivalence Theorem states that a design $\xi^\star$ is I-optimal if:

$$f^t(T) M^{-1}(\xi^\star) B M^{-1}(\xi^\star) f(T) - Tr[M^{-1}(\xi^\star) B] \leq 0 \quad \forall T \in \mathcal{X}. \quad (1)$$

The equality is reached on the support points of the design (Goos *et al.*, 2016).

2.1 Algorithms for I-optimal designs

The problem of finding optimal designs for non-linear models is often hard or untractable to take on analitically. With that consideration, much effort has been put on numerical methods to find optimal designs.

For this work, a version of the Wynn-Fedorov algorithm and a genetic algorithm have been implemented.
The Wynn-Fedorov algorithm has been adapted for I-optimality, with a few commonly added heuristics. This algorithm consist on sequentially choosing points to add to the design that maximizes the expression of the Equivalence Theorem (1), with a certain weight decreasing on each iteration (Goos *et al.*, 2016).
Since this algorithm usually has a slow convergence rate, the use of a genetic algorithm was considered. A genetic algorithm is a search metaheuristic that is inspired by the theory of natural evolution. This kind of algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring for the next generation.
The genetic algorithm starts by generating and evaluating (according to the objective function, $I(\xi)$) an initial population of random designs. The upper bound for the number of points can be set to 7, due to Caratheodorys Theorem, which states that an optimal design can be found with at most $k(k+1)/2+1$, with k the number of unknown parameters. In each generation, the population is randomly combined in pairs, and a crossover operator is applied to each pair, in order to generate two offspring solutions by mixing random points and weights. Then, a series of small random mutations are carried on in each offspring, iteratively improving them by modifying some points and weights. Then, for the next population, the best two individuals from each pair of parents and their respective two offspring solutions are chosen. The stopping condition of the genetic algorithm is a certain number of generations without improving the best design found so far, or when the optimal design is reached (via the Equivalence Theorem). During the whole process, the best individual is stored, and its criterion function is stored as a benchmark. If after a certain number of iterations there is no improvement, the algorithm stops, and the optimality of the design is verified via the Equivalence Theorem.

3 Results

In I-optimality the region of interest for the prediction must be chosen, and a certain probability distribution over that region assumed. For this particular model, the chosen region of interest is around the change of state temperature, $T_{max} = 100°C$. As for the probability distribution, two different options have been tested: the uniform distribution and a symmetric triangular distribution, which emphasises the interest on the central part of the region.
For the uniform distribution, the designs of Table 1 have been calculated. While for the triangular distribution, the designs of Table 2 have been calculated.

TABLE 1. I-optimal design for uniform distribution

$\mathcal{R}$	x_1	ω_1	x_2	ω_2	x_3	ω_3
98-102	34.8217	0.0448625	86.5664	0.134595	100.	0.820542
90-110	33.6143	0.170485	84.6011	0.333001	100.	0.496514
80-120	33.2688	0.262212	83.9214	0.377867	100.	0.359921
60-140	33.0619	0.351941	83.8008	0.382857	100.	0.265203
40-160	33.0224	0.397804	83.8096	0.374876	100.	0.227319

TABLE 2. I-optimal design for triangular distribution

$\mathcal{R}$	x_1	ω_1	x_2	ω_2	x_3	ω_3
98-102	35.021	0.0328981	86.7807	0.101693	100.	0.865409
90-110	33.9156	0.136586	85.0865	0.295118	100.	0.568297
80-120	33.4647	0.228313	84.1820	0.362912	100.	0.408775
60-140	33.148	0.329009	83.8345	0.382853	100.	0.288138
40-160	33.0636	0.380855	83.8119	0.377867	100.	0.241278

These designs are useful as a starting point for choosing designing industry-level experiments. Moreover, they serve as a benchmark, as the efficiency of industry design can be tested against theirs.

Acknowledgments: This work was sponsored by Ministerio de Economa y Competitividad MTM2016-80539-C2-1-R, by JCCM and FEDER SBPLY/17/180501/000380, by the Spanish Government TIN2016-79190-R and by the Principality of Asturias FC-GRUPIN-IDI/2018/000176.

References

Dortmund Data Bank (2020). www.ddbst.com

Fedorov, V. V. and Leonov, S. L. (2014). *Optimal Design for Nonlinear Response Models*. Boca Raton: CRC Press.

Goos, P., Jones, B. and Syafitri, U. (2016) I-Optimal Design of Mixture Experiments. *Journal of the American Statistical Association*, **111**, 899 – 911.

Wisniak, J. (2001). Historical Development of the Vapor Pressure Equation from Dalton to Antoine. *Journal of Phase Equilibria*, **22**, 622 – 630.

Joint analysis of nonlinear longitudinal and time-to-event data: application to predicting pregnancy outcomes

Rolando de la Cruz[1], Marc Lavielle[2], Cristian Meza[3], Vicente Nuñez-Antón[4]

[1] Universidad Adolfo Ibañez, Chile
[2] INRIA-Ecole Polytechnique, France
[3] Universidad de Valparaíso, Chile
[4] University of the Basque Country UPV/EHU, Spain

E-mail for correspondence: cristian.meza@uv.cl

Abstract: Nonlinear mixed effects models are statistical models containing both fixed and random effects. They are particularly useful in settings where repeated measurements are made on the same statistical units (longitudinal data), or where measurements are made on clusters of related statistical units. Observations in the same unit/cluster cannot be considered independent and mixed effects models constitute a convenient tool for modeling unit/cluster dependence. Nonlinear mixed effects models are commonly used in longitudinal data analysis since they can cope with missing observations and unbalanced data, and take into account individual variations from a common pattern. A commonly encountered complication in the analysis of longitudinal data is the variable length of follow-up due to interval censoring. This can be further exacerbated by the possible dependency between the time-to-event data and the longitudinal measurements. This paper proposes a combination of a nonlinear mixed effects model for the longitudinal measurements and a parametric model for the time-to-event data. The dependency is handled via latent variables, which are naturally incorporated. Estimation procedures based on the Stochastic Aproximation of the EM algorithm (SAEM) are proposed.

Keywords: Joint modeling; Longitudinal data; Nonlinear mixed models; Time-to-event data.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

1 Joint model formulation

Joint models can be used as a class of statistical methods for modeling longitudinal data and time-to-event (TTE) data together. In a biometrics setting, we often have, for a set of patients, time-to-event data of interest, for instance, the loss of the fetus during pregnancy. One may be interested in modeling the process inducing the event using, for example, a suitable selected (time-dependent or not) hazard function to describe the instantaneous chance of an event occurrence. Simultaneously, for each patient, we may be able to measure a longitudinal outcome and model its progression. It is common that a given longitudinal biomarker has a real influence on the TTE process. Mbogning et al. (2015) proposed a nonlinear mixed effects framework to jointly model longitudinal and repeated TTE data using a parametric mixed effects hazard model for the repeated event times, so that the link between both types of data, longitudinal and TTE data, is the conditional expectation of the longitudinal observation given the random effects or, more simply, function of the predicted longitudinal biomarker. In this work, we follow the idea in Mbogning et al. (2015), but we handle the dependency in the longitudinal and TTE data via latent variables, which are naturally incorporated (i.e., only some random individual effects are included in the survival model). Moreover, we consider a nonlinear mixed model for the longitudinal data and a parametric model to explain the TTE data, where both parts share a common parameter. In the case of the TTE data, the recorded observations are the times at which events occur. Here, we consider that the event can be interval censored. In addition, we assume that the responses under study are repeatedly measured for each of the m units over a period of time. For the i-th unit, $i = 1, \ldots, m$, observation times are restricted to a unit-specific time interval $[0, T_i]$; that is, observation times are interval censored in a time interval. For the longitudinal part, let y_{ij} be the response for the i-th unit at time t_{ij}. We consider the longitudinal data arising from a nonlinear mixed effects model with unit-specific random effects. More specifically,

$$\begin{aligned} y_{ij} &= f(t_{ij}, \Psi_i) + \epsilon_{ij}, \quad i = 1, \cdots, m,\ j = 1, \ldots, n_i, \quad \text{with} \\ \Psi_i &= \Psi_{pop} + \beta z_i + \eta_i, \end{aligned} \tag{1}$$

where Ψ_{pop} are population parameters, β is a set of coefficients, z_i a vector of individual covariates, the η_i's are the random effects, with $\eta_i \sim \mathcal{N}_d(0, \Gamma)$, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, and η_i and ϵ_{ij} are independent. For the TTE part, we need to concentrate on the specification of the survival, $S(t)$, and the hazard, $h(t)$, functions. Under our population approach, these functions are subject-specific functions and we will use parametric models for the TTE analysis (i.e., they depend on subject-specific parameters, Ψ_i), so that $S(t, \Psi_i) = P(T_i > t; \Psi_i)$ and $h_i(t) = h(\Psi_i, t)$. We will use the Weibull model defined for individual i, which shares individual parameters with

Model (1), so that $h_i(t) = \gamma \times \Psi_i \times \left(t^{\beta_w - 1}\right)$, where Ψ_i is a subject-specific effect for the i-th individual. We propose the use of a stochastic approximation version of the EM algorithm (SAEM) (Delyon et al., 1999) via the Monolix software to obtain maximum likelihood estimates of $\theta = (\gamma, \beta_w, \beta, \Psi_{\text{pop}}, \sigma^2, \Gamma)$. Moreover, if the simulation step cannot be directly performed, we propose to combine the SAEM algorithm with a Markov Chain Monte Carlo (MCMC) procedure (Kuhn and Lavielle, 2004).

2 Application: the pregnant women dataset

Data were collected during a clinical trial in a privately assisted reproduction center in Santiago, Chile. The data set consists of repeated measures of β-HCG concentration levels taken over a period of two years on 173 different pregnant women divided in two groups: (i) pregnancies with a normal development that came to term without important complications (124 individuals); and (ii) a group of abnormal pregnancies with serious anomalies that ended up with the loss of the fetus (49 individuals). Measurements were recorded at different times for each woman during the first trimester of pregnancy (first 80 days). It is well known that the β-HCG concentration levels in the two groups follow different patterns. The event here is the time of occurrence of the loss of the fetus in the abnormal group, which occurs within 10 days after the last measurement. We propose to fit this joint model to all longitudinal data (normal and abnormal groups), using the Weibull model defined above for the abnormal group via a random effect, a_i, so that the joint model defined in equations (2)-(3) is labelled as model $\mathcal{M}_1$:

$$y_{ij} = \frac{a_i}{1 + \exp\left\{-\frac{(t_{ij} - b_i)}{c_i}\right\}} + \varepsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \cdots, n_i, \tag{2}$$

$$h_{i^\star}(t) = \gamma \times a_{i^\star} \times \left(t^{\beta_w - 1}\right), \tag{3}$$

where $i^\star$ indicates that individuals belong to the abnormal group, $\varepsilon_{ij} \sim N(0, \sigma^2)$ and $a_i \sim N(a_{\text{pop}}, \sigma_a^2)$, $b_i \sim N(b_{\text{pop}}, \sigma_b^2)$, $c_i \sim N(c_{\text{pop}}, \sigma_c^2)$. For the abnormal group, we know that the event occurs in a period of time not exceeding 10 days after the last measurement, so the unknown event time for individual is in the interval $[l_i \leq T_i \leq l_i + 10]$, where l_i is the last measurement time for the i-th woman in the abnormal group. The parameter vector is then $\theta = (\gamma, a_{\text{pop}}, b_{\text{pop}}, c_{\text{pop}}, \sigma^2, \sigma_a^2, \sigma_b^2, \sigma_c^2)$. Here, random effects $\phi_i = (a_i, b_i, c_i)$ are treated as missing data. We propose to use a Stochastic Approximation of the EM algorithm (Delyon et al., 1999) using the Monolix software to obtain the maximum likelihood estimate of θ. Additionally, we also consider a second model which includes the variable group in parameter a_i. For this purpose, let (z_i) be a sequence of latent variables such that $z_i = 0$ if the i-th woman belongs to the normal group, and $z_i = 1$

otherwise. The statistical model for the individual parameter a_i is given by equation (4) in the joint models (2) and (3), which we label as model $\mathcal{M}_2$.

$$a_i = a_{\text{pop}} + \beta_a \times z_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma_a^2). \tag{4}$$

Finally, following Mbogning et al. (2015), we propose a third model which relates the predicted longitudinal biomaker, here the predicted concentration of β-HCG hormone levels in the the hazard function as:

$$h_{i^\star}(t) = \gamma \times Cc_{i^\star}(t) \times \left(t^{\beta_w - 1}\right), \tag{5}$$

where $Cc_{i^\star}(t)$ is the predicted β-HCG concentration for individual $i^\star$, belonging to the abnormal group, at time t. The specification in the joint models (2) and (5) will be labelled as model $\mathcal{M}_3$. Table 1 shows the estimated parameters and the BIC and AIC values obtained with the SAEM algorithm for the three models described above. We observe that the best model, based on these criteria, is model $\mathcal{M}_2$. Additionally, we could compute the predicted interval for the Kaplan Meier estimator, which can be obtained by Monte Carlo simulation via Monolix in Model $\mathcal{M}_2$.

TABLE 1. Estimated parameters obtained using the SAEM algorithm-Monolix.

Parameters	Model $\mathcal{M}_1$	Model $\mathcal{M}_2$	Model $\mathcal{M}_3$
a_{pop}	4.58 (.0487)	4.79 (.0523)	4.59 (.0537)
β_a	-	-0.759 (.0916)	-
b_{pop}	15.88 (.557)	15.7 (.594)	15.69 (.581)
c_{pop}	7.2 (.469)	7.42 (.544)	7.41 (.609)
γ	$1.1e^{-6}$ ($2.48e^{-6}$)	$1.32e^{-6}$ ($1.12e^{-6}$)	$5.9e^{-6}$ ($6.39e^{-7}$)
β_w	3.5 (.602)	3.46 (.226)	3.07 (.0814)
σ	0.262 (.0178)	0.249 (.0206)	0.437 (.0403)
σ_a	0.437 (.0391)	0.323 (.0391)	4.54 (.488)
σ_b	4.47 (.485)	4.14 (.86)	0.405 (.203)
σ_c	0.893 (.276)	1.67 (.807)	0.275 (.0193)
BIC	741.34	671.81	742.52
AIC	706.33	633.65	707.52

References

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Stat.*, **27**, 94 – 128.

Mbogning, C., Bleakley, K. and Lavielle, M. . (2015). Joint modeling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the SAEM algorithm. *Journal of Statistical Computation and Simulation*, **85(8)**, 1512 – 1528.

A multivariate geometric distribution for lifetimes of n-components series systems

Ricardo Puziol de Oliveira[1], Jorge Alberto Achcar[1]

[1] University of São Paulo, Brazil

E-mail for correspondence: rpuziol.oliveira@gmail.com

Abstract: System reliability studies usually assume independent lifetimes for the components in the estimation of the reliability of the system. This assumption in general is not reasonable in many engineering applications, since it is possible the presence of some dependence structure among the lifetimes of the components which could affect the evaluation of the reliability of the system. In the present study, it is assumed a dependence structure for the components and provided a new method to estimate the reliability of a n-component series system using a multivariate geometric distribution derived from the Marshall-Olkin method used due to its simplicity and flexibility.

Keywords: Bayesian analysis, Marshall-Olkin, multivariate geometric distribution, reliability analysis, series system.

1 Introduction

In the analysis of the reliability of component systems, an analyst first describes the overall design of the system in the form of a functional block diagram of reliability. In this way, a series system is a component configuration usually assumed in engineering studies such that if any one of the system components fails, the entire system fails. Associated to each system component there is a response given by a random variable that could be binary (fail/no fail) or denoted by its lifetime (a positive value). For a n-component series system with lifetimes associated to each component denoted respectively by $T_j\,(j = 1, \ldots, n)$, the reliability function of the system at a fixed time t, under independence assumption, is given by,

$$\mathrm{R}(t) = \mathrm{P}(\min(T_1, \ldots, T_n) > t) = \mathrm{R}_1(t) \ldots \mathrm{R}_n(t) \tag{1}$$

However, since in many practical situations in reliability engineering studies the lifetimes $T_j\,(j = 1, \ldots, n)$ are usually correlated which could affect

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

the estimation of the reliability of the system (see, for example, Singh and Billinton, 1977; Blanchard et al., 1990; Aggarwal, 2012), the reliability function of the system at a fixed time t, under dependence assumption, is given by,

$$\mathrm{R}(t) = \mathrm{P}(T_1 > t_1, \ldots, T_n > t_n) = \mathrm{R}_{1,\ldots,n}(t_1, \ldots, t_n), \tag{2}$$

where $\mathrm{R}_{1,\ldots,n}(t_1, \ldots, t_n)$ denotes a multivariate reliability function for the lifetimes T_j $(j = 1, \ldots, n)$.
The main goal of this paper is to introduce a multivariate geometric distribution in the estimation of the reliability of series system under a Bayesian approach assuming the multivariate n-component series system lifetime data and the dependence structure. The obtained inference results for the reliability of the system are compared to the usual approaches not considering the dependence structure.

2 Derivation of the Multivariate Geometric Distribution

In this study, the defining properties of the multivariate geometric distribution are based on models in which a two-component system fails according to the occurrences of fatal shocks to each one of the components or for all of the components. The first approach related to this idea introduced in the literature was proposed by Marshall and Olkin (1967) from where the authors introduced a multivariate exponential distribution.
Suppose that the components of a two-component system fail after receiving an overall fatal shock. Independent Poisson processes $U_1(t, \theta_1), U_2(t, \theta_2)$, $U_{12}(t, \theta_{12})$ govern the occurrence of fatal shocks. Events in the process $U_1(t, \theta_1)$ are fatal shocks transmitted to component 1, events in the process $U_2(t, \theta_2)$ are fatal shocks transmitted to component 2, and events in the process $U_{12}(t, \theta_{12})$ are fatal shocks transmitted equally and independently to both components. Therefore if $X = \min(U_1, U_{12})$ and $Y = \min(U_2, U_{12})$ denote, respectively, the lifetimes of the first and second components. In this case, the probability of the system is working until an overall failure is given by,

$$\mathrm{P}(X > x, Y > y) = \theta_1^x \theta_2^y \theta_{12}^{\max(x,y)} \tag{3}$$

The probability given by (3) is known in the literature as the Basu-Dhar bivariate geometric distribution introduced by Basu and Dhar (1995). Inferences and some computational aspects for this distribution under a Bayesian approach in the presence of censoring and covariates are introduced by Achcar et al. (2016); de Oliveira and Achcar (2018). An implementation of this distribution in R software is given by the package BivGeo introduced by Oliveira and Achcar, 2019. Similar arguments produce the

n-dimensional geometric distribution given by,

$$\begin{aligned} \mathrm{P}(X_1 > x_1, \ldots, X_n > x_n) &= \prod_{i=1}^{n} \theta_i^{x_i} \cdot \prod_{i=1<j}^{n} \theta_{ij}^{\max(x_i, x_j)} \ldots \cdot \\ &\times \theta_{12\ldots n}^{\max(x_1, x_2, \ldots, x_n)}, \end{aligned} \tag{4}$$

where $0 < \theta_i < 1, i = 1, \ldots, n$ and $0 < \theta_{ij}, \ldots, \theta_{12\ldots n} \leq 1, i = 1, \ldots, n; j = 2, \ldots, n; i < j$.

3 A numerical simulated data analysis

As an example of statistical analysis for the series systems, in this section it is presented the Bayesian Monte Carlo estimators (use of MCMC methods) for the reliability function R(t) for 2-components; 3-components and 4-components series systems. The obtained results are summarized in Figure 1 in which it is presented the plots of the estimated reliability functions and the 95% credible intervals for the reliability functions assuming the MVG distribution. Based on this simulated dataset, the reliability function for the system can also be estimated. For the specified time, t = 1, the true reliability value is obtained R(1) = 0.7695 (2-components), R(1) = 0.6063 (3-components) and R(1) = 0.4184 (4-components). The estimated Bayesian estimators based on the simulated Gibbs samples for R(1) are presented in Table 2 for each sample size assuming the MVG model and assuming the independence structure.

TABLE 1. Bayes estimators for $R(1)$ for each simulated dataset for each series system under dependence and independence assumption.

Dependence Assumption				Independence Assumption			
Sample	2-comp.	3-comp.	4-comp.	Sample	2-comp.	3-comp.	4-comp.
20	0.7319	0.5076	0.2912	20	0.7008	0.4387	0.2026
50	0.7822	0.5707	0.3466	50	0.7410	0.4813	0.1762
100	0.7776	0.5450	0.3748	100	0.7413	0.4597	0.2236
150	0.7652	0.5841	0.3960	150	0.7237	0.4700	0.2132
300	0.7687	0.5994	0.4161	300	0.7281	0.4927	0.2227

4 Conclusion

In this study, it is possible to conclude based on the illustrated simulated data application that the use of MVG distribution lead to more accurate results assuming the dependence structure than the approach assuming independence structure with univariate geometric distributions for the system lifetimes.

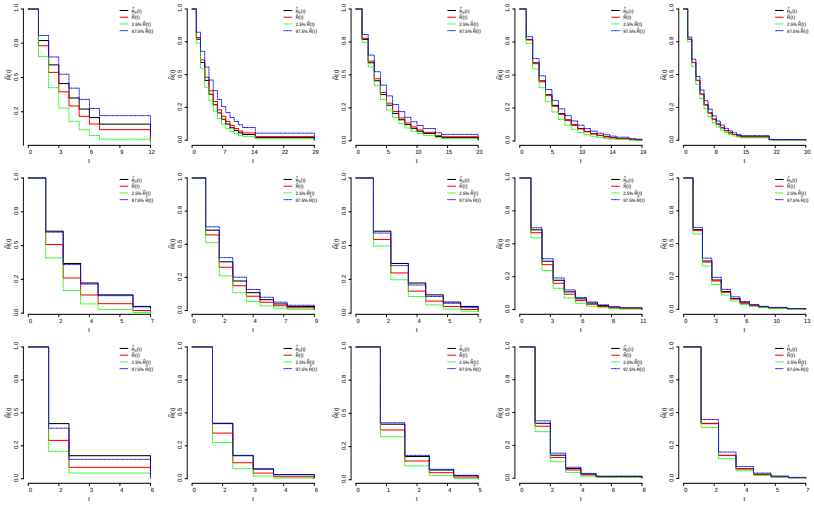


FIGURE 1. Empirical reliability function, mean estimate and 95% credible intervals for the reliability function of the MVG model assuming the multivariate lifetimes of each considered series system (top to bottom: 2-components → 4-components) for each sample size (left to right: $n = 20 \rightarrow 300$).

References

Marshall, A. W. and Olkin, I. (1967). A multivariate exponential distribution. *Journal of the American Statistical Association*, **62**, 30 – 44.

Singh, C. and Billinton, R. (1977). System reliability, modelling and evaluation. Hutchinson, London.

Blanchard, B. S., Fabrycky, W. J., and Fabrycky, W. J. (1990). Systems engineering and analysis. Prentice Hall Englewood Cliffs, NJ.

Aggarwal, K. (2012). Reliability engineering. Springer Science & Business Media.

Achcar, J., Davarzani, N., and Souza, R. (2016). Basu-Dhar bivariate geometric distribution in the presence of covariates and censored data: a Bayesian approach. *Journal of Applied Statistics*, **43**, 1636 – 1648

de Oliveira, R. P. and Achcar, J. A. (2018). Basu-Dhar's bivariate geometric distribution in presence of censored data and covariates: some computational aspects. *Electronic Journal of Applied Statistical Analysis*, **11**, 108 – 136.

Basu, A. P. and Dhar, S. (1995). Bivariate geometric distribution. *Journal of Applied Statistical Science*, **2**, 33 – 44.

On the selection of number of knots in linear regression splines with free-knots

Gioia Di Credico[1], Francesco Pauli[1], Nicola Torelli[1]

[1] University of Trieste, Italy

E-mail for correspondence: gioia.dicredico@deams.units.it

Abstract: Linear regression splines are useful tools to describe departures from linearity in several real applications. Location of knots can be seen as change points in the relationship between the variables. In a Bayesian context, we analyze the variation of the Stochastic Search Variable Selection approach previously proposed in Di Credico et al. (2018), focusing on the impact of the hyperparameters choice on the estimation of the correct number of knots.

Keywords: Linear regression splines; Free-knots; SSVS.

1 Problem and methods

Generalized linear models (GLM) are flexible tools to describe a linear relationship between the response, transformed by the link function, and some continuous covariates in the linear predictor. In many real applications, data show that this linearity assumption might be too restrictive or unable to describe the real connection among the variables. We may take as an example some epidemiological studies where the dose-response relationship highlights a saturation effect at high dose levels resulting into a change in the effect of one (or more) covariate included in the model. Assuming that the underlying relationship between the outcome and one continuous predictor can be well approximated by a piece-wise linear function, we can relax the linearity assumption modeling the continuous covariate by a spline function of degree one, that is

$$\eta = Z\alpha + f(x),$$

where η represents the linear predictor, Z is the matrix of covariates that enter linearly in the model, α is the vector of regression coefficients, x

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

is a continuous covariate, and $f : \mathbb{R} \to \mathbb{R}$ is a suitable linear piece-wise polynomial with joint points called knots (Ruppert et al., 2003).
In some problems, the number and location of the knots may have an important and meaningful interpretation, turning their estimation into one of the main objectives of the analysis. However, considering locations and number of knots as parameters adds several layers of complexity to the estimation problem. Moreover, without adding a penalization term, the spline function may lead to extremely different estimates depending on the choice of these two quantities.
We focus on those situations in which it is reasonable to assume a limited number of change points and we choose the truncated linear basis functions to represent the spline f, that is

$$f(x) = \beta_0 + \beta_1 x + \sum_{i=1}^{K} \gamma_i (x - k_i)_+,$$

where β_0, β_1 and γ_i for $i = 1, \ldots, K$ are the spline coefficients, K is the number of knots, k_i is the location of the i^{th} knot, and $(x - k_i)_+$ is the truncated linear function. This representation allows us to have a direct interpretation of the knot locations as change points of the slope.
Regression and spline coefficients, and knot location parameters can be simultaneously estimated using Monte Carlo Markov Chains (MCMC) methods (Carpenter et al., 2017; Di Credico, 2018). Estimation of the number of knots represents a more challenging problem since it is directly linked with the dimension of the parameter space.
We analyze the two step procedure proposed by Di Credico (2018) to estimate the number and location of knots, focusing on the first step, that is, the selection of the number of knots. Briefly, a variation of the Stochastic Search Variable Selection (SSVS) approach by George and McCulloch (1995) is used to select the right number of knots in a possibly over-parametrized model. Since each knot location parameter k_i is uniquely linked to a spline coefficient γ_i, the variable selection is performed adopting spike and slab prior distributions on the coefficients γ_i

$$\gamma_i \sim \lambda_i N(0, \sigma_{sl}) + (1 - \lambda_i) N(0, \sigma_{sp}), \quad i = 1, \ldots, K$$

where $\lambda_i \in (0, 1)$, is the mixing proportion of the mixture distribution, $\sigma_{sl} > 0$ and $\sigma_{sp} > 0$ are the standard deviations respectively of the slab and of the spike mixture components. The choice of the values for σ_{sl} and σ_{sp} needs to be carefully evaluated paying attention to the scale of the data and the link function.
The choice of the prior distribution for the parameter λ_i is a critical point of the methodology. Instead of a standard Bernoulli distribution, we specify a Beta distribution on the mixing proportion parameters

$$\lambda_i \sim Beta(a_i, b_i), \quad i = 1, \ldots, K$$

where $a_i \in (0,1)$ and $b_i \geq a_i$. The main advantage of a continuous prior distribution is the improvement of the mixing of the MCMC sampler due to the less restrictive geometry of the posterior distribution space to be explored (Rinta-aho and Sillanpää, 2019). When $a_i = b_i$, the smaller their value, the higher the concentration of the density function of λ_i on 0 and 1. While, the higher the value of b_i, the higher the concentration of the density function on 0. Di Credico et al. (2018) showed an improvement of the algorithm performances when each hyperparameter b_i is modeled as a function of the knot location k_i, and $a_i = 0.5, \forall i$. Since the abuntant knots are pushed towards the boundary of the predictor range, the function on b_i moves towards 0 the distributions of the λ_i associated to those knots. Rinta-aho and Sillanpää (2019) proposed to specify $b = 1 - a$, choosing to set $a \sim Unif(0,1)$. We decided to compare the previously tested approaches with the one proposed by Rinta-aho and Sillanpää (2019), and try to include in the prior distribution on λ_i the information about the knot location.
Prior distributions for the regression parameters α and the spline parameters β_0 and β_1 are defined as weakly informative. Prior distributions on the knot location parameters k are defined as Uniform on the predictor range, subject to ordering constraint to ensure identifiability. The final number of knots is selected examining the posterior distribution of the mixing proportion parameter λ_i together with the posterior distributions of the knot locations (Di Credico, 2018). Analysis were performed using the software R (R Core Team, 2016) and Stan (Stan Development Team, 2017).

2 Results and conclusions

Using synthetic data from a linear, logistic and Poisson regression with two true knots, we tested the impact of several definitions of prior distributions on the mixing proportion parameters λ, fixing the number of estimated knots to 2, 5 and 10.
As expected, when the density of the prior distributions is very highly concentrated, e.g. $a \leq 0.1$ or $b \leq 0.1$, we experienced performance degradation of the algorithm. Diagnostic tools reported divergences of the algorithm (Carpenter et al., 2017), and poor mixing of the chains, that easily got stuck on specific areas of the posterior distribution. Comparing the two specifications of the hyperparameter: $b = 1 - a$ and $b \geq a$; gives very similar results in terms of the marginal posterior distributions of the knot locations k_i. Whereas, in the latter case, the marginal posterior distributions of λ_i give clearer indications about the choice of the number of knots. In both specifications, using the knot locations in the definition of b facilitates the interpretation of simulation results in selecting the number of knots. Moreover, the algorithm is faster and diagnostic statistics suggest better quality of the simulations if compared with the ones obtained without the knot location function on b.

Future steps involve the study of the connection between the Beta density prior distribution defined on λ and the choice of a scale mixture of Normal distributions. The role of the variances of the spike and slab mixture components will be also explored.

References

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., et al. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, **76(1)**, 1 – 32.

Di Credico G. (2018). *Some developments in semiparametric and cross-classified multilevel models*. Ph.D. Thesis, http://paduaresearch.cab.unipd.it/1157.

Di Credico G., Pauli F., and Torelli N. (2018) Bayesian estimation of number and position of knots in regression splines. In: *Book of short Papers SIS 2018*, Palermo, Italy.

George, E.I., and McCulloch, R.E. (1995). Stochastic search variable selection. *Markov chain Monte Carlo in practice*, **68**, 203 – 214.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rinta-aho, M.J. and Sillanpää M.J. (2019). Stochastic search variable selection based on two mixture components and continuous-scale weighting. *Biometrical Journal*, **61**, 729 – 746.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric regression*. Cambridge university press.

Stan Development Team (2017). Stan modeling language users guide and reference manual. Version 2.15.0.

Looking for growth curves in the situation designed by François Cretté de Palluel (1788)

Antoine de Falguerolles[1]

[1] Retired senior lecturer, Université de Toulouse (III), France

E-mail for correspondence: antoine@falguerolles.net

Abstract: This poster presentation aims at getting advice on candidates for growth curves observed under a constrained design (Latin-square).

Keywords: Growth curves; Experimental design; History.

1 Introduction

At the turn of the 19th century, Agriculture Societies were active, farming experiences were discussed at meetings or published, good practices circulated. This was the case in Great-Britain under the steering of Arthur Young (1741 - 1820) and sir John Sinclair (1754 - 1835), to name the most popular. It was also the case on the Continent. However, these experiments do not lend themselves to modern statistical treatment as pioneered by sir Ronald A. Fisher in the early 20th century: in most instances it is impossible to extract from the publications a data set in the format which statisticians are familiar to.
A publication of the French Franois Cretté de Palluel (1741 - 1798) provides a known exception: the aim, design, data and results of the experiment which he conducted near Paris are reported in an exceptional tabular format (Palluel, 1788). Frequently referred in agricultural statistics papers, the introduction of a Latin-square design has somewhat obliterated Palluel's primary objective which was to characterize growth curves conditioned on breeds and feeds (in the context of censored observations).
The aim of this poster presentation is to recall the effectiveness of Palluel's presentation of the data in tabular form, to exhibit graphical representations of his data, discuss preliminary exploratory analyses, but mostly to ask for advice on the type of parametric growth curves suitable for handling such data.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Motivation of the experiment

On 31 July 1788, Palluel gave a talk at the *Société royale d'agriculture* seated in Paris on the design and result of an experiment he had conducted on sheep feeds in barn fattening. The talk was immediately published (Palluel, 1788) and republished (in English) in Arthur Young's *Annals of Agriculture* (1790). The practical motivation for Palluel was the profit that could be generated by barn fattening of sheep when the price of meat was high. But when the price of cereals is high, less costly feeds have to be experienced and proposed. Palluel's interests proved to be premonitory: France had a disastrous harvest of cereals in the summer 1788 with a cataclysmic hailstorm which hit France on Sunday, 13 July, destroying the crops on about $100,000$ *hectares*. Additionally, winter 1788-1789 was very severe.

3 The experiment

In Palluel's experiment 4 sheep breeds (Île de France or local, Beauce, Champagne, Picardy) were fed on 4 feeds (Potatoes, Cereals, Turnips, Beets) during 4 months at most. (Like the French agronomist Antoine Parmentier (1737-1813), Palluel advocated potatoes in all possible forms to a reluctant French population.) Sheep were killed after one, two, three, or four months of fattening. At the end of each month, the weights and carcass compositions of the slaughtered sheep were recorded, as well as the weights of all remaining sheep. Thus Palluel had information on the overall weight growth of sheep for any given breed, feed (his primary objective) and duration. He had also some information on the structural aspects of the observed growth. Palluel wanted to conduct the experiment in such a way that all breeds, feeds, fattening durations were represented. To reduce the 64 sheep which the experiment called for on an a priori basis, Palluel cleverly designed a $\frac{1}{4}$ replicate of a 4^3 factorial, or a 4×4 Latin-Square (Susan Wilson, 2009, p. 8).

TABLE 1. At the end of each month, weight and composition of the carcass is observed on 4 units representing the 4 different breeds and the 4 different feeds, but never in the same association.

Period 1	Period 2	Period 3	Period 4
Breed Feed	Breed Feed	Breed Feed	Breed Feed
1 1	1 2	1 3	1 4
2 2	2 3	2 4	2 1
3 3	3 4	3 1	3 2
4 4	4 1	4 2	4 3

4 Data

Palluel's data set is given in the form of two tables (*Tableau* 1 and *Tableau* 2) which occupy each a full page in landscape orientation and are not reproduced here for lack of space. The rows are the sheep which are referenced by a number; the columns are the measured weights. *Tableau* 2 reports the initial weights and the weights at killing (2 columns), and the associated compositional weights of the carcass (9 columns) for each sheep at its killing; rows corresponds to the 16 combination of feed and breed. *Tableau* 1 consists in the column concatenation of three sub-tables: 1) weights at initial date, at intermediate dates, and at killing (5 columns); 2) monthly incremental weights (4 columns); 3) total weight increments (1 column). The 16 rows are ordered by feed and, within feed, by breed. Some cells are structurally empty due to a planned duration of fattening shorter than 4 months.
Interestingly, Palluel inserted in *Tableau* 1 4 lines which report the sum over each breed of the monthly weight increments. The concatenation of these 4 statistics forms a 4×4 two-way table (feed $\times$ duration of fattening)seems to have guided Palluel's conclusions. Similar two-way tables could have been formed: breed $\times$ duration of fattening, feed $\times$ breed. This suggests to construct a square table analogous to a Burt table giving in this context the sums of the response (here monthly weight increment) for all two-way cross classifications of the factors. (As in Burt tables diagonal blocks are diagonal, with terms formed from the corresponding one-way classification.) This is shown below in Table 2 (upper triangular part).

5 Analyses

5.1 Linear modelling

A linear modelling approach is an obvious strategy in this context. The response is either weights or monthly incremental weights until slaughtering and the cofactors are feed, breed and duration. Various forms of serial dependence for the error term can be tested. Interactions are also an issue.

5.2 Exploratory analysis of two-way interactions

When investing two-way interactions in contingency tables, multiple correspondence analysis (MCA) is a useful practice. Its core is to derive a significant low rank approximation to the Burt matrix of counts which highlights the structure of two-way interactions. Is such an approach conceivable in this context? Following the Palluel's intuition, two Burt clones can be now constructed as reported here in Table 2: one for the sums (or averages) of the response and one for the counts of observations leading to theses sums (averages). These have to be combined in analyses mimicking MCA to be discussed.

5.3 Reduced rank multiplicative interactions

It is nowadays standard to introduce reduced rank multiplicative interactions in a hierarchy of linear models. This could be done in GLIM long ago (generalised bilinear models) and is easily performed now with the excellent R *gnm* package developed by Heather Turner and David Firth (as consulted in 2020).

TABLE 2. The upper triangular part, by symmetrization, gives the Burt clone for the sums of monthly weight increments. The lower triangular part, by symmetrization, gives the Burt clone for the counts of associated responses; the counts are given in parenthesis to emphasize their different meaning. The diagonal gives the diagonal terms for both clones. All statistics typed in bold characters were introduced by Palluel. Some typographic errors have been corrected.

	Feed Potatoes	Feed Turnips	Feed Beets	Feed Cereals	Breed local	Breed Beauce	Breed Champagne	Breed Picardy	Weight increment 1 month	Weight increment 2 months	Weight increment 3 months	Weight increment 4 months
Potatoes	**70.00** (10)				10.00	24.25	14.75	21.00	**50.50**	**13.25**	**4.25**	**2.00**
Turnips		**67.50** (10)			18.00	15.00	16.00	18.50	**58.50**	**7.00**	**1.50**	**0.50**
Beets			**71.50** (10)		22.00	15.25	13.25	21.00	**48.00**	**17.50**	**4.00**	**2.00**
Cereals				**92.5** (10)	32.00	22.50	22.00	16.00	**59.00**	**18.50**	**11.00**	**4.00**
local	(1)	(2)	(3)	(4)	82.00 (10)				55.25	12.75	10.00	4.00
Beauce	(4)	(1)	(2)	(3)		77.00 (10)			47.50	20.25	7.25	2.00
Champagne	(3)	(4)	(1)	(2)			66.00 (10)		52.25	10.25	3.00	0.50
Picardy	(2)	(3)	(4)	(1)				76.50 (10)	61.00	13.00	0.50	2.00
1 month	(4)	(4)	(4)	(4)	(4)	(4)	(4)	(4)	216.00 (16)			
2 months	(3)	(3)	(3)	(3)	(3)	(3)	(3)	(3)		56.25 (12)		
3 months	(2)	(2)	(2)	(2)	(2)	(2)	(2)	(2)			20.75 (8)	
4 months	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)				8.50 (4)

6 Discussion

Still, a smarter approach to Palluel's data would be to find a family of parametric growth curves which could take into account the structure of the cofactors and the censoring. I need help in that matter.

References

Cretté de Palluel, F. (1788). Mémoire sur les avantages et l'économie que procurent les racines employées à l'engrais des moutons à l'étable. In: *Mémoire d'agriculture, d'économie rurale et domestique*. Paris: Cuchet, 17 – 23. (English translation (1790). On the Advantages and Economy of Feeding Sheep in the House with Roots. In: *Annals of Agriculture and Others useful Arts collected by Arthur Young*, **14**, 133 – 139.)

Turner, H. (2020). *gnm: Generalized Nonlinear Models. R package. Version 1.1-1.*

Wilson, S. R. (2009). Modern Biometry. In: *Biometrics, vol. 1.* Eds by Susan R. Wilson and Conrad Burden, EOLSS Publications, 1 – 33.

Modeling Mothers' yearly earnings after returning from maternity leave with a Bayesian distributional regression model

Tim Föckersperger[1], Helga Wagner[1]

[1] Department for Applied Statistics, Johannes Kepler University Linz, Austria

E-mail for correspondence: tim.foeckersperger@jku.at

Abstract: The goal of this paper is to estimate and compare the effects of potential covariates on mothers' yearly earnings in the first and the sixth year after the maternity leave period in Austria. To go beyond modeling only the mean yearly salary as a function of covariates, we analyze the data with a Bayesian distributional regression model where the response variable is assumed to follow a zero-adjusted Gamma distribution. This allows to model the proportion of mothers with zero income as well as the Gamma-specific parameters in terms of covariates.

Keywords: Bayesian Distributional Regression; MCMC; Zero-adjusted Gamma

1 Introduction

In Austria many mothers delay returning to the labor market after the period in which they get maternity leave benefits. Thus, a non-negligible proportion of mothers has no (or very low) earnings even years after their last child-birth. To analyze effects of covariates on earnings of mothers we do not restrict the analysis to mothers with positive income but use a Bayesian distribution regression model that allows to model zero as well as non-zero earnings. We use the zero-adjusted Gamma distribution (ZAGA) which is a mixture of a point mass at zero to model the proportion of mothers with no earnings and a Gamma distribution to model earnings of employed mothers and allow all three parameters of the ZAGA, the weight of the point mass at zero as well as the parameters of the Gamma distribution, to vary with covariates.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Data

Our analysis is based on the yearly earnings of $n = 30107$ mothers, who gave birth to their last child within a 2-year period starting from July 2000 when a fundamental extension of the maternity leave benefits from 18 to 30 months became effective in Austria. To investigate the development of mothers' earnings over time we will analyze the incomes in the first as well as in the sixth full year after the maximum maternity leave period of 30 months. Zero earnings are observed for mothers who are not employed in the respective years. Additionally, we define earnings which are below the inflation-adjusted minimum yearly wage as zero. As potential covariates we have information on the type of employment (indicator for being a blue collar worker) and earnings of the mother in the year before birth, as well as the age and additionally for mothers who are employed after the child birth whether they returned to the same employer, worked part or full time and started working before or after the job protection period of 24 months.

3 Bayesian distributional regression

In Bayesian distributional regression the data $\mathbf{y}$ are modeled as realizations of a K-parametric distribution D

$$y_i|\mathbf{x_i} \sim D(\theta_1(\mathbf{x_i}), ..., \theta_K(\mathbf{x_i})) \tag{1}$$

with parameters θ_k, k = 1, ..., K. In order to connect the covariates to a parameter a monotonic, twice differentiable link function $h_k(\cdot)$ is applied that guarantees to preserve the domain restriction of a parameter. Further, as in additive models the linear predictor is assumed to consist of a sum of unspecified functions $f_{\theta_k,j}(\cdot)$, $j = 1, ..., J$, which allow to model the effects of the covariates $\mathbf{X}$ as linear, non-linear, spatial or as random effects, i.e.

$$h_k(\theta_k(\mathbf{x_i})) = \eta_{\theta_k,i} = f_{\theta_k,1}(\mathbf{x_i}; \beta_{\theta_k,1}) + ... + f_{\theta_k,J}(\mathbf{x_i}; \beta_{\theta_k,J}) \tag{2}$$

where $\boldsymbol{\beta}_{\theta_k} = (\beta_{\theta_k,1}, ..., \beta_{\theta_k,J})$ are the regression parameters whose structure depend on the type of covariate(s) and prior assumptions about $f_{\theta_k,j}(\cdot)$.

4 Zero-adjusted Gamma Distribution

The zero-adjusted Gamma distribution (ZAGA) is a mixture of a point mass at zero and a Gamma distribution. Its pdf is defined as

$$f(y|\mu, \sigma, \nu) = \begin{cases} \nu & \text{if } y = 0 \\ (1-\nu)\frac{1}{(\mu\sigma^2)^{1/\sigma^2}\Gamma(1/\sigma^2)} y^{1/\sigma^2-1} \exp(-\frac{y}{\mu\sigma^2}) & \text{if } y > 0 \end{cases} \tag{3}$$

with $0 \leq \nu \leq 1$ and $\mu > 0$, $\sigma > 0$. The advantage of the alternative parametrization of the Gamma component is that the expected value

$E[y|y > 0] = \mu$ and the skewness $\gamma[y|y > 0] = 2\sigma$ are linear functions of the parameters μ and σ which allows to interpret the corresponding regression effects as effects on the conditional expectation and skewness. Generally, the ZAGA is a suitable choice as its point mass at zero allows to model the proportion of unemployed mothers and its continuous component accommodates the right skewness of the earnings. For the parameter ν a logit link is applied and analogously to GLMs, a loglinear link function is used for the parameter μ as well as for σ to guarantee restriction to their domain.

5 Results

To model mothers' yearly earnings after the maternity benefit period we use only her age, the reference earnings and the employment type before birth as covariates for ν, whereas all six covariates are considered to model μ and σ. To allow for possible non-linear dependencies, the reference wage and the age are modeled with penalized splines.
To estimate the model parameters we employ MCMC methods with standard prior distributions as implemented in the *R* function bamlss(). Results are based on 18000 MCMC draws after a burnin of 2000 and an additional thinning of 40. Table 1 reports estimates of the posterior means of the exponentiated effects of the categorical covariates as well as the 2.5 % and 97.5% quantiles of their posterior distributions. For ν the reported estimates are multiplicative effects on the odds ratio of being unemployed and for μ multiplicative effects on the expected mean salary for employed mothers. Further, as the skewness is a linear function of σ, the estimated effects on σ can also be interpreted as multiplicative effects on the skewness of the wage distribution.

Covariates		**Year 1**			**Year 6**		
		ν	μ	σ	ν	μ	σ
Blue Collar	$\exp(\beta)$	1.277	0.847	0.926	1.249	0.785	0.914
		[1.207, 1.354]	[0.837, 0.858]	[0.905, 0.946]	[1.154, 1.332]	[0.775, 0.794]	[0.896, 0.933]
Same Employer	$\exp(\beta)$		0.995	0.877		0.975	0.899
			[0.976,1.014]	[0.850, 0.901]		[0.959, 0.993]	[0.877, 0.924]
< 24 Months	$\exp(\beta)$		0.695	1.576		0.759	1.335
			[0.683, 0.707]	[1.517, 1.641]		[0.745, 0.774]	[1.282, 1.391]
Part time	$\exp(\beta)$		0.785	0.850		0.731	0.849
			[0.776, 0.794]	[0.834, 0.867]		[0.722, 0.739]	[0.834, 0.864]

TABLE 1. Exponentiated effects of the categorical covariates on the parameters

In the first full year after the maternity benefit period the odds of being unemployed is 1.277 times higher for blue compared to white collar workers. Working part-time, being a blue-collar worker before birth and returning to work before 24 months decreases the expected mean salary substantially; σ and hence the skewness is smaller for blue collar workers and mothers working part-time but considerably larger for mothers who were returning to work early. Effects are similar in year 6 with some exceptions: the effect of being a blue collar worker on the expected income is smaller in year 6

than in year 1 which means the earning gap between blue and white collar workers has increased; also the effect of an early return to the labor market on the mean has increased, but its effect on the skewness has decreased. To visualize the effects of the reference wage and the age the exponentiated non-linear effects (centered at the mean of the corresponding covariate) are displayed in Figure 1.

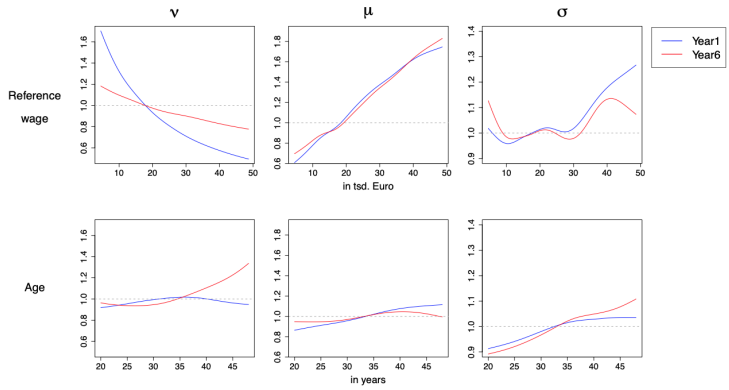


FIGURE 1. Exponentiated effects of the reference wage and the age on the parameters

The odds ratio of being unemployed decreases with the reference wage but this effect is much less pronounced in year 6 than in year 1. There is no effect of the age on the odds ratio of being unemployed in year 1, and a slightly increased odds for older mothers in year 6. The expected mean earnings increase with the reference wage whereas the effect of the age is negligible in both years. Also the effects on the skewness are similar in both years: It is higher for low and high reference wages and increases with age. Finally, we conclude that modeling yearly earnings of mothers with a Bayesian distributional regression model allows for a more detailed insight on the effects of covariates.

References

Jacobi L., Wagner H. and Frhwirth-Schnatter S. (2016). Bayesian Treatment Effects Models with Variable Selection for Panel Outcomes with an Application to Earnings Effects of Maternity Leave. *Journal of Econometrics*, **193 (1)**, 234-250.

Umlauf N., Klein N. and Zeileis A. (2018). BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond). *Journal of Computational and Graphical Statistics*, **27 (3)**, 612-627.

Umlauf, N. and Kneib, T. (2018). A primer on Bayesian distributional regression. *Statistical Modelling*, **18 (34)**, 219247.

Analyzing Likert-Type Data using Penalized Non-Linear Principal Components Analysis

Aisouda Hoshiyar[1]

[1] Helmut Schmidt University, Hamburg, Germany

E-mail for correspondence: aisouda.hoshiyar@hsu-hh.de

Abstract:
We consider a survey on animal ethics and sustainability consisting of various Likert-type items. Although this kind of (ordinal) data often occurs in the social sciences, in case of principal components analysis (PCA) those data are either treated as numeric implying linear relationships between the variables at hand, or nonlinear PCA is applied where the obtained coefficients are sometimes hard to interpret. We therefore revisit penalized nonlinear PCA for ordinal variables as an intermediate between the mentioned methods used so far. The new approach offers both better interpretability as well as better performance on validation data.

Keywords: Ordinal Variables; Prinicipal Compontens Analysis; Optimal Quantification; Smoothing.

1 Introduction

At IWSM 2013, Gertheiss and Kiers presented the idea of penalized nonlinear principal components anaylsis (PCA) as an intermediate between standard, linear PCA, simply using the levels of ordinal variables as numerical input, and optimal scaling as, e.g., described by Linting *et al.* (2007). In short, the general aim of PCA is to reduce the observed variables to a number of uncorrelated linear combinations - called principal components - while explaining as much of the variability in the original data as possible. The extended nonlinear approach respects the scale level of ordinal variables through the process of optimal quantification. The objective is achieved by assigning numerical values to the ordered levels via nonlinear transformations - the quantifications. However, the found quantifications often result in overfitting the (training) data, which worsens

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

the performance and generalization to new data. In addition, the obtained quantifications are sometimes erratic and thus hard to interpret. Therefore, Gertheiss and Kiers (2013) introduced an additional penalty term penalizing nonlinearity in the coefficients. As an intermediate between standard linear PCA and fully nonlinear PCA, the proposed approach offers both better interpretability of the nonlinear transformation as well as better performance on validation data.

The general idea is as follows: Nonlinear PCA minimizes the criterion $L(\Phi, Y, A) = \sum_j \sum_i (\phi_{ij} - \sum_r y_{ir} a_{rj})^2$ as a function of matrices A, Y and Φ, with $(\Phi)_{ij} = \phi_{ij} = \varphi_j(x_{ij})$ and $i = 1, ..., n$; see Linting *et al.* (2007). A and Y correspond to loadings and respective PC scores when using the transformed variables, $r = 1, ..., m$ and m corresponds to the number of PCs to be extracted. Scaling function φ_j, $j = 1, ..., p$, can also be represented by the vector $\theta_j = (\theta_{j1}, ..., \theta_{jk_j})^T$ where θ_{jl} is the value that is assigned to category l of the jth variable, k_j denotes the highest level of variable j. With linear scaling function $\varphi_j(x_{ij})$, the approach is equivalent to usual PCA using the group labels $1, 2, ..., k_j$. So for a trade-off between the latter approach and optimal scaling in its pure form, deviations from linearity are penalized when fitting φ_j. More precisely, a second-order difference penalty is used in terms of

$$J(\theta) = \sum_{j=1}^{p} \sum_{l=2}^{k_j - 1} (\theta_{j,l+1} - 2\theta_{jl} + \theta_{j,l-1})^2.$$

2 Application: Animal Ethics

We consider a survey conducted by the Department of Animal Sciences, University of G$\acute{}$ottingen. The data set consists of 2000 observations of 33 ordinally scaled variables addressing sustainability indicators with regard to animal welfare, human health, and environmental issues. Each statement of agreement is measured on a five-point Likert scale with: 1 strongly agree, 2 agree, 3 undecided, 4 disagree, 5 strongly disagree. We perform the proposed method initially with $m = 6$ resulting from the scree test after also comparing to the corresponding plots for $m = 5$ and $m = 7$. Figure 1 illustrates the estimated coefficients of selected variables for different values of the penalty parameter λ. The black lines refer to unpenalized nonlinear PCA (i.e., $\lambda = 0$), the red dashed lines refer to $\lambda = 1$, and the green dotted lines to $\lambda = 10$. It is noticable that with an increasing penalty parameter quantifications become increasingly linear, with the latter being equivalent to standard linear PCA using just the class labels. For the variable "Drive less car" in Figure 1 (right) the impact of the penalty can be seen noticably with regularization towards linearity. On the other hand, it is observed (Figure 1, left) that also non-monotonic effects can be discovered, which is a clear benefit of nonlinear PCA over usual (linear) PCA in general. When

using the method proposed, coefficients θ_j are smoother than for unpenalized nonlinear PCA, which is convincing, as wiggly coefficients are hard to interpret. In addition, the possibility of incorporating constraints enforcing monotonicity is provided, as this assumption is reasonable for some practical applications.

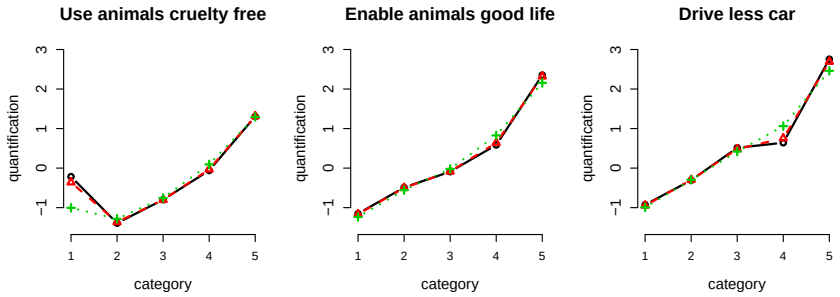


FIGURE 1. Category quantifications for $\lambda = 0$ (solid black), $\lambda = 1$ (dashed red), $\lambda = 10$ (dotted green).

To obtain an optimal amount of shrinkage, the smoothing parameter λ was determined based on fivefold cross-validation (the optimal smoothing parameter is indicated as a dashed line in Figure 2). Based on this procedure, the performance of the quantification rule is measured by the proportion of variance that is explained by the first $m = 6$ principal components. The proportion of variance explained as a function of λ is demonstrated in Figure 2 on a logarithmic scale for both the training sample as well as the validation sample. Cross-validation shows that results of nonlinear PCA can be improved by using the suggested penalized fitting algorithm. Although penalized scaling functions are less complex, and thus easier to interpret, performance does not deteriorate on both training and validation data up to a certain lambda value.
To obtain the final scaling rule, however, a distinct λ-value needs to be chosen. For that purpose, cross-validation results as given in Figure 2 (right) can be used. For the sustainability data, we would use $\lambda \approx 1$, where the proportion of variance explained on the test data reaches its maximal value.

3 Concluding remarks

In this article, we revisited an extension of nonlinear principal components analysis for ordinal data with two crucial benefits over both, the linear and the fully nonlinear version: The ability of discovering and handling nonlinear and even non-monotonic relationships between variables along

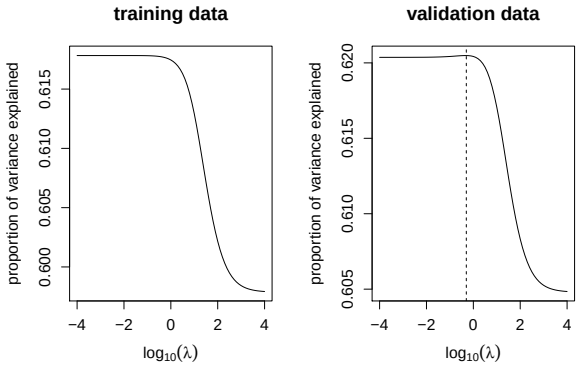


FIGURE 2. Mean proportion of variance explained by the first 6 principal components; left: training sample, right: validation sample.

with respecting the ordinal scale of the data, while avoiding overfitting as well as offering better interpretability of the estimated coefficients. Our preliminary results on real (and simulated) data suggest that penalized nonlinear PCA is a promising and convincing framework for dimension reduction of ordinally scaled data sets.

Acknowledgments: I thank Daniel Mörlein (Department of Animal Sciences, Georg August University, Göttingen) for providing the data on animal ethics and related information. Furthermore, I want to thank Jan Gertheiss for valuable discussion of the manuscript.

References

Gertheiss, J. and Kiers, A.L. (2013). Penalized Non-Linear Principal Components Analysis for Ordinal Variables. *In V.M.R. Muggeo, V. Capursi, G. Boscaino, and G. Lovison (eds.): Proceedings of the 28th International Workshop on Statistical Modelling*, **3**, 607-610.

Gertheiss, J. and Oehrlein, F. (2011). Testing linearity and relevance of ordinal predictors. *Electronic Journal of Statistics*, **5**, 1935-1959.

Linting, M., Meulmann, J. J., van der Kooij, A. J. and Groenen, P. J. F. (2007). Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, **12**, 336-358.

Bayesian Inference for modelling the Uncertainty by a Mixture Model for rating data

Maria Iannario[1], Claudia Tarantola[2]

[1] University of Naples Federico II, Italy
[2] University of Pavia, Italy

E-mail for correspondence: maria.iannario@unina.it

Abstract: In this paper we perform Bayesian quantitative analysis of the CUP model, which is a two-component mixture model recently introduced for the analysis of ordinal data. It combines a standard cumulative model with a discrete Uniform distribution used to take into account the uncertainty in the rating process. Since the posterior distribution of the parameters of interest is not in a closed form, MCMC methods are used to simulate from it. The performance of the proposed methodology has been evaluated via real data offering practical suggestions for using this approach in social-science, medicine or economic settings when an ordinal response is provided.

Keywords: Bayesian inference; CUP model; Ordinal data.

1 Method and setting

The extension of the class of cumulative models by introducing a component of uncertainty to improve the fitting and the interpretation of response process on ordinal scale has been recently proposed by Tutz *et al.* (2017). This new type of model combines, via a mixture, the standard cumulative model with a component gathering the uncertainty in the rating process. The mixture has been denoted as the CUP model that is a Combination of a discrete Uniform distribution and a Preference component.
Formally, let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ be a random sample generated by an ordinal random variable on the support $\{1, \ldots, m\}$, where m is a known integer. We interpret Y_i as the rating/preference expressed by the i-th subject about a specific item. For each i-th subject, we collect information $(y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \ldots, n$, where y_i is the observed value of the rating and

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

$\boldsymbol{x}_i$ is a row vector of the matrix $\boldsymbol{X}$ including a suitable set of covariates. The stochastic and systematic components of a CUP model for $j = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, n$ are

$$\begin{cases} Pr\left(Y_i = j \mid \boldsymbol{x}_i\right) = \pi_i \left[F(\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}) - F(\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta})\right] + (1 - \pi_i)\, p_j^U ; \\ \\ \pi_i = \dfrac{\exp(\boldsymbol{z}_i\boldsymbol{\gamma})}{1 + exp(\boldsymbol{z}_i\boldsymbol{\gamma})}, \end{cases}$$

where $F(\cdot)$ is the inverse of a suitable *link* function and $\boldsymbol{z}_i$ is a row vector of the matrix of covariates $\boldsymbol{Z}$. The vector $\boldsymbol{z}_i$ may have some elements in common with $\boldsymbol{x}_i$, that is some covariates can play a rule in both parts of the model. We refer to $\mathcal{I} = (\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = ||\mathcal{I}_i||_{i=1,\ldots,n}$ as the *information set*, where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$. Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$ are the vectors of regression coefficients and $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_{m-1} < \alpha_m = +\infty$ are the thresholds of the latent variable Y^* surrounding the observed discretized Y. The logistic link function is commonly used, yielding

$$\begin{aligned} logit\left[Pr(Y_i \leq j|\boldsymbol{x}_i)\right] &= F^{-1}\left(Pr\left(Y_i \leq j|\boldsymbol{x}_i\right)\right) \\ &= \alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}\,, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{1}$$

From equation (1), we obtain

$$\log\left[\frac{Pr\left(Y_i \leq j|\boldsymbol{x}_i\right)}{Pr\left(Y_i > j|\boldsymbol{x}_i\right)}\right] = \alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}\,.$$

The systematic part of the model saves the traditional definition of a predictor -as in cumulative models- but considers as well parameters π_i -as in the family of CUB models, see Piccolo (2003)- to weight for the uncertainty component.

Model (1) implies a constant relationship between the cumulative probability and the covariates. For given $\boldsymbol{x}_i$, the logit is altered only by the intercepts α_j which are different for each category $j = 1, 2, \ldots, m-1$. This is known in the literature as the proportional odds model. The names derives form the fact that log-odds ratio for two sets of explanatory variables depends only on the distance between them. Alternatively, in the CUP family several different models as adjacent categories or continuation ratio models may be considered.

Tutz *et al.* (2017) emphasize the ability of this class of models to capture multimodality and empirical overdispersion with a better fitting than cumulative and CUB models which are traditionally used in the ordinal data context. They also underline the added value of considering the uncertainty part in the process of analysis. A recent extension replaces the Uniform distribution by a more flexible one which is centered in the middle of the response categories. The resulting model allows to distinguish between a tendency to middle categories and a tendency to extreme categories by taking into account different response styles (Tutz and Schneider, 2019).

2 Bayesian Inference

In a Bayesian perspective suitable prior distributions should be assigned on the parameters of interest. We assign independent normal priors on each element of the vector $\boldsymbol{\beta}$ ($\beta_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_B^2)$, for any $k = 1, \ldots, p$) and each entry of the vector $\boldsymbol{\gamma}$ ($\gamma_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_G^2)$, for any $k = 1, \ldots, q$). In order to maintain the stochastic ordering of the intercepts we let $\alpha_1 \sim \mathcal{N}(0, \sigma_A^2)$ and

$$(\alpha_j | \alpha_{j-1}) \sim \mathcal{N}(0, \sigma_A^2) I(T_{j-1}, \infty),$$

for $j = 2, \ldots, m-1$, where $I(T_{j-1}, \infty)$ signifies that the distribution is truncated in the region (T_{j-1}, ∞) (i.e. it is a lower-truncated normal distribution) with $T_{j-1} = \alpha_{\alpha_{j-1}}$. As an alternative to the previous approach one can use doubly-truncated normal priors (Congdon, 2005) or an ordered Uniform distribution (Ishwaran, 2000). A more sophisticated approach can be obtained re-parameterising the model mapping the constrained parameters $\boldsymbol{\alpha}$ to a set of unconstrained variables $\boldsymbol{\xi}$, on which we can assign a suitable prior distribution (see for example Fahrmeier and Tutz, 1994; Albert and Chib, 1997).

We assume that $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ parameters are a priori independent. The prior distributions $\mathcal{P}(\boldsymbol{\beta})$ and $\mathcal{P}(\boldsymbol{\gamma})$ are defined in $(-\infty, \infty)$ whereas

$$\mathcal{P}(\boldsymbol{\alpha}) = \mathcal{P}(\alpha_1) \prod_{j=2}^{m-1} \mathcal{P}(\alpha_j | \alpha_{j-1});$$

thus $\mathcal{P}(\alpha_j | \alpha_{j-1})$ is defined in the range (α_{j-1}, ∞) for $j = 2, \ldots, m-1$ (see Congdon, 2005, among others). This ensures stochastic ordering for any values of $\boldsymbol{\alpha}$.

Given a sample of n respondents, the posterior distribution of the parameters of the model is given by

$$\mathcal{P}\big((\boldsymbol{\alpha}, \beta, \gamma) | \mathcal{I}\big) \propto L(\boldsymbol{\alpha}, \beta\gamma; \mathcal{I}) \mathcal{P}(\boldsymbol{\alpha}) \mathcal{P}(\boldsymbol{\beta}) \mathcal{P}(\boldsymbol{\gamma}) \tag{2}$$

where $L\big((\boldsymbol{\alpha}, \beta\gamma); \mathcal{I}\big)$ is the likelihood function (see Tutz *et al.* 2017) and $\mathcal{P}(\boldsymbol{\alpha}), \mathcal{P}(\boldsymbol{\beta})$ and $\mathcal{P}(\boldsymbol{\gamma})$ are the prior distributions described above.

Since the posterior distribution is not in a standard form we rely on an MCMC sampler. Once a starting value of the parameter vector has been provided we iteratively sample from the joint posterior distribution by means of a three step Metropolis-Hasting algorithm (update α, udate β, update γ); see Iannario and Tarantola (2020) for the details.

Normal random walk proposal distributions for each β_k are chosen

$$q(\beta_k^{(t)} \to \beta_k^{(t+1)}) : \beta_k^{t+1} | \beta_k^t \sim \mathcal{N}(\beta_k^t, \sigma_{\mathcal{P}B}^2),$$

where $\sigma_{\mathcal{P}B}^2$ is the proposal variance. Same consideration for γ_k where random walk proposal distributions are selected

$$q(\gamma_k^{(t)} \to \gamma_k^{(t+1)}) : \gamma_k^{t+1} | \gamma_k^t \sim \mathcal{N}(\gamma_k^t, \sigma_{\mathcal{P}G}^2),$$

where $\sigma^2_{\mathcal{P}G}$ is the related proposal variance.
For the intercepts, α_j, truncated uniform random-walk proposals are picked, such that $q(\alpha_k^{(t)} \to \alpha_k^{(t+1)}) : \alpha_j^{(t+1)} | \boldsymbol{\alpha}(t)$ yields

$$\begin{cases} \mathbb{U}\left(\alpha_j^{(t)} - \tau_\alpha, min\left[\alpha_j^{(t)} + \tau_\alpha, \alpha_{j+1}^{(t)}\right]\right); & if \quad j = 1, \\ \mathbb{U}\left(max\left[\alpha_j^{(t)} - \tau_\alpha, \alpha_{j-1}^{(t)}\right], min[\alpha_j^{(t)} + \tau_\alpha, \alpha_{j+1}^{(t)}\right]\right); if & \quad j = 2, \ldots, m-2, \\ \mathbb{U}\left(max\left[\alpha_j^{(t)} - \tau_\alpha, \alpha_{j-1}^{(t)}\right], \alpha_j^{(t)} + \tau_\alpha\right); & if \quad j = m. \end{cases}$$

where $\tau_\alpha > 0$ controls the size of the maximum unconstrained move away from the current value at each iteration.
An illustrative example includes a CUP model of mental health problems and depressive symptoms in later life (SHARE data available at http://www.share-project.org/home0.html). It represents a challenging application of the CUP model and underlines the potentiality of the Bayesian approach.

References

Albert, J. H. and Chib, S. (1997). Bayesian methods for cumulative, sequential and two-step ordinal data regression models. *Technical report*.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. Wiley.

Fahrmeier, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer.

Iannario, M. and Tarantola, C. (2020). A Bayesian Mixture Modeling for rating data with an uncertainty component. *Manuscript*.

Ishwaran, H. (2000). Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. *The Canadian Journal of Statistics*, **28**, 715 – 730.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85 – 104.

Tutz, G., Schneider, M., Iannario, M. and Piccolo, D. (2017). Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification,* **11**, 281 – 305.

Tutz, G. and Schneider, M. (2019). Flexible uncertainty in mixture models for ordinal responses. *Journal of Applied Statistics*, **46**, 1582 – 1601.

The determinants of discards in fisheries: A country approach with GAMs methodology

Belén Inguanzo[1], María-José Gutiérrez[1], Susan Orbe[2]

[1] University of the Basque Country (UPV/EHU). FAEII and BiRTE. Avd Lehendakari Aguirre 83, 48015 Bilbao, Spain
[2] University of the Basque Country (UPV/EHU). Applied Economics III (Econometrics and Statistics) and BiRTE. Avd Lehendakari Aguirre 83, 48015 Bilbao, Spain

E-mail for correspondence: belen.inguanzo@ehu.eus

Abstract: Discards are defined as the fishery catches returned to the waters, dead or alive. The aim of this analysis consists on identifying the main socioeconomic, technical and biological variables explaining the discards produced by countries from 1962 to 2013. For this purpose, the analysis relies on General Additive Models (GAMs). The significance of the economic, technical and biological variables on the estimations of this analysis shows the complex nature of discards, intending to provide a broad idea on the additional aspects that should be kept in mind when designing effective international policies to minimize discards.

Keywords: Fisheries conservation; Discards; General Additive Models; Ordinary Least Squares.

1 Introduction

The social benefits derived from fishing activity are associated to increasing collateral damages. One of these damages are the discards. Discards refer to the organisms of both commercial and non-commercial value that are caught during commercial fishing operations and returned to the sea, often dead or dying (Feekings *et al.*, 2012).
Despite being reintroduced in the trophic chain as food for other species, discards may compromise the multi-species balance of ecosystems by altering the different fish stock sizes or modifying the features of the environment when they decompose (Clucas, 1997). Ethically, discards can be seen as wasted products that could be consumed or used otherwise (Blanco *et al.*, 2007, Diamond and Beukers-Stewart, 2011).

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

This analysis aims to contribute to the elaboration of effective international initiatives targetting the minimization of discards by providing a basic idea on the socioeconomic, technical and biological factors underlying this type of catches at a global level.

2 Methodology and data

Initially, GAMs (Hastie and Tibshirani, 1990) are used to observe how socioeconomic and technological variables (the relative value of discards, the gear composition of the fishing fleet, the harvested areas, the fish demand, the alternatives to fish production and the economic size of the countries) determine the level of countries' discards. The mgcv package (Wood, 2017) in R (R Core Team, 2017) is used for its estimation, assuming a Gaussian family and an identity link function.

Once the unexplained trend in discards is obtained, the analysis uses an Ordinary Least Square regression to estimate the impact of biological and anthropogenic variables (anomalies in the global Sea Surface Temperature, anomalies in the global ocean heat and the lagged global catches) on the evolution of discards.

Data on the catches of countries is taken from the Sea Around Us project (Pauly *et al.*, 2015). Information on the aquaculture production, the fish consumption and exports is provided by the FAO (2019a, 2019b and 2019c). The Gross Domestic Product of countries in constant 2010 US dollars is extracted from the World Bank (2019). The EPA (2019 a and b) is the source for the anomalies in sea surface temperature and the ocean heat.

3 Results

TABLE 1. Determinants of discards.

Dependent variable: Discards	Estimate	Std. Error	T-value	P-value
Intercept	140468.0598	20967.2320	6.6994	< 0.0001
Relative value of discards	-13779.4028	2389.8734	-5.7657	< 0.0001
Fish exports	-0.0283	0.0038	-7.3706	< 0.0001
Percentage of landings from EEZ	-763.4189	207.1190	-3.6859	0.0002
Landings group1	0.0468	0.0029	16.1823	< 0.0001
Landings group2	0.3627	0.0052	69.2179	< 0.0001
Landings group3	0.3051	0.0545	5.6010	< 0.0001
Landings group4	-1.9062	1.1507	-1.6566	0.0977
Fish consumption	-0.0579	0.0013	-43.0788	< 0.0001
Aquaculture	-0.0009	0.0001	-7.6204	< 0.0001

Table 1 and Figure 1 present the estimation of the GAM. In order to interpret the impact of each variable, it is assumed that the remaining factors remain constant (*ceteris paribus*). The negative sign in the coefficient of the relative value reflects the economic incentives of fishermen to land

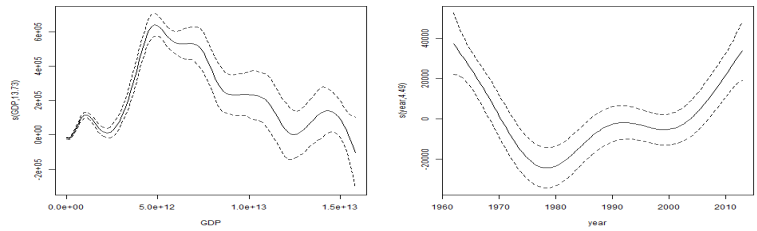


FIGURE 1. GDP and time effects on discards

more profitable catches. The negative signs of *exports*, *consumption* and *aquaculture* highlight the importance of the fish demand in reducing the discards of countries. Supporting the theory in the tragedy of the commons, harvesting in coastal waters is associated with the production of lower amounts of discards as implied by the negative sign of the percentage of landings coming from EEZ. Being composed by the gears producing the largest discards, the effect on discards of increasing the landings from the second group (bottom and pelagic trawls, dredge and long distance small scale) is the largest one. The effect of *GDP* on discards changes depending on its level. The evolution of *year* indicates that there have been global factors contributing to the evolution of discards from 1962 to 2013. The impact of these factors has varied over the period analyzed.

TABLE 2. Determinants of the trend in discards.

Dependent variable: Trend in discards	Estimate	Std. Error	P-value
Intercept	$3.5635e+04$	$2.1109e+04$	0.0946
SST anomaly	$-4.3872e+04$	$1.5411e+04$	0.0054
Heat anomaly	$5.0873e+03$	$6.4884e+02$	$6.165e-12$
Lagged catches	$-4.6241e-04$	$2.0856e-04$	0.02897

Table 2 shows the influence of the variables included in the OLS regression on the evolution of discards. In order to interpret the impact of each variable, it is assumed that the remaining factors remain constant. The response of the discards trend to changes in the temperature of waters depends on the depth. While anomalies in the ocean heat produce variations of the same sign in the discards trend, anomalies in the sea surface temperature cause variations of the opposite sign. The negative sign of lagged catches reflects that larger exploitation may diminish the future size of fish stock, decreasing the possibilities to high grade or even reducing the fishing activity.

Acknowledgments: This work was funded by the the Basque Government (BiRTE, IT-1336-19). Gutiérrez and Inguanzo also acknowledges the financial support from the Spanish Ministry of Economy, Industry and Competitiveness (ECO2016-78819-R, AEI/FEDER, UE).

References

Blanco, M., Sotelo, C.G., Chapela, M.J. and Pérez-Martín (2007). Towards sustainable and efficient use of fishery resources: present and future trends. *Trends in Food Science & Technology*, **18**, 29 – 36.

Clucas, I. (1997). A study of the options for utilization of bycatch and discards from marine captures fisheries. *FAO Fisheries Circular*.

Diamond, B. and Beukers-Stewart, B.D. (2011). Fisheries Discards in the North Sea: Waste of Resources or a Necessary Evil?. *Reviews in Fisheries Science*, **19**, 231 – 245.

Environmental Protection Agency (EPA) of the United States (2019a). *Climate Change Indicators: Ocean Heat*.

Environmental Protection Agency (EPA) of the United States (2019b). *Climate Change Indicators: Sea Surface Temperature*.

Feekings, J., Bartolino, V., Madsen, N. and Catchpole, T. (2012). Fishery Discards: Factors Affecting Their Variability within a Demersal Trawl Fishery. *PLoS ONE*, **7**, 1 – 9.

Food and Agriculture Organization of the United Nations (2019a). *Global Aquaculture Production*. Fisheries and Aquaculture Department.

Food and Agriculture Organization of the United Nations (2019b). *Consumption of Fish and Fishery Products*. Fisheries and Aquaculture Department.

Food and Agriculture Organization of the United Nations. (2019c). *Fishery Commodities and Trade*. Fisheries and Aquaculture Department.

Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*. John Wiley & Sons, Inc.

Pauly D., Zeller D., Palomares M.L.D. (Editors) (2020). *Sea Around Us Concepts, Design and Data*. seaaroundus.org.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

World Bank (2017). *World Development Indicators*. The World Bank Group.

Goodness of fit for complete and right-censored data. The R package `GofCens`.

Klaus Langohr[1], Mireia Besalú[2], Guadalupe Gómez Melis[1]

[1] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain
[2] Department of Genetics, Microbiology and Statistics, Universitat de Barcelona, Barcelona, Spain

E-mail for correspondence: `klaus.langohr@upc.edu`

Abstract: We present the R package `GofCens`, which implements both graphical tools and statistical tests to check the goodness of fit of parametric models for complete and right-censored data.

Keywords: Censored data; Goodness of fit; R package.

1 Introduction

Goodness-of-fit techniques are important to test the validity of parametric models and to provide indications that the modeling assumptions are reasonable. As an example, consider the accelerated failure time model:

$$Y = \log(T) = \mu + \boldsymbol{\beta}'\boldsymbol{X} + \sigma W, \tag{1}$$

where T is a possibly right-censored survival time, $\boldsymbol{\beta}$ and $\boldsymbol{X}$ are the parameter and covariate vectors, σ is the scale parameter, and W is the error term distribution, which is determined by the parametric choice for T. For example, if T follows a Weibull distribution, W is the standard Gumbel distribution. The validity of the model-based inference relies on this parametric assumption.
To check the parametric assumption in (1) based on a sample $(y_i = \log(t_i), \boldsymbol{x}_i)$, $i = 1, \ldots, n$, the following residuals can be used:

$$r_i = \big(y_i - (\hat{\mu} + \hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)\big)/\hat{\sigma}.$$

Residuals r_i are right-censored whenever t_i is right-censored. While methods are well developed for complete data, research for right-censored data

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

is still ongoing and, in both cases, integrated tools to check the goodness of fit of model (1) are needed. The R package `GofCens`, which is developed by the authors and will soon be available on CRAN, provides a large variety of goodness-of-fit methods for complete and right-censored data.

2 Goodness-of-fit methods

2.1 Graphical tools

Both probability and cumulative hazard plots are useful tools to check if a certain distribution is an appropriate choice for the data at hand.
The Probability-Probability plot (P-P plot) depicts the empirical distribution, $\widehat{F}(t)$, which is obtained with the Kaplan-Meier estimator if data are right-censored, versus the theoretical cumulative distribution function (cdf), $\widehat{F}_0(t)$. If the data come from the chosen distribution, the points of the resulting graph are expected to lie on the identity line.
The Stabilised Probability plot (SP plot), which is a transformation of the P-P plot, stabilises the variance of the plotted points. If $F_0 = F$ and the parameters of F_0 are known, $\widehat{F}_0(t)$ corresponds to the cdf of a uniform order statistic, and the arcsin transformation stabilises its variance. If the data come from distribution F_0, the SP plot will resemble the identity line.
The Quartile-Quartile plot (Q-Q plot) represents the sample quantiles versus the theoretical ones, that is, it plots t versus $\widehat{F}_0^{-1}(\widehat{F}(t))$. Hence, if F_0 fits the data well, the resulting plot will be a straight line.
A drawback of the Q-Q plot is that the plotted points are not evenly spread. Waller and Turnbull (1992) proposed the Empirically Rescaled plot (EP plot), which plots $\widehat{F}_u(t)$ against $\widehat{F}_u(\widehat{F}_0^{-1}(\widehat{F}(t)))$, where $\widehat{F}_u$ is the empirical cdf of the points corresponding to the uncensored observations. Again, if F_0 fits the data well, the ER plot will resemble the identity line.
Like probability plots, cumulative hazard plots can be used to assess the goodness of fit of a distribution. These plots are based on a transformation of the cumulative hazard function, $A(\Lambda(t))$, that is linear in either t or $\log(t)$. Complete and right-censored data are used to compute the Nelson-Aalen estimator of $\Lambda(t)$, but $A(\hat{\Lambda}(t))$ versus either t or $\log(t)$ is plotted only with the values of the uncensored observations. If the data come from the distribution under study, the points are expected to lie on a straight line.

2.2 Statistical tests

No general asymptotic optimality theory exists for this very difficult problem (Lehmann and Romano (2005)); in fact, any test can achieve high asymptotic power or perform uniformly well against local or contiguous alternatives when the family of possible alternatives is large (Janssen (2000)). Kolmogorov-Smirnov and chi-squared goodness-of-fit tests encompass the

most used analytical tests. However, due to the lack of good power of these tests, graphical techniques should be used together with these tests.
Goodness-of-fit tests have been developed for complete data and are based either on the empirical distribution function or on chi-squared-type tests. Preliminary extensions to account for right-censored data were proposed by Barr and Davidson (1973), who modified Kolmogorov-Smirnov statistics for censored or truncated data. Koziol and Green (1976) developed Cramér-von Mises-type statistics based on the product-limit empirical distribution function when the data are subject to random censorship.

3 The R package `GofCens`

The R package `GofCens` provides functions for 10 different laws (normal, logistic, Gumbel, lognormal, log-logistic, Weibull, exponential, beta, exponential power, and exponentiated Weibull) that perform the following goodness-of-fit methods:

- Probability and quantile plots: function `probPlot`.
- Cumulative hazard plots: function `cumHazPlot`.
- Kolmogorov-Smirnov test: function `KScens`.
- Kolmogorov-Smirnov, Cramér-von Mises, and Anderson Darling tests: function `gofcens`.

All functions can be used with complete and right-censored data, and provide the parameter estimates of all distributions. For this purpose, the package takes advantage of the `fitdistcens` function of the `fitdistrplus` package (Delignette-Muller and Dutang (2015)).
An example of the function `KScens` with right-censored survival times of patients who had suffered a myocardial infarction is shown in the following output. The data are from the Worcester Heart Attack Study (available at `ftp://ftp.wiley.com/public/sci_tech_med/survival`).

```
> KScens(whas500$lenfol,whas500$fstat,"weibull")$test
KS      p.value
1.41197 0.03599
> KScens(whas500$lenfol,whas500$fstat, "loglogistic")$test
KS      p.value
1.58043 0.01318
> KScens(whas500$lenfol,whas500$fstat, "gumbel")$test
KS        p.value
37.61688  0.00000
```

Among the three distributions compared —Weibull, loglogistic, and Gumbel distribution—, the Weibull law seems the most reasonable model given that the value of the test statistic (`KS`) is the smallest.

To confirm that the Weibull distribution is a good parametric choice, we apply the `probPlot` function, which draws four probability plots. According to these plots shown in Figure 1, this distribution, indeed, seems to be a good choice.

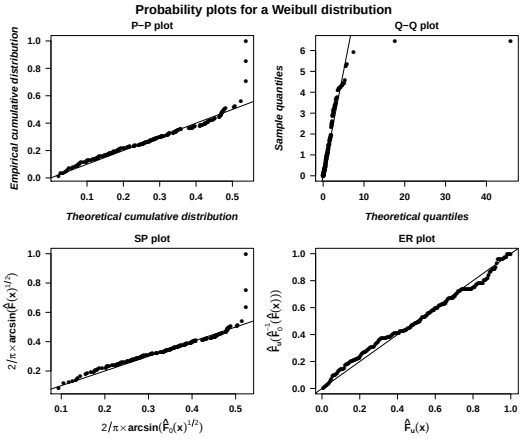


FIGURE 1. Probability plots drawn with function `probPlot`.

Acknowledgments: Grants MTM2015-64465-C2-1-R (*Ministerio de Economía y Competividad*, Spain) and 2017 SGR 622 (*Departament d'Economia i Coneixement, Generalitat de Catalunya*, Spain).

References

Barr, D.R. and Davidson, T. (1973). A Kolmogorov-Smirnov Test for Censored Samples. *Tecnometrics*, **15**, 739 – 757.

Delignette-Muller, M. and Dutang, C. (2015). `fitdistrplus`: An R Package for Fitting Distributions. *Journal of Statistical Software*, **64**, 1 – 34.

Janssen, A. (2000). Global power functions of goodness of fit tests. *Annals of Statistics*, **28**, 239 – 253.

Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses*. New York: Springer-Verlag.

Koziol, J.A. and Green, S.B. (1976). A Cramér-von Mises Statistic for Randomly Censored Data. *Biometrika*, **63**, 465 – 474.

Waller, L. and Turnbull, B. (1992). Probability plotting with censored data. *American Statistician*, **46**, 5 – 12.

Early diagnosis of sepsis from clinical data using the competing risk approach

Xinyi Liu[1], Ardo van den Hout[1]

[1] University College London, United Kingdom

E-mail for correspondence: xinyi_liu.17@ucl.ac.uk

Abstract: Sepsis is one of the leading causes of death in the hospitals and it is of great importance to diagnose sepsis as early as possible. In this work we investigate the early diagnosis of sepsis using the approach of survival analysis. In particular, we described 'sepsis' and 'nonsepsis' as two competing events and modeled the disease progression using a multi-state model.

Keywords: Muti-state model; Survival analysis; Early diagnosis .

1 Introduction

Sepsis is a life-threatening condition that occurs when the body overreacts to an infection and will cause tissue damage, organ failure, or death. Early detection and antibiotic treatment of sepsis are critical for improving outcomes for patients with sepsis. Therefore there exists a need to detect and treat sepsis as early as possible.

It is common to treat the sepsis early detection problem as a classification problem and solve it using the machine learning techniques. In works by Morrill et al. (2019) and Chang et al. (2019), complicated feature engineering methods were proposed and established classification algorithms were directly applied. But these techniques lack the explanabilty.

In this work we solved the classification of 'sepsis' and 'nonsepsis' using survival analysis. More details about the relationship between the classification problem and the survival analysis was discussed by Ripley and Ripley (2001). We modeled the the two events 'diagnosed to be sepsis' and 'diagnosed to be nonsepsis' as two competing events. In particular, we used a multi-state model to describe the disease progression of a patient from entering the hospital to the time the events occurs.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Data

The dataset used in this study are the clinical data of ICU patients from two hospitals, with observations of 20,336 and 20,000 patients. The dataset is available from the Physionet Challenge 2019.
Observations of each patient i are denoted by $[X_{i1}, \ldots, X_{i,t}, \ldots, X_{i,T_i}, Z_i]$, where $Z_i \in \mathcal{R}^{6\times 1}$ represents the demographic variables of the patient, and $X_{i,t} \in \mathcal{R}^{34\times 1}, t = 1, 2, 3, \ldots, T_i$ represents the time-varying covariates measured hourly. Moreover, for each patient, there is a vector of label Y indicates the onset of sepsis at each hour.

3 Problem Description

The aim of this study is to develop a method which could: identify the risk of sepsis for a patient at each hour $t, 0 < t < T_i$; and make a (0 or 1) prediction of the label $y_{i,t}$ at each hour based on its historical data $[X_{i,1}, \ldots, X_{i,t}, Z_i]$.
In order to measure the performance of the prediction models in terms of both the accuracy and the capacity of early diagnose, a normalized utility score was proposed in the Pysionet Challenge:

$$U_{normalized} = \frac{U_{total} - U_{noprediction}}{U_{optimal} - U_{noprediction}} \tag{1}$$

$$U_{total} = \sum_{i=1}^{N} \sum_{t=1}^{T_i} U(i,t) \tag{2}$$

For non-septic patients, the reward for the true negative prediction is 0 and the penalty for the false positive is 0.05. For the septic patients, sepsis prediction 12 hours before the onset of sepsis is slightly penalized, but the nonsepsis prediction after the onset of sepsis is increasingly penalized. Figure 1 shows an example of the utility function of a septic patient.

4 Multi-state model for sepsis early diagnose

In survival analysis, typically the censoring is assumed to be non-informative. However in our problem the censoring is informative because patients dropped out once they were diagnosed to be non-septic. To tackle this problem, we proposed to model the diagnosis of 'sepsis' and 'nonsepsis' as two competing events. Furthermore, it is natural to model the competing risks using the multi-state model, as discussed by Andersen et al. (2002).
In our multi-state model we defined three states (displayed in Figure 2): State 1 - under risk; State 2 - onset of sepsis; State 3 - diagnosed to be non-septic/dropout. Transitions are permitted are from State 1 to State 2,

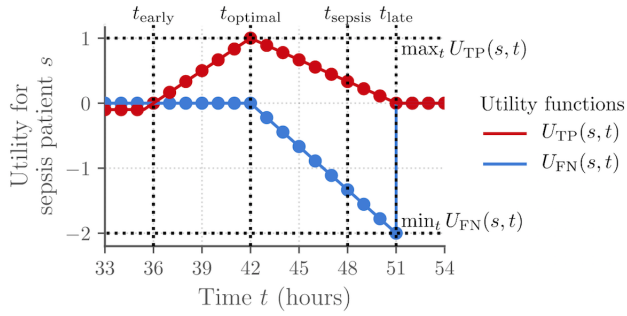


FIGURE 1. Utility Functions for a septic patient with the onset time of sepsis $t_{sepsis} = 48$ hrs, adopted from Reyna et al. (2019).

and State 1 to State 3. The model can be described using the transition probabilities $p_{rs}(t_i, t_j)$, which is the probability that an individual moves from the state r to state s within the time interval (t_i, t_j).
For the prediction, according to the definition of the utility score above, it is clear that the reward of the prediction depends on whether a patient is a septic patient or not, furthermore it is related to the onset time of the sepsis. Therefore it is natural to introduce a time-varying loss. We proposed that we can make predictions which minimize the expected total loss:

$$\begin{aligned} E[L|\mathcal{H}(t)] \quad &= \sum_{t_{sep}=t+1}^{\infty} L_+[P_{1,2}(0, t_{sep}) - P_{1,2}(0, t_{sep}-1)] \\ &+ \sum_{t_{sep}=1}^{t} L_-^{(t)}(t_{sep})[P_{1,2}(0, t_{sep}) - P_{1,2}(0, t_{sep}-1)] \end{aligned} \tag{3}$$

Where $\mathcal{H}^i(t) = [X_{i,1}, \ldots, X_{i,t}, Z_i]$ represents the observations of the patient i until time t. t_{sep} is the time that the onset of sepsis appears. For simplicity the utility score function is used as the loss function. L_+ represents the false positive loss, and $L_-^t(t_{sep})$ represent the true negative loss at time t, if the sepsis onset time is t_{sep}. $P_{1,2}(0, t_{sep})$ is the transition probability from state 1 to state 2 during the time interval $(0, t_{sep})$
In the multi-state model, in order to incorporate the effects of covariates, the transition intensity function from the state r to state s is defined as

$$q_{rs} = q_{rs}^{(0)} \exp(\beta_{rs}^T X(t) + \alpha_{rs}^T Z + \gamma_{rs} t) \tag{4}$$

and the contribution to the log-likelihood of each patient i between the time interval t_j, t_{j+1} is therefore

$$\log L_{i,j} = (t_{j+1} - t_j) q_{S(t_j)S(t_j)} + \log q_{S(t_j)S(t_{j+1})} 1_{\{s(t_j) \neq s(t_{j+1})\}} \tag{5}$$

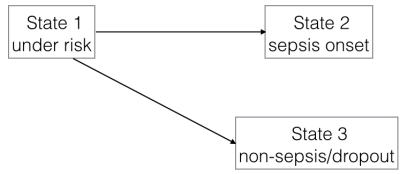


FIGURE 2. Transitions in the three-state model for the sepsis early diagnosis.

The parameters of this model is estimated by maximizing the likelihood function using the 'msm' package in R. For the predicting, since time-varying covariates were used, we assume that the transition intensities are piecewise constant and estimate the transition probability for each patient using the function provided in the 'msm' package.
However when estimating the transition probabilities $P_{rs}(t_i, t_j)$ for the time intervals while the observations $\mathcal{H}(t)$ are not yet available, i.e $t < t_j$, the above estimating function is not applicable. For the further work, we plan to extend the current model by incorporating the modeling of the time-varying covariates to improve the accuracy of the early diagnosis.

References

Morrill, J., Kormilitzin, A., Nevado-Holgado, A., Swaminathan, S., Lyons, T. (2019). The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. *Computing in Cardiology*.

Chang, Y., Rubin, J., Boverman, G., Vij, S., Rahman, A., Parvaneh, S. (2019). A Multi-Task Imputation and Classification Neural Architecture for Early Prediction of Sepsis from Multivariate Clinical Time Series. *Computing in Cardiology*.

Ripley, B. D., Ripley, R. M. (2001). Neural networks as statistical methods in survival analysis. *Clinical Applications of Artificial Neural Networks*, 237 255.

Reyna, M.A., Josef, C.S., Jeter, R., Shashikumar, S.P., Nemati, S., Westover, M.B., Clifford, G.D., Sharma, A.(2019). Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Critical Care Medicine*.

Andersen, P. K., Abildstrom, S. Z., Rosthj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research*, **11(2)**, 203 215.

Site 'dumpability': Where is illegal dumping in forests, and does signage help reduce it?

Samantha Low-Choy[12], Vicki Hall[34], Tobias Probst[35], Cameron Williams[16], Daniela Vasco[1]

[1] Griffith University, Mt Gravatt, Australia
[2] Environmental Futures Research Institute, Nathan, Australia
[3] Department of Environment and Heritage Protection, Brisbane, Australia
[4] Secretariat of the Pacific Regional Environment Programme, Fiji
[5] Department of Agriculture and Fisheries, Queensland, Australia
[6] Newcastle University, UK

E-mail for correspondence: s.low-choy@griffith.edu.au

Abstract: Previous research had identified factors affecting people's decisions to dump household waste, illegally, in forests. This information helped redesign signage aiming to deter dumping. The main question was: Where should signage be placed? From the perspective of state forest managers: What kinds of locations in the state forest are more 'dumpable' and tend to attract more illegal dumping than others? This study was one of the first to address the environmental context of dumping within a forest, which aggregates the psychological context of motivations of many individuals. Due to this novelty, we used expert elicitation techniques to formulate a conceptual model, which guided design of field data collection, both before and after introducing signage into the forest. Importantly, this also engaged a range of stakeholders with the project aims. Signage locations were pre-determined by regulators and forresters. Information from three phases of surveillance in the forest was analysed using a 'multimethod' statistical approach, starting with models more familiar to stakeholders: examining main effects via regression with smoothing splines; and high-order interactions via regression trees. Model-based clustering via Bayesian mixture models permitted insight directly relevant to the research questions, and found that: highly dumpable sites occurred both in close or far proximity to waste collection sites with varying profiles describing seclusion. Overall findings across model kinds confirmed that the 'nudging' sign motifs were most effective whilst evidence was conflicting regarding effectiveness of didactic signage.

Keywords: Mixture model; regression; multimethod; surveillance; Illegal dumping.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

1 Introduction

This project was motivated by the desire, by regulators and other stakeholders, to reduce illegal dumping of household waste in forests. Previous research (e.g. Marteau et al., 2011) into the motivations for dumping led to creation and community testing of several sign motifs, each designed to reduce dumping in forests in different ways: highest community preference (Owl motif); a didactic message with strong instructions (Stamp motif), 'nudging' messages noting surveillance (Camera, Report motifs), and aspirational messages encouraging protection of wildlife habitat (Fire and Home motifs). The challenge here was to evaluate how these new signs performed, which required intensive surveillance of the forest before and after sign installation.

The case study location was the Beerburrum Forest area, a pine plantation near the towns of Woodford, Beerburrum, Beerwah, Landsborough, and Caboolture. The property is operated by HQ Plantations (HQP) under a 99 year lease from Queensland Government. The area has numerous stakeholders and users. The research questions were:

RQ1 How do experts (with knowledge of illegal dumping in this forest) characterise sites where illegal dumping occurs?

RQ2 How does statistical analysis of field data describe site dumpability?

RQ3 Do signage interventions reduce illegal dumping?

Expert elicitation identified factors (RQ1) crucial for designing the field sampling protocol, to survey incidence of dumping before and after introduction of signage. Statistical models evaluated the effect of interventions (RQ3), whilst adjusting for site characteristics (RQ2) in different ways.

2 Conceptual model, via Expert elicitation

Expert elicitation was conducted in 2–3 hour sessions with different stakeholders: regulators in State Government; forresters in Beerburrum Forest area; city council officers involved in managing household waste; Queensland Police officers; recreational users of the forest. They were asked to brainstorm factors relating to dumpability of sites in Beerburrum, rank these factors, then characterise one or two key profiles of dumpable sites. This information was encoded in several waves. The final conceptual model is shown here.

3 Methods and Materials

Based on expert input, variables identified for data collection included: location; dumped material (type, scatter or freshness); proximity to waste

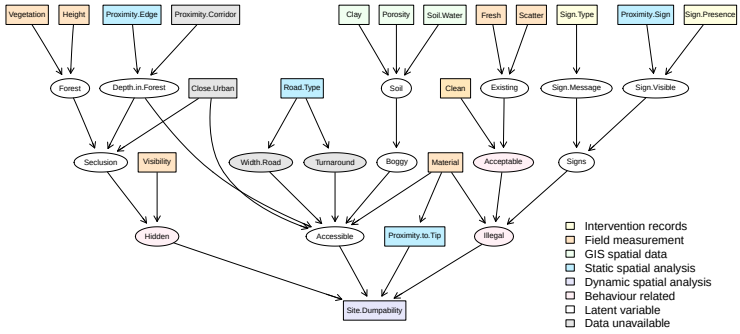


FIGURE 1. Detailed conceptual model: factors affecting site dumpability elicited from stakeholders with expertise on illegal dumping in Beerburrum Forest area.

collection ('tips') or to forest edge; vegetation type and horizontal visibility; soil type (not discussed here). Surveillance of the forest was conducted via vehicle on road sections, first at baseline, then in two phases after introducing signage, throughout 2 sectors of the park. Locations and numbers of 6 signage motifs in this pilot experiment were dictated by operational constraints, but distributed proportionally in 9 areas of the forest, always containing the community-preferred Owl design as a reference motif. Each signage motif had low replication, particularly in regard to combinations of site-specific environmental predictors. Thus a multimethod approach to modelling was used to examine effects of site characteristics and interventions on site-changes, so that different 'templates' (models) could describe effects in different ways. Analysis started with models familiar to stakeholders: examining main effects via regression with smoothing splines; and high-order interactions via regression trees. Imbalanced design made it difficult to interpret main effects or interactions. Model-based clustering via Bayesian infinite mixture models permitted insight relevant to research questions. We used Profile Regression in the PReMiuM package in R, suitable for categorical and continuous covariates (Liverani et al., 2015).

4 Results

Intensive spatial analysis was required to encode transects of search effort and note locations of dumpsites, leading to estimates of rates of dumping by road section. Changes experienced by road sections over a two-month period mostly ranged from 0 to 2 dumpsites per km (dpk), with some sites experiencing larger changes of approx. 5dpk. Overall evidence from modelling suggested that signs were plausibly slightly effective in general, reducing dumping by 0.3 dumpsites per km. Interpretation of the two simpler models (regression and regression trees) was far from straightforward

TABLE 1. Profiles of changes in rate of dumping (after 2mth) across dumpsites. Dumpsites were allocated with highest post. prob. via mixture model to one of fourteen profiles, with 2-mth change in rate of dumping as the response, sign type as the intervention, and all other variables defining clusters. Results shown for 4 profiles with highest dumprates, noting (**no.**) of sites, and posterior modes.

Cl		**Dumped material**			**Proximity to**			**Vege-**
#	**No.**	**Type**	**Scatter**	**Fresh**	**Tip**	**Edge**	**Visibility**	**tation**
	326	*Biggest increase in rate of dumping*						
16	10		↓Not		H			
15	26				VL			
14	35			↑Yes	L		VH	
13	255		↑Yes		L	VL	VL	↑Pine
								↑Grass

due to the imbalance of design; Bayesian mixture model results provided a useful alternative (Table 1).

Evidence regarding the most didactic Stamp design was conflicting: either found least (model-based clustering) or more effective (regression trees). Thus further studies are required with more nuanced design to control for important interactions. However the Camera and Report signs designed to nudge behaviours (Marteau et al., 2011) were consistently found more effective (flexible regressions, model-based clustering). Also the Home and Fire signs encouraging positive behaviours to protect wildlife were not found effective for protecting areas deep in the forest (regression tree). These comparisons were made possible due to the use of the Owl motif as a reference sign motif. Altogether this evidence suggests that signs may be effective, further work is required with more targeted experimental design informed by these results.

Acknowledgments: Special thanks to members of the Illegal Dumping unit in QEHP. This project was funded by a Research Contract through Griffith Enterprise, funded by QEHP.

References

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M. & Richardson, S. (2015). PReMiuM: An R package for profile regression mixture models using Dirichlet processes *Journal of Statistical Software*, 64(7), 130.

Marteau, T.M., Ogilvie, D., Roland, M., Suhrcke, M. & Kelly, M. P. (2011). 'Judging nudging: Can nudging improve population health?', *British Medical Journal*, 342: d228.

Relevance of Semantic-Enriched in Information Retrieval Models

Kenan M Matawie[1], Sargon Hasso[2]

[1] Western Sydney University, Australia
[2] Loyola University Chicago, UL LLC, Northbrook, IL USA

E-mail for correspondence: k.matawie@westernsydney.edu.au

Abstract: Improving document relevance in Information Retrieval has been recently the focus of many research projects and papers. Such investigations and developments are very helpful and essential for everyday private and commercial decisions making process. The main parts of this improvement are the scoring models such as BM25 and the evaluation of the performance of these techniques such as the rank-based models, e.g. MAP, NDCG and RBP. Here we are focusing on the semantic enrichment of the documents using specialised dictionary that will improve the score and rank of the search results. This enrichment is analysed and presented using TREC data and utilizing Lucene full-text search library.

Keywords: Relevance; IR models; Semantic Enrichment; BM25.

1 Introduction

One of the important stages of the Information retrieval (IR) process is indexing, where documents are analysed and indexed to be prepared for searching process. Relevance plays an essential role to determine the matching level and the order of the documents retrieved to satisfy user's query. Traditional relevance process is based on keyword searching and this can be enriched and further supported with semantic-based searching using synonym and specialised/custom thesauri. While the two approaches still use individual keywords or terms in the indexing and searching processes, the keyword-based attempts to match word for word, while the semantic-based will try to match a keyword against a set of keywords that are semantically related. The latter is achieved using a thesaurus. In this paper, we will discuss our findings using semantic-based searching using domain-specific thesaurus.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

In section 2, we briefly explain the scoring, ranking, and the evaluation model. In section 3, we describe a methodology we used in this research and we provide a brief description about the NASA thesaurus we used as a specialty thesaurus used for synonym expansion during indexing. An analysis and evaluation of the test results are discussed in section 4. We summarize and provide directions for future research in section 5.

2 Scoring, Ranking and Evaluation Models

We use Lucene (Apache Software Foundation) search engine to evaluate statistical information models with different indexing configurations. Lucene uses a scoring function to determine how relevant a document is to a given user's query. Lucene allows us to use one of several implementations of these scoring models. Our intent is figuring out how accurate, i.e. relevant, the retrieved documents (d) are to the user's information need (q). It is important to determine the contribution of the term, i.e. word, to the document. This is calculated by using language models based on a given d. Most of the models are based on the maximum likelihood estimate of the relative counts. However, we will only present one model here, i.e. best Match family (BM25) by Jones et al. (2000), since this is the scoring model we used for evaluation as implemented by Lucene:

$$f(q,d) = \sum_{w \in q \bigcap d} c(w,q) \frac{(k+1)c(w,d)}{c(w,d) + k(1 - b + b\frac{|d|}{avdl})} log \frac{M+1}{df(w)} \qquad (1)$$

where $b \in [0,1]$ is part of the *normalizer* term $1 - b + b\frac{|d|}{avdl}$; $avdl$ denotes average document length.
Various evaluation models are used and all are score/rank based functions. We used all these three evaluation models but here we will focus on Rank-Biased Precision (RBP) Moffat and Zobel (2008)as it is the most recent, suitable and effective precision evaluation model:

$$RBP_q = (1-p)\sum_{i=1}^{N} r_i p^{i-1} \qquad (2)$$

where r_i is the ith relevance judgement, i is the ith document rank (with 1 as the highest document rank), N is the number of documents and $p \in [0,1]$ is the probability function parameter.

3 Semantic Enrichment

In a previous work, Matawie and Hasso (2018), we have described the methodology we used to generate the results discussed in this paper. Briefly,

it consists of using elasticsearch to index TREC aviation data set NASA (2012).
To improve relevancy of the result sets returned by search engine, we used several other parameters that influence how the indexed documents are textually analyzed, stored, and returned during searching.
Finally, We used the *Rank-biased Precision* (RBP) mentioned above, as an evaluation criterion to measure the quality across recall levels among all algorithms, i.e. relevant and non-relevant as judged by human experts (in this paper assumed to be 0.8).
The use of thesaurus in search engine expands the search engine capability from term-centric to meaning-centric. The inclusion of thesaurus, which clusters words around concepts, in the search engine allows returning documents that are similar in meaning. In our previous work used a generic English language-based thesaurus, e.g. WordNet (2010). The generic language-based thesauri may have limited impact on searching highly specialized documents on subjects like aviation and medical texts. In the current research, we have used NASA Thesaurus (2012). It contains the authorized NASA subject terms used to index and retrieve materials in the NASA Technical Reports.

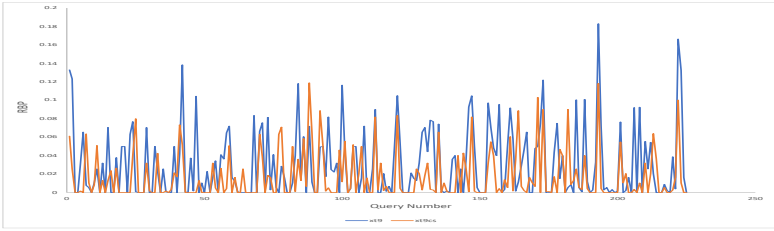


FIGURE 1. *RBP* averages of 225 Queries with and without enhancement (Orange and blue lines respectively, xt9-xt9cs). Title of the documents not included (xt9)

4 Analysis and Evaluation

We set out to examine the effect of using domain-specific thesaurus, i.e. NASA Thesaurus, on the relevancy of the retrieved documents during search. In general, augmenting the search engine with synonym expansion capability should improve the relevance of the retrieved documents.
Our test results, however, didn't improve the relevancy of the retrieved results as illustrated in Figures 1. In fact, it made it slightly worse. This is because, we believe, is a result of not using highly specialized thesaurus. We resorted to using it because it was the most accessible resource at this time.

Surely, it includes terms applicable to avionics, the subject covered in our TREC document set, and other terms from other engineering and scientific disciplines. The nature of synonym expansion process is that documents are enriched with a set of terms that may act as a noise and render a document as being irrelevant. As a result, the BM25 ranking function will penalize such documents because of artificially inflated document noise. In the absence of highly specialized thesaurus and as part of our continuing research, we are looking at a methodology that generates a corpus-specific thesaurus to be used as a source for synonym expansion capability of the search engine.

5 Conclusion

We have developed a methodology that comprises a testing platform using Elasticsearch engine to generate and evaluate the test results using *Rank-biased Precision* (RBP) evaluation model. We can configure the Elasticsearch engine using different analyzers, different lemmatizers, different thesauri for synonym expansion, and different ranking functions on the TREC aviation data set. When applied on the TREC aviation data set, we obtained different test results. The relevance of each returned search results was analyzed and compared using RBP evaluation criteria. This way, we can tell which search engine configurations gives us the closest match to the human experts evaluation criteria. In this paper, we evaluated a configuration based on the NASA thesaurus, i.e. specialized, that we adapted to work in Elasticsearch engine. We compared the results with similar configuration but with English-language thesaurus, i.e. generic. This is an ongoing research, more analysis and relevance modelling enhancement approaches will be included in the final version of this paper.

References

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.

Matawie, K. and Hasso, S. (2018) Evaluating Statistical Information Retrieval Models with Different Indexing Enhancement Strategies. In *Proceedings of the 33rd International Workshop on Statistical Modelling*, Bristol, UK, July 1520, 2018.

NASA (2012) *NASA Thesaurus: Hierarchical Listing with Definitions.* Vol 1, NASA.

Moffat, A., Zobel, J. (2008) *Rank-biased Precision for Measurement of Retrieval Effectiveness.* ACM Transactions on Information Systems (TOIS), 27(1).

Bayesian Spatial Conditional Overdispersion Models: Application to infant mortality

Mabel Morales Otero[1], Vicente Núñez-Antón[1]

[1] Department of Econometrics and Statistics, University of the Basque Country UPV/EHU, Bilbao, Spain

E-mail for correspondence: mabel.morales@ehu.eus

Abstract: In this work we revise Bayesian generalized conditional models for spatial count data with overdispersion. We show their usefulness by fitting them to infant mortality rates from Colombian regions. These models assume that the overdispersion present in the data may be caused partially from the spatial dependence that exists among the spatial units. Therefore, regression structures are specified both for the conditional mean and for the dispersion parameter, including also spatial neighborhood structures in the model. We work on the case of spatial count data which follow a Poisson distribution, and focus our attention on the spatial generalized conditional normal Poisson model. Models have been fitted with the use of the Markov Chain Monte Carlo (MCMC) algorithms within the context of Bayesian estimation methods.

Keywords: Bayesian models; Spatial models; Overdispersion; Count spatial data.

1 Introduction

When working with count data, it is common to use generalized linear models (GLM) to fit the distribution of the response variable. However, regression models for count data often present overdispersion, a phenomenon that arises when the real variance of the data is larger than the one specified in the model. This could cause the standard errors to be underestimated, resulting in an incorrect inferential process. One of the main causes for overdispersion is the possible existing correlation between the values of the response variable for the different units, which is very common in the presence of spatial data. Consequently, the spatial dependence that may exist among the different locations must be taken into account in order to

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

produce reliable inference processes from the estimations. We have fitted spatial generalized conditional overdispersion models, originally proposed by Cepeda-Cuervo et al. (2018). In these models, regression structures are specified both for the conditional mean and for the overdispersion parameter. The spatial dependence present in the data is captured by including a spatial lag in both regression structures. In this way, the researcher can have information about the type of spatial association that is present in the data. To show the usefulness of the aforementioned models, we apply them to a dataset including infant mortality rates from different regions of Colombia, as well as a number of variables that we will use as covariates.

2 Spatial generalized conditional overdispersion models

The most common approach for modelling overdispersion is to include an additional dispersion parameter in the GLM. In generalized overdispersion models for count data, regression structures are specified both for the conditional mean and for the overdispersion parameter. However, these models do not provide information about the strength of the spatial association present in the data. Spatial conditional overdispersion models were proposed by Cepeda-Cuervo et al. (2018) in order to be able to estimate this effect. This effect is modelled by proposing the use of the spatial weights matrix to compute the spatial lag of the variable under study, which is included in the model with a parameter that estimates the intensity of the spatial association. The spatial structure of the neighborhood is defined by the weights matrix $\mathbf{W} = [w_{ij}]$, with elements w_{ij} given by the weights that reflect the intensity of the dependence between regions i and j. In general, $w_{ij} = 1/n_i$, if region j belongs to the neighborhood of region i, with n_i being the number of first or second order adjacent regions for region i; and $w_{ij} = 0$ otherwise. The spatial conditional overdispersion regression model assumes that the spatial variable under study, $Y_i,\ i = 1, \ldots, n$, conditioned on the values of all of its neighbor, but not including the i-th region itself (i.e., $Y_{\sim i}$), has a overdispersed conditional distribution denoted by $f(y_i|y_{\sim i}),\ i = 1, \ldots, n$, where the conditional mean and the dispersion parameter follow given regression structures that, besides some covariates affecting the response variable, also include spatial lags of the variable under study. The conditional overdispersion density function follows either a Poisson or a binomial distribution, leading to the (generalized) spatial conditional Poisson, negative binomial, normal Poisson, binomial, beta binomial and binomial normal regression models, respectively.

Let us consider the Poisson normal model for overdispersed count data, in which the overdispersion is included in the model with the use of a normally distributed random effect term in the mean model. In this way,

$$g(\lambda_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \nu_i, \tag{1}$$

where $g(.)$ is usually the logarithm function, $\mathbf{x}_i$ is the $q \times 1$ vector of explanatory variables for the i-th observation, $\boldsymbol{\beta}$ is a $q \times 1$ vector of unknown regression parameters, and $\nu_i \sim N(0, \tau)$. In this model, $(Y_i|\nu_i),\ i = 1, \ldots, n$, follows a Poisson distribution with mean $\lambda_i = E(Y_i|\nu_i)$. In the spatial conditional normal Poisson model, if $Y_i,\ i = 1, \ldots, n$, represent area count data from different regions or areas, a portion of the existing overdispersion can be explained by the neighborhood spatial structure assumed by the researcher, so this model assumes that $(Y_i|Y_{\sim i}, \nu_i)$ follows a Poisson distribution with mean $\lambda_i = \mathrm{E}(Y_i|Y_{\sim i}, \nu_i)$, and

$$g(\lambda_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \rho\mathbf{W}_i\mathbf{y} + \nu_i, \tag{2}$$

with ρ being the parameter explaining the first order spatial association in the mean model, $\mathbf{W}_i$ is the i-th row of the $n \times n$ weight matrix $\mathbf{W}$, and $\mathbf{y}$ is the $n \times 1$ vector of the observed values of the response variable under study. Finally, in the generalized spatial conditional normal Poisson model, the conditional mean μ_i and the variance terms in the random effect distribution, σ_i^2's, follow regression structures given by

$$\log(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \rho\mathbf{W}_i\mathbf{y} + \nu_i \ \text{ and } \ \log(\sigma_i^2) = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \eta\mathbf{W}_i\mathbf{y}, \tag{3}$$

where $\mathbf{z}_i$ is the $q_\phi \times 1$ vector of explanatory variables for the i-th observation, $\boldsymbol{\gamma}$ is a $q_\phi \times 1$ vector of unknown regression parameters, and η is the parameter explaining the first order spatial association in the dispersion model.

3 Application

The dataset considered here has been obtained from the National Statistics Department of Colombia and corresponds to 32 departments (regions). Some of the variables available for each one of the geographical units are: infant mortality rate, which is the number of children under one year of age who died per 1000 born alive in 2005 (i.e., variable IMR), the percentage of the population that had basic services not being satisfactorily attended to for the year 2005 (i.e., variable NBI) and the resources (in thousands) provided for academic achievement or education and integral attention for young children per household in the year 2005 (i.e., variable Rec), among others. For the prior distributions, we assume independent normal distribution, $N(0, 10^5)$, for all of the regression parameters. In the specific application considered here, after 10000 iterations and a burn in period of 2000 samples, the chains showed strong signs of convergence. The best fitting model for this data was the generalized spatial conditional normal Poisson model with DIC and BIC values of 200.1 and 187.8, respectively, and mean and variance regression models given by:

$$\log(\mu_i) = \beta_0 + \beta_1\mathrm{Rec}_i + \nu_i, \quad \nu_i \sim N(0, \sigma_i^2) \tag{4}$$

$$\log(\sigma_i^2) = \gamma + \eta\mathrm{NBI}_i + \rho\mathbf{W}_i\mathbf{y}, \tag{5}$$

with the corresponding estimates reported in Table 1.

TABLE 1. Parameter estimates, together with their standard deviations for the generalized spatial conditional normal Poisson model fitted to the infant mortality data

	β_0	β_1	γ	η	ρ
Estimate	3.003	-1.294×10^{-03}	-12.679	0.112	0.203
SD	0.105	9.306×10^{-04}	4.265	4.267×10^{-02}	9.877×10^{-02}

4 Conclusions

We have reviewed generalized spatial conditional overdispersion models for count data and applied them to the study of infant mortality rates in the departments of Colombia. They have provided a good fit and were able to explain the overdispersion and spatial association present in the data. From our study, for the Poisson case, we can conclude that the proposed models fit better than other models that are not taking into account the overdispersion or the well known intrinsic conditional autoregressive (ICAR) models. More specifically, for the infant mortality rates data, the best fitting model was the generalized spatial conditional normal Poisson model.

Acknowledgments: This work was supported by Ministerio de Economía y Competitividad, Agencia Estatal de Investigación (AEI), Fondo Europeo de Desarrollo Regional (FEDER), and the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group) under research grants MTM2013-40941-P (AEI/FEDER, UE), MTM2016-74931-P (AEI/FEDER, UE), IT-642-13 and IT1359-19, as well as from Ministerio de Ciencia, Innovación y Universidades under Ayudas para contratos predoctorales para la formación de doctores (FPI) 2017, reference BES-2017-079940.

References

Cepeda-Cuervo, E., Córdoba, M. and Núñez-Antón, V. (2018). Conditional overdispersed models: application to count area data. *Statistical Methods in Medical Research*, **27(10)**, 2964 – 2988.

Hinde, J. and Demétrio, C.G.B. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, **27(2)**, 151 – 170.

Quintero-Sarmiento, A, Cepeda-Cuervo, E. and Núñez-Antón, V. (2012). Estimating infant mortality in Colombia: some overdispersion modelling approaches. *Journal of Applied Statistics*, **39(5)**, 1011– 1036.

A note on (basic) Principal Components Analysis

Vito M.R. Muggeo[1]

[1] Università di Palermo, Italy

E-mail for correspondence: vito.muggeo@unipa.it

Abstract: This paper provides some issues, known but somewhat little stressed, on using the conventional covariance or correlation matrix when performing the simple PCA. The paper also proposes a new simple alternative by providing some evidence, via real-data analysis and some simulation experiments, supporting the proposal.

Keywords: eigendecomposition, multivariate analysis.

1 Introduction

Principal Component Analysis (PCA) is probably one of the most old, known and widespread statistical tool of multivariate analysis across disciplines. PCA backdates to Pearson (1901) and Hotelling (1933) in the early twentieth century, and nowadays several authoritative books are available for graduate students and researchers working in different areas. An exhaustive listing of all textbooks discussing PCA is a tough task and practically unfeasible. We just refer to Jolliffe (2002) for a comprehensive and modern introduction and some extensions. While several and challenging extensions of PCA have been discussed, this note just aims at providing some thoughts and comments about computing the principal components in the simple case. We underline some issues which characterize the usual approaches to PCA and then we propose a new one which is based on a relatively little used variability measure which does not appear to have ever been discussed.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Methodology of PCA

The settings are the following: data refer to p numerical variables observed on n units: Let $x_1, \ldots x_j, \ldots, x_p$ the n-dimensional vectors arranged in the $n \times p$ data matrix X. Let $var(X) = S = [s_{jk}]$ the covariance matrix wherein $s_{jj} = s_j^2 = var(X_j)$ and $s_{jk} = cov(X_j, X_k)$. PCA aims to find the 'best' linear combination of observed variables; 'best' refers to capability to account for the *whole* variability in data as much as possible such that the found component has marginal variability larger than the single Xs. More formally, let $y = Xa$ the linear transformation of the Xs where the unknown $a = (a_1, \ldots, a_p)$ are found by maximizing $var(Xa) = a^T Sa$. Identifiability constraints, such that $a^T a = 1$, have to be added to make the problem well determined. The unique solution comes from the eigendecomposition of S, such that a is an eigenvector with corresponding eigenvalue $\lambda = var(Xa)$. The p eigenvectors are sorted according to the eigenvalues $\lambda_1^x \geq \lambda_2^x \geq \ldots \geq \lambda_p^x$ leading to the p uncorrelated principal components $y_1^x = Xa_1, y_2^x = Xa_2, \ldots, y_p^x = Xa_p$. If X does not include *redundant* columns, i.e. variables with no variability or expressed as linear combination of others, S has full rank, and $\lambda_j^x > 0$ for each j.

Altogether the p principal components explain the *whole* variability, sometimes referred as *inertia*, as expressed by $\text{trace}(S) = \sum_j s_j^2$, but typically just a few components are retained to account for most of the whole variability: the cumulative portion of variance explained by the first k, say, principal components $\sum_j^k \lambda_j / \text{trace}(S)$ is employed to assess how many principal components could be kept. Retaining few components with 'good' portion of explained variability, no less than 70% or 80% probably, is usually considered successful for PCA.

In practice, phenomena under investigations are very complex with relevant variables X_j observed on different scales or different units of measure. While eigendecomposition of S can be still carried out, the additive expressions for the principal components ($\sum_j a_j X_j$) themselves and the total variance ($\sum_j s_j^2$) get weird and difficult to admit from a substantive viewpoint, since these involve quantities on different units of measure. The usual recommendation reported in all textbooks is to consider the standardized variables $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$ being $\bar{x}_j$ the covariate mean and s_j the standard deviation. Hence, given the standardized data matrix $Z = [z_{ij}]$, one performs eigendecomposition of the corresponding covariance matrix $var(Z) = cor(X) = R = [r_{jk}]$ to get eigenvectors b_j and eigenvalues λ_j^z which, in turn, lead to components $y_j^z = Zb_j$ having variances $var(y_j^z) = \lambda_j^z$. Using the correlation matrix $R = [r_{jk}]$ allows to get dimensionless quantities which can be summed fairly.

While the canonical standardization 'z' allows to include variables on different scales straightforwardly, we guess the price to be paid is high: information about the marginal variability is destroyed, since each standardized variable has unit variance.

Actually we need a transformation making unitless the variables, while preserving information on the whole variability in data, namely covariabilities and marginal variances as well. The proposed transformation fulfilling the aforementioned points is

$$u_{ij} = \frac{x_{ij}}{|\bar{x}_j|} - \text{sgn}(\bar{x}_j) \qquad j = 1, 2, \ldots, p. \tag{1}$$

The transformed variables U_j have zero means with corresponding covariance matrix $V = [v_{jk}]$ such that

$$var(U_j) = v_j = \frac{s_j^2}{|\bar{x}_j|^2} \qquad \text{and} \qquad cov(U_j, U_k) = v_{jk} = \frac{s_{ij}}{|\bar{x}_j \bar{x}_k|}. \tag{2}$$

Namely the elements on the main diagonal are the coefficients of variations (cv) of the original X_j and the off-diagonal elements can be called 'coefficients of covariation'. Curiosly, while the cv is well known in explorative analysis as a measure of unitless variability, the corresponding 'covariability' does not appear to have been discussed. Similarly to the matrix of variance-covariance, we could name V the matrix of variation-covariation coefficients, or more simply the covariation matrix.
Our proposal is to run PCA via eigendecomposition of the matrix V leading to eigenvectors c_j and eigenvalues λ_j^u. Hence the resulting principal components are $y_j^u = Uc_j$ where U is the data matrix obtained via transformation (1), and $var(y_j^u) = \lambda_j^u$. Of course $\lambda_j^u / \sum_j \lambda_j^u$ is the portion of total variance explained by y_j^u. The intuition behind (1) is that *all* information about variability in data is exploited: correlation is preserved, but unlike the traditional standardization z, the likely different marginal variabilities are also involved in the determination of the principal components and corresponding variance. As a conseguence, we conjecture the first principal components based on V will capture larger portions of the whole variability of data.

3 Empirical evidence

To gain empirical evidence about using the covariation rather than the correlation, some simulation experiments were run: we generate n observations from p multinormal variables with means ranging in $(-10, 10)$, marginal standard deviations in $(1, 50)$ and two kinds of correlation matrix: Uniform ($\rho_{jk} = \rho$) and Toeplitz $\rho_{jk} = \rho^{|j-k|}$. Five values for $\rho = \{0.1, 0.2, 0.5, 0.7, 0.9\}$, three values for $p = \{5, 10, 25\}$ and three sample sizes $n = \{30, 50, 100\}$ leading to 90 scenarios, overall. We contrast performance of PCA based on matrices R and V via the differences between portions of total variance explained by the first principal components reported in Figure 1. Patterns are rather clear and easy to interpret: PCA

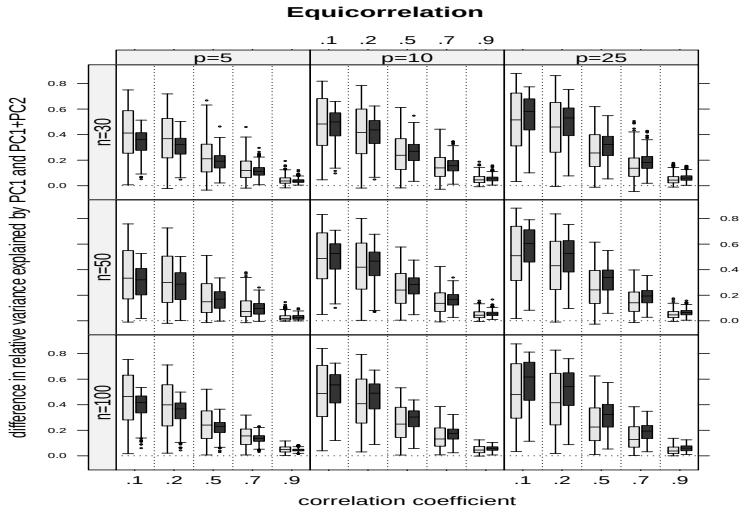


FIGURE 1. Differences between PCA based on V and R. The boxplots refer to differences of relative variances explained by PC1 (light grey box) and PC1+PC2 (dark grey box). Positive value indicate that PCA based on V is able to explain larger portions of relative inertia.

using the proposed scaling (1) is always able to explain higher portions of total variability than the conventional standardization, especially when only the first principal component is kept. While sample size does not matter as expected, outperformance gets higher when the number of involved variables increases and correlations lessen. When covariability in data is scarce, i.e. low correlation coefficients, the marginal variabilities get considerable and they matter in determining principal components able to 'grab' most of whole variability in data. When the correlation structure is strong, differences alleviate.

References

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosphical Magazine*, **2**, 559 – 572.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417 – 441.

Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer-Verlag, New York.

Cholesky-based multivariate Gaussian regression

Thomas Muschinski[1,2], Georg J. Mayr[2], Thorsten Simon[1,2], Achim Zeileis[1]

[1] Department of Statistics, Universität Innsbruck, Innsbruck, Austria
[2] Department of Atmospheric and Cryospheric Science, Universität Innsbruck, Innsbruck, Austria

E-mail for correspondence: Thomas.Muschinski@uibk.ac.at

Abstract: Multivariate Gaussian regression has applications in many fields, but is made difficult by the high model complexity and positive-definite requirement on the estimated covariance. We implement multivariate Gaussian regression through a Cholesky-based reparameterization of the covariance matrix. The distributional parameters—the means and the entries of the Cholesky factor—can be made to depend on covariates through flexible additive predictors, allowing for nonlinear variations in mean and covariance. The reparameterization is compared to reference methods for estimating a fixed covariance. An application for weather prediction (surface temperature) illustrates the flexibility of the approach.

Keywords: Covariance modeling; Cholesky decomposition; Multivariate Gaussian; MCMC simulation.

1 Cholesky-based multivariate Gaussian regression

Multivariate modeling has a wide range of applications from longitudinal analyses of biomarker data to postprocessing of numerical weather predictions. Employing multivariate Gaussian distributions in the framework of distributional regression allows one to specify very flexible models. For the bivariate Gaussian case, the correlation may be modeled directly (e.g. Klein et al. 2015), but for higher dimensions two main difficulties occur: (i) high complexity resulting from the large number of distributional parameters and (ii) ensuring a positive definite covariance. To tackle the latter issue, we factorize the covariance by the Cholesky decomposition (Pourahmadi 2011). To deal with its high complexity, we regularize the Cholesky-based multivariate Gaussian regression models (Umlauf et. al 2018).

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

The Cholesky decomposition of a positive definite symmetric matrix Σ has the form

$$\Sigma = LL^{\mathrm{T}} \qquad \text{and} \qquad \Sigma^{-1} = L^{-1\mathrm{T}}L^{-1}, \tag{1}$$

and is unique if the main diagonal of the lower triangular L is positive. The log-likelihood of a multivariate Gaussian distribution for the k-dimensional observation vector y can then be written in terms of μ and L^{-1} by

$$\ell(\mu, L^{-1}|y) = -\frac{k}{2}\log(2\pi) + \log(|L^{-1}|) - \frac{1}{2}(y-\mu)^{\mathrm{T}}(L^{-1})^{\mathrm{T}}L^{-1}(y-\mu). \tag{2}$$

We designate the nontrivial elements of $L^{-1\mathrm{T}}$ by λ_{ij}, with $i \leq j$, and link all distributional parameters to additive models:

$$\mu_i = \eta_{\mu,i}, \quad \log(\lambda_{ii}) = \eta_{\lambda,ii}, \quad \text{and} \quad \lambda_{ij} = \eta_{\lambda,ij} \quad \text{for} \quad i < j. \tag{3}$$

The reparameterization is available as a family for the R package **bamlss** (Umlauf et. al 2018) that implements optimizers for regularized estimation.

2 Simulation study

We test the proposed regression method with data simulated from a known multivariate Gaussian distribution of dimension 10. The distribution has zero mean, heteroscedastic marginal variances $\Sigma_{ii} = i$ and a first order autoregressive correlation matrix with $\rho = 0.5$.

Two different model setups are used to estimate the true distributional parameters from 50 simulated y and the process is repeated 1000 times. In Model 1, all η_i (see Eq. 3) are modeled as intercepts only. Model 2 is the same as Model 1 except that off-diagonal entries of L^{-1} (i.e. λ_{ij}, $i \neq j$) are regularized with a ridge penalty.

The estimates' representations of the true covariance and precision is evaluated using the spectral norm of the corresponding matrix differences, and compared to three reference methods for covariance estimation: (i) the sample covariance, (ii) a shrinkage covariance estimate and (iii) the graphical lasso (glasso).

The unregularized Model 1 has similar performance to the sample covariance; the regularized Model 2 performs better than both the shrinkage estimate and glasso (Fig. 1). For estimating a stationary covariance structure, the proposed multivariate distributional regression approach performs well despite the number of distributional parameters (65) exceeding the number of simulated vectors used for estimation (50). The true strength of the method, though, lies in the flexible manner in which distributional parameters can be modeled on covariates.

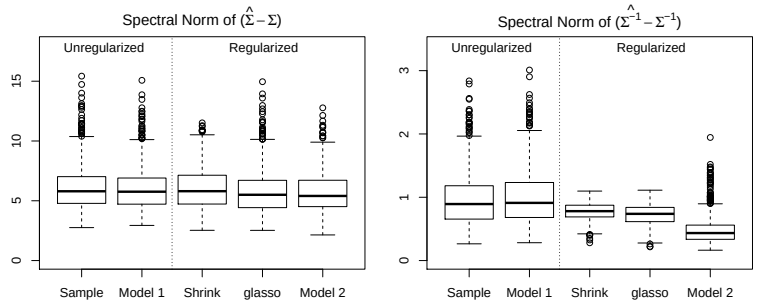


FIGURE 1. Spectral norm of differences between the estimated and true covariance (left) and precision matrices (right). Boxplots represent 1000 simulations. Smaller values indicate that the estimated covariance (precision) matrices are closer to the truth.

3 Multivariate forecasting of surface temperatures

The goal of numerical weather prediction is forecasting future atmospheric states from current observations using governing physical equations. The resulting predictions are postprocessed by statistical methods to improve their skill. For forecasting the temporal evolution of surface temperature over several future (lead) times, the error correlation between lead times must be considered. Our proposed method accomplishes this task with a multivariate approach by postprocessing the predictions (GEFS reforecasts, Hamill 2013) for several lead times simultaneously.

To illustrate, we model 00 UTC surface temperature at Innsbruck, Austria, for 8 lead times ($+8\,\mathrm{d}$, $+9\,\mathrm{d}$, …, $+15\,\mathrm{d}$) with an 8-dimensional Gaussian distribution. Seasonal variations in both predictive skill and error correlations are permitted by letting Cholesky factor entries depend on the day of the year (`yday`) and mean parameters have a linear dependency on the corresponding forecasts $\texttt{ens}_i$, but with seasonally varying coefficients:

$$\begin{aligned} \mu_i &= (\beta_{0,i} + f_{0,i}(\texttt{yday})) + (\beta_{1,i} + f_{1,i}(\texttt{yday})) \cdot \texttt{ens}_i \\ \log(\lambda_{ii}) &= \beta_{0,ii} + f_{ii}(\texttt{yday}) \\ \lambda_{ij} &= \beta_{0,ij} + f_{ij}(\texttt{yday}), \end{aligned} \tag{4}$$

where f are nonlinear cyclical functions of `yday`.

Five years of data were used to estimate the model parameters and reveal pronounced seasonal cycles in the effects of the μ models (Fig. 2). Each of the modeled λ_{ij} are also allowed to have such seasonal dependencies, which are significant for $i = j$ and also for several λ_{ij} with lag 1 (i.e. $j = i + 1$). At higher lags, the seasonal effects become insignificant.

Seasonally varying Cholesky factor estimates ($\widehat{L^{-1}}$) result in distinct $\widehat{\Sigma}$ for every `yday`. Taking $\widehat{\Sigma}$ for January 1 and July 1, we see that not only are variances in winter nearly twice as large as in summer, the errors are also more

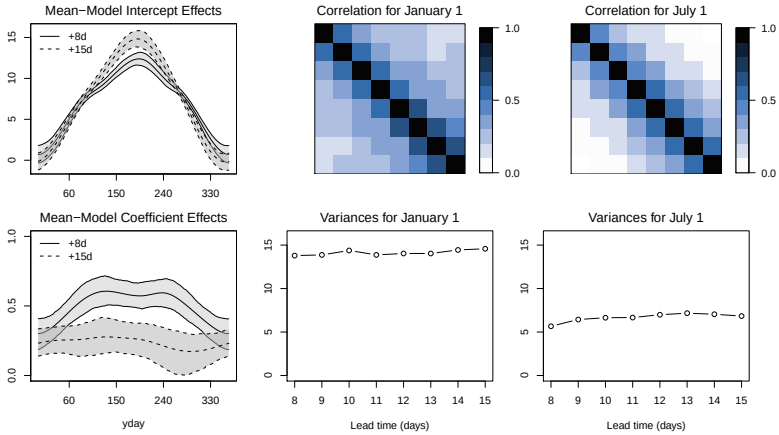


FIGURE 2. **Left column:** Estimated mean-model effects for $\beta_{0,i} + f_{0,i}(\texttt{yday})$ in Eq. 4 (top) and $\beta_{1,i} + f_{1,i}(\texttt{yday})$ (bottom). **Center column:** Correlation matrix (top) and marginal variances (bottom) calculated from the Cholesky factor estimated for January 1. **Right column:** Correlation and variances for July 1.

strongly correlated (Fig. 2). This is the benefit of the proposed multivariate Gaussian regression method: flexible mean and covariance estimation, while ensuring positive-definiteness and enabling data-driven regularization.

Acknowledgments: This project was funded by the Austrian Science Fund (FWF, grant no. P 31836). We thank the Zentralanstalt für Meteorologie und Geodynamik (ZAMG) for providing the observational data.

References

Hamill, Bates, Whitaker et al. (2013). *NOAA's second-generation global medium-range ensemble reforecast dataset.* B. Am. Meteorol. Soc., **94(10)**, 1553–1565. doi: 10.1175/BAMS-D-12-00014.1.

Klein, Kneib, Klasen and Lang (2015). *Bayesian structured additive distributional regression for multivariate responses.* J. Roy. Stat. Soc. C, **64(4)**, 569–591. doi: 10.1111/rssc.12090.

Pourahmadi (2011). *Covariance estimation: The GLM and regularization perspectives.* Statistical Science, **26(3)**, 369–387. doi: 10.1214/11-STS358

Umlauf, Klein and Zeileis (2018). *BAMLSS: Bayesian additive models for location, scale, and shape (and beyond).* J. Comput. Graph. Stat, **3**, 612–627. doi: 10.1080/10618600.2017.1407325.

Bayesian hierarchical modelling of stellar clusters

Javier Olivares[1], Hervé Bouy[1], Luis Manuel Sarro[2], Estelle Moraux[3], Ángel Berihuete[4]

[1] Laboratoire d'Astrophysique de Bordeaux, CNRS, Pessac, France.
[2] Depto. de Inteligencia Artificial , UNED, Madrid, Spain.
[3] Institut de Planétologie et d'Astrophysique de Grenoble, CNRS, Grenoble, France.
[4] Dept. Statistics and Operations Research, University of Cádiz, Cádiz, Spain.

E-mail for correspondence: `javier.olivares-romero@u-bordeaux.fr`

Abstract: Bayesian hierarchical models are useful tools for the statistical description of diverse types of phenomena. Here we present three applications of these types models for the study of stellar clusters distributions. In particular for the inference of their luminosity, spatial, and distance distributions.

Keywords: Statistical models; Astrophysics; Stellar clusters.

1 Introduction

Stellar clusters are natural laboratories where astrophysical theories can be tested. These stellar systems are formed after the collapse of a giant molecular cloud due to its own gravitational potential and possibly some external perturbation. The collapse stops once the dense cores ignite into stars, and the radiation pressure of their light expels the gas and dusts of their cocoons.
Due to their common origin, the stars in these stellar systems share, up to a certain extent, their chemical composition, age, velocity, and distance relative to the observer. The homogeneity of their properties together with their relatively large size of their populations, which range from hundreds up to thousands, make them easily identifiable from the rest of the heterogeneous Galactic population.
The Bayesian hierarchical formalism (see Gelman and Hill 2006, and references there in) allows the simultaneous inference of individual and popula-

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

tion parameters by mimicking the hierarchy present in the data. In addition, they can minimize the impact of the prior when the parameters of the later are incorporated into the model and inferred from the data as well. Here, we present three examples of Bayesian hierarchical models (hereafter BHM) designed to infer the luminosity (i.e. a proxy for the mass), spatial, and distance distributions of stellar clusters in the solar neighbourhood.

2 Luminosity distribution

In Olivares et al. (2018b), we created a BHM to simultaneously disentangle the cluster population form that of the Galactic field, and to derive a parametric representation of the distribution of stellar luminosities. The prior distribution is established based on previously known cluster members. The likelihood takes into account the heteroscedastic uncertainties and missing value sources of data sets with hundreds of thousands of stars. Due to its high computational cost, its implementation uses graphical processor units. The inference process is carried out in two steps. First, a swarm-intelligence approach (Particle Swarm Optimizer, Blackwell & Bentley 2002) is used to obtain a maximum-a-posteriori solution. Then, an affine invariant Markov Chain Monte Carlo method (Foreman-Mackey et al. 2013) is used to sample the posterior distribution of the model parameters.

3 Spatial distribution

In Olivares et al. (2018a), we constructed a BHM to infer the spatial distribution of stellar clusters (its projection in the plane of the sky and perpendicular to the line of sight). A set of common statistical and astrophysical distributions are used to describe the sky position of the stars. The parameters of these distributions are inferred using a Nested Sampling approach (Skilling 2006). The advantage of the latter is that it delivers, in addition to samples from the posterior distribution of the parameters, the Bayesian evidence of the model. This evidence provides a solid foundation to select between competing models (Trotta et al. 2006).

4 Distance distribution

In Olivares et al. (2020), we constructed a BHM to simultaneously infer individual distances to stars, and the population parameters of the cluster, like its location, scale size, and some extra parameters. The probabilistic graphical model associated with this BHM is shown in Fig. 1. In this model, a prior distribution for the distance is proposed from a set of common statistical and astrophysical families. The measurements of the stars are Gaussian distributed but not independent; they are spatially correlated.

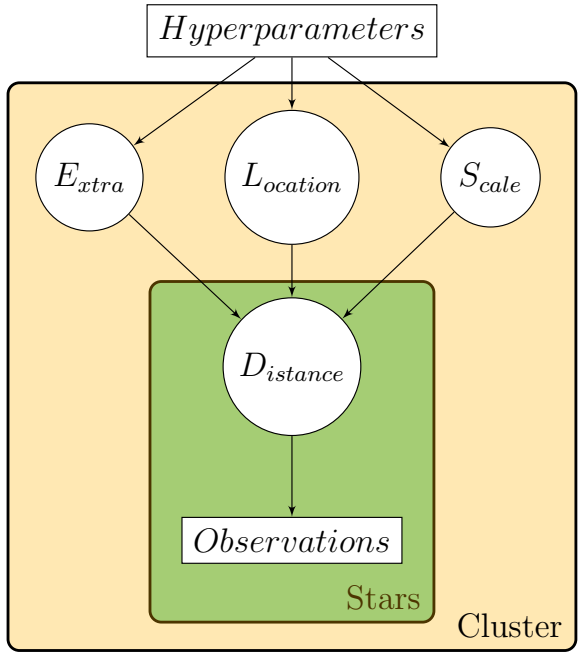


FIGURE 1. Probabilistic graphical model used to infer the distance distribution in stellar clusters. The model shows parameters at the two levels of the hierarchy: stars and cluster. Circular nodes represent inferred values while rectangular ones represent given values.

Thus the likelihood is a multivariate Gaussian distribution. The model is implemented in a probabilistic programming language (PyMC3, Salvatier et al. 2016) that performs automatic differentiation thus enabling the use of the Hamiltonian Monte Carlo sampler (Duane et al. 1987). The advantage of the latter is that the thousands of model parameters can be inferred in a few minutes using a personal computer.

5 Conclusions

The Bayesian hierarchical formalism combined with comprehensive data modelling and computationally intensive approaches, like the use graphical processors and automatic-differentiation algorithms, enables researchers to: i) compare and reject competing hypothesis, and ii) elicit the information hidden in the overwhelming flood of data of today surveys. The work shown here is an example of the success that can be expected from the exchange of methods and techniques provided by interdisciplinary collaborations, in this particular case between astronomers and mathematicians.

Acknowledgments: This research has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (grant agreement No 682903, P.I. H. Bouy), and from the French State in the framework of the Investments for the future Program, IdEx Bordeaux, reference ANR-10-IDEX-03-02

References

Blackwell, T.M., and Bentley, P.J. (2002). Dynamic search with charged swarms. In *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation*, **9**, 19 – 26.

Duane, S., et al (1987). Hybrid Monte Carlo. *Physics Letter B*, **195**, 2, 216 – 222.

Foreman-Mackey, D., et al. (2013). *emcee*: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, **125**, 925.

Gelman, A. and Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.

Olivares, J., et al. (2018a). The seven sisters DANCe. III. Projected spatial distribution *Astronomy & Astrophysics*, **612**, A70

Olivares, J., et al. (2018b). The seven sisters DANCe. IV. Bayesian hierarchical model. *Astronomy & Astrophysics*, **617**, A15

Olivares, J., et al. (2020). *Kalkayotl*: A cluster distance inference code., Submitted to *Astronomy & Astrophysics*.

Salvatier, J., Wiecki, T.V., and Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55.

Skilling, J. (2006). Nested sampling for general Bayesian computation., *Bayesian Analysis*, **1**, 833 – 859.

Trotta, R. (2006). Cosmological Bayesian Model Selection. In: *Statistical Problems in Particle Physics, Astrophysics and Cosmology*. London: Imperial College Press.

Joint model for bivariate responses using left-truncated data in aging research

Shengning Pan[1], Ardo van den Hout[1]

[1] University College London, United Kingdom

E-mail for correspondence: shengning.pan.18@ucl.ac.uk

Abstract: In aging research, the change of cognitive function over time is of interest. We construct a bivariate shared random-effects joint model to investigate it. Generally in a cognitive test, researchers use non-negative integers to reflect the level of cognitive function. We apply a bivariate binomial distribution in the joint model to investigate two test scores at the same. Moreover, to deal with the attrition, we use the Weibull hazard model and the Gompertz hazard model. The joint models are applied to the English Longitudinal Study of Ageing (ELSA) data.

Keywords: Joint model; Bivariate binomial distribution; Cognitive function.

1 Introduction

In aging research, it is important to investigate the changes in individual cognitive function. Cognitive function is the individual's ability to process information, which mainly contains learning and problem-solving ability. Scientists provide advice on whether old people need care by analyzing the relationship between cognitive function decline and aging (Van den Hout and Muniz-Terrera, 2016). Normally, the data related to cognitive function are longitudinal data. The most common causes of dropout are dementia and death in this type of data, which can be treated as an event of interest in corresponding survival analysis. Therefore, the cognitive-related data can be analyzed using the joint model. Moreover, scientists sometimes use more than one test to investigate cognitive ability in one study. Since all of the responses are important to such research, it is necessary to construct a multivariate model.

The bivariate longitudinal model and the survival model are joint by sharing a random effect, and we assume that these two models are independent

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

given the random effect. Since the cognitive function is usually measured by non-negative integers, the bivariate extension of the binomial distribution proposed by Altham and Hankin (2012) is used for the longitudinal model. The Weibull hazard and the Gompertz hazard are used for the survival models. We apply the models to analyze the English Longitudinal Study of Ageing (ELSA) data. The data are related to verbal learning and recall. Individuals are required to learn ten words and recall these words at two different time points (immediate and later), but within the same interview; see Van den Hout and Muniz-Terrera (2018).

2 Models

This section introduces the bivariate binomial distribution used for the longitudinal model and the Weibull, Gompertz hazard models for the survival part. After that, the marginal likelihood function is presented. For what follows, assume that the random effect is b.

2.1 Longitudinal model

We assume that the longitudinal responses for two scores are $y^{(1)}$ and $y^{(2)}$ at any time (or age) t. The bivariate binomial distribution is used to analyze the responses (Altham and Hankin, 2012):

$$p(Y_1 = y^{(1)}, Y_2 = y^{(2)}) = \frac{g(Y_1 = y^{(1)})g(Y_2 = y^{(2)})\phi^{y^{(1)}y^{(2)}}}{C}$$

$$g\left(Y_j = y^{(j)}\right) = \binom{m}{y^{(j)}} p_{Y_j}^{y^{(j)}} (1 - p_{Y_j})^{(m-y^{(j)})} \theta_{Y_j}^{y^{(j)}(m-y^{(j)})},$$

where $0 < p_{Y_1}, p_{Y_2} < 1$, $\theta_{Y_j}, \phi > 0$, and

$$C = \sum_{y^{(1)}=0}^{m} \sum_{y^{(2)}=0}^{m} g(Y_1 = y^{(1)})g(Y_2 = y^{(2)})\phi^{y^{(1)}y^{(2)}}.$$

The probabilities p_{Y_j} are linked to time using a logistic regression model:

$$p_{Y_j} = \frac{\exp(\eta_0^{(j)} + \eta_1^{(j)} t + b^{(j)})}{1 + \exp(\eta_0^{(j)} + \eta_1^{(j)} t + b^{(j)})}.$$

2.2 Survival model

The event of interest is death. The Weibull hazard model and the Gompertz model are used to construct the survival part of the joint model:

$$\text{Weibull} : h(t_{last}) = \exp\left(\beta + \alpha(\eta_0^{(j)} + b^{(j)})\right) \tau t_{last}^{(\tau-1)} \tag{1}$$

$$\text{Gompertz} : h(t_{last}) = \exp\left(\beta + \alpha(\eta_0^{(j)} + b^{(j)}) + \gamma t_{last}\right), \tag{2}$$

where $\tau, \gamma > 0$, t_{last} is the last recorded age for the corresponding individual. Equations (1) and (2) assume that the random intercept in the longitudinal model would impact the risk of death via $\alpha_j(\eta_0^{(j)} + b^{(j)})$.

2.3 Marginal likelihood function given left truncation

Left truncation, also called *delayed entry*, occurs when individuals have been at risk before entering the study (Wienke, 2010). In aging research, since the event of interest is death, individuals can be included in the data only if they have not experienced the event before they enter the study. If we do not deal with the left truncation, the estimation is based on the assumption that individuals were not at risk of dying before the start of the study. Therefore, the left truncation needs to be taken into account in the model estimation.

For individual i, $i = 1, \ldots, N$, the corresponding longitudinal responses are $\boldsymbol{y}_i = (\boldsymbol{y}_i^{(1)}, \boldsymbol{y}_i^{(2)})$, $\left(\boldsymbol{y}_i^{(j)} = (y_{i1}^{(j)}, \ldots, y_{in_i}^{(j)})\right)$ at age $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{in_i})$, where $j = 1, 2$ is the jth method of measuring cognitive function, n_i is the number of observations for each individual. Let $\boldsymbol{\omega}$ represent all the parameters in the joint model except the random effects, t_{i1} for baseline age. The likelihood contribution of individual i conditionally on truncation time t_{i1} is:

$$L_i(\boldsymbol{\omega}|\boldsymbol{y}_i, t_i, T \geq t_{i1}) = p(\boldsymbol{y}_i, t_i|T \geq t_{i1}, \boldsymbol{\omega}) = \frac{p(\boldsymbol{y}_i, t_i|\boldsymbol{\omega})}{p(T \geq t_{i1}|\boldsymbol{\omega})}, \tag{3}$$

where $p(T \geq t_{i1}|\boldsymbol{\omega})$ is the survivor function evaluated at t_{i1}. For the shared random-effects model, the denominator in (3) can be written as:

$$p(T \geq t_{i1}|\boldsymbol{\omega}) = \int p(T \geq t_{i1}|\boldsymbol{b}_i, \boldsymbol{\omega}) p(\boldsymbol{b}_i|\boldsymbol{\omega}) d\boldsymbol{b}_i.$$

Assuming the independence between responses given the random effect, so that the marginal likelihood function is:

$$\begin{aligned} p(\boldsymbol{y}_i, t_i|\boldsymbol{\omega}) &= \int p\left(\boldsymbol{y}_i|\boldsymbol{t}_i, \boldsymbol{b}_i\right) p\left(t_{last\ i}|\boldsymbol{b}_i, \delta_i\right) p\left(\boldsymbol{b}_i\right) d\boldsymbol{b}_i, \\ &= \log \sum_{i=1}^{N} \int \left[\prod_{k=1}^{n_i} f(y_{ik}^{(1)}, y_{ik}^{(2)}|t_{ik}, b_{ji})\right] f\left(t_{last\ i}|\boldsymbol{b}_i, \delta_i\right) f\left(\boldsymbol{b}_i\right) d\boldsymbol{b}_i, \end{aligned}$$

where $t_{last\ i} = t_{in_i}$ is the last recorded age. Parameter $\delta_i = 0$ means alive at the last observation and $\delta_i = 1$ means death. Distribution $p\left(\boldsymbol{y}_i|\boldsymbol{t}_i, \boldsymbol{b}_i\right)$ represents the longitudinal model and $p\left(t_{last\ i}|\boldsymbol{b}_i, \delta_i\right)$ is the survival model:

$$p\left(t_{last\ i}|\boldsymbol{b}_i, \delta_i\right) = h\left(t_{last\ i}|\boldsymbol{b}_i\right)^{\delta_i} P(T \geq t_{last\ i}).$$

We define the random effect $\boldsymbol{b}_i \in \mathbb{R}^p$ by $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \sum)$, where $\sum$ is a $p \times p$ covariance matrix.

3 Application

We apply the model to analyze the ELSA data. The bivariate responses are the number of immediately recalled words '$y^{(1)}$' and the number of later recalled words '$y^{(2)}$'. We code the marginal log-likelihood function in *R*. The corresponding parameters are estimated using the *optim* function with the *Nelder-Mead* algorithm.

Hazard distribution	$-2LL$	AIC	α
Weibull: $j = 1$	29286.58	29312.58	-1.216
Gompertz: $j = 1$	28972.61	**28998.61**	-1.608
Weibull: $j = 2$	29281.87	29307.87	-1.822
Gompertz: $j = 1$	29010.75	29036.75	-1.013

TABLE 1. AIC and estimated αs for shared random-effects models, where j refers to immediately recall and later recall

The AIC values for joint models with Gompertz hazard model are smaller than joint models with Weibull hazard model, when they share the same random intercept $\eta_0^{(j)} + b^{(j)}$. The estimation of αs follows our expectation: the risk of death will be relatively low if individuals have a good cognitive function at the baseline age.

References

Altham, P. M., & Hankin, R. K. (2012). Hierarchical generalized linear models. *Journal of Statistical Software*, **46**(12), 1 – 23.

Van den Hout, A., & Muniz – Terrera, G. (2016). Joint models for discrete longitudinal outcomes in aging research. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **65**(1), 167 – 186.

Van den Hout, A., & Muniz – Terrera, G. (2018). Hidden three-state survival model for bivariate longitudinal count data. *Lifetime Data Analysis*, 1 – 17.

Wienke, A. (2010). *Frailty Models in Survival Analysis*, Chapman and Hall/CRC.

Spatio-temporal and hierarchical modelling of high-throughput phenotypic data

Diana M. Pérez [1], María Xosé Rodríguez-Álvarez[1,2], Martin P. Boer [3], Emilie J. Millet [3], Fred A. van Eeuwijk [3]

[1] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[3] Biometris, Wageningen University & Research, Wageningen, The Netherlands

E-mail for correspondence: dperez@bcamath.org

Abstract: We present a full spatio-temporal and hierarchical data modelling approach for the analysis of high-throughput phenotypic data. We use the recently proposed SpATS approach as the base model, and extend it to the spatio-temporal case, also considering a three-level hierarchical data model (plants nested in genotypes, nested in populations). We illustrate our approach using data from a high-throughput phenotypic platform.

Keywords: Agricultural experiments; high-throughput phenotyping platforms; P-splines; spatio-temporal model; three-level hierarchical model.

1 Introduction

Plant breeding programmes aim to improve food production and to enhance nutrition through genetic improvement. Recent technological developments have improved data acquisition through high-throughput phenotypic platforms. With these platforms, researchers have now access to large and detailed datasets on multiple traits (phenotypes) for many plants and genotypes, long time-series of repeated measurements, under different environmental and management conditions, to cite a few.
In this work, we aim to model the longitudinal evolution of the genetic effect on a given phenotype, while correcting for the environmental effects. In particular, we generalise the two-stage modelling strategy presented in Pérez et al. (2019) to a full and one-stage spatio-temporal approach. We use the spatial SpATS model (Rodríguez-Álvarez et al., 2018) as the base model

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

and extend it to the spatio-temporal case, considering a three-level hierarchical data structure (plants nested in genotypes, nested in populations). We use (tensor-products of) cubic B-splines and second order penalties (P-splines, Eilers and Marx, 1996), and their representation as mixed-effect models. Our approach tackles a great computational challenge due to the large amount of observations and the large number of parameters to be estimated. To speed up computation, we take the advantage of the array structure of the data through Generalised Linear Array Models (GLAM, Currie et al. 2006). Also, the computational time is further improved by exploiting the sparse structure of the matrices involved in the model.

2 Spatio-temporal Data Modelling Approach

Let $y_i(t)$ denote the observed phenotype of interest of the i-th plant at time t, which is modelled as follows,

$$y_i(t) = \underbrace{f_{p(i)}(t) + f_{g(i)}(t) + f_i(t)}_{\text{3-level longitudinal effects}} + \underbrace{f_{r(i)}(t) + f_{c(i)}(t) + f(r(i), c(i), t)}_{\text{Spatio-temporal effects}} + \varepsilon_i(t),$$

where $f_{p(i)}(\cdot)$ models the evolution over time of the p-th population, $f_{g(i)}(\cdot)$ and $f_i(\cdot)$ are random processes associated with genotype g and plant i, respectively; $f_{r(i)}(\cdot)$ and $f_{c(i)}(\cdot)$ are random processes associated with row r and column c, respectively, and $f(\cdot, \cdot, \cdot)$ is a spatio-temporal three-dimensional surface defined over the row and column positions (r and c), and time t. Finally, $\varepsilon_i(\cdot)$ is a white noise measurement error with variance σ^2. Each univariate function ($f_{p(i)}(t)$, $f_{g(i)}(t)$, $f_i(t)$, $f_{r(i)}(t)$, and $f_{c(i)}(t)$) is modelled using cubic B-spline basis functions, and $f(r(i), c(i), t)$ using the tensor product of marginal cubic B-spline bases. Smoothness is achieved by imposing a second-order difference penalty on the regression coefficients. We use the connection between P-splines and mixed models through the parametrisation proposed by Lee and Durban (2011). Here, the smooth functions are sums of linear and non-linear components, and smoothing parameters are "replaced" by variance components. For $f_{g(i)}(\cdot)$, $f_i(\cdot)$, $f_{r(i)}(\cdot)$ and $f_{c(i)}(\cdot)$ we penalise (assume random) the linear component (intercept + slope). Note that it implies that we have one variance component per population, and three variance components for genotypes, plants, rows and columns (associated, respectively, with the intercept, the slope and non-linear effect). Finally, the smoothness of the spatio-temporal surface is controlled by three variance components (for row, column and time).

3 Application to the PhenoArch Platform

The data analysed here corresponds to an experiment conducted in the PhenoArch platform (Cabrera-Bosquet et al., 2016). The data set consists

of 35 leaf area measurements on 1680 plants of 180 genotypes from 4 populations of maize ($1680 \times 35 = 58800$ observations, including missing data). For the results shown here, B-spline bases of dimension 11 were used to model $f_{p(i)}(t)$, $f_{g(i)}(t)$, $f_i(t)$, $f_{r(i)}(t)$, and $f_{c(i)}(t)$, and of dimension 8 for each marginal of $f(r(i), c(i), t)$. This configuration yielded a total of 21877 regression coefficients (both fixed and random), and 20 variance components. Model estimation took approximately 70 minutes. Computations were performed in (64-bit) `R` 3.6.3, and a 2.40GHz $\times$ 4 Intel® Core™ i7 processor computer with 15.6GB of RAM and Ubuntu 16.04 LTS OS.
Figure 1 shows the estimated genotypic deviations (the main level of decision for these experiments) for five genotypes per population selected for illustration. We compare the results with (1) the estimated genotypic deviations obtained with the two-stage approach by Pérez et al. (2019), and (2) the genotypic BLUPs obtained from the SpATS analysis of each measurement time separately. In order to characterise the genotypes, the first and second-order derivatives of the (estimated) plant trajectories were obtained and some features extracted (as in Hurtado et al., 2012). Figure 2 shows the maximum growth rate and acceleration rate of the leaf area for all genotypes in one population.

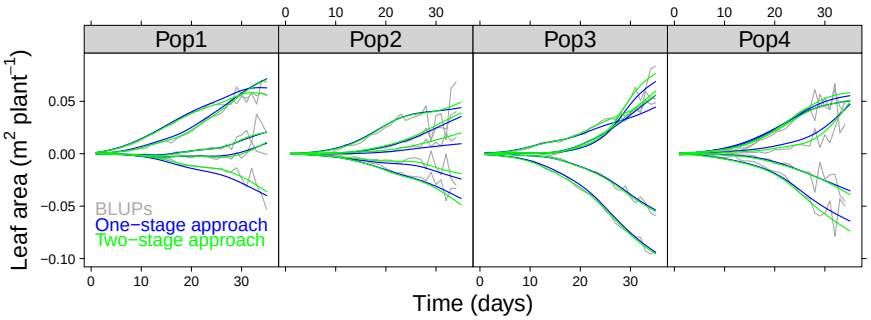


FIGURE 1. **Estimated genotypic deviations** $\hat{f}_{g(i)}(t)$ for five genotypes per population (as illustration). The blue lines correspond to the results using proposal presented here, the green lines to the results using to the two-stage approach, and the gray lines are the genotypic BLUPs obtained from SpATS.

Acknowledgments: This research was supported by the Basque Government through the BERC 2018-2021 program, by the Spanish Ministry of Science, Innovation and Universities through BCAM Severo Ochoa accreditation SEV-2017-0718, through project MTM2017-82379-R funded by (AEI/FEDER, UE) and through EU project H2020 731013 (EPPN2020). We thank Llorenç Cabrera-Bosquet and François Tardieu for sharing with us the PhenoArch data.

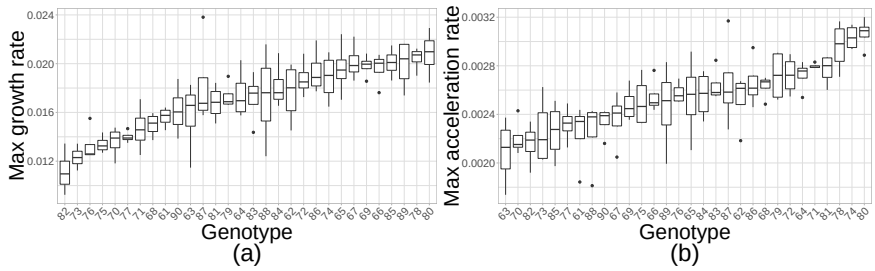


FIGURE 2. **Genotypic characterization** from the first and second derivatives curves of the plant trajectories for the genotypes in one population (as illustration). Figure (a): box-plot of the maximum growth rates of the leaf area, extracted from the first-order derivative curves, for the plants of each genotype. Figure (b): box-plot of the maximum acceleration rates of the leaf area, extracted from the second-order derivative curves, for the plants of each genotype. Genotypes are ordered according the median value for the respective feature.

References

Cabrera-Bosquet, L., Fournier, C., Brichet, N. , Welckerand, C., Suard, B., and Tardieu, F. (2016). High-throughput estimation of incident light, light interception and radiation-use efficiency of thousands of plants in a phenotyping platform. *New Phytologist*, **212**, 269 – 281.

Currie, I.D., Durban, M., and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society B*, **68(2)**, 259-280.

Eilers, P. H. C., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89-102.

Lee D-J. and Durban M (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, **11**, 49-69.

Hurtado, P.X., Schnabel, S.K., Zaban, A., Vetelinen, M. , Virtanen, E., Eilers, P. H., van Eeuwijk, F. A., Visser, R. G. and Maliepaard, C. (2012). Dynamics of senescence-related QTLs in potato. *Euphytica*, **183(3)**, 289-302.

Pérez, D.M., Rodríguez-Álvarez, M.X., Boer, M.P., Millet, E.J. and van Eeuwijk F.A. (2019). A two-stage approach for the spatio-temporal analysis of high-throughput phenotypic data. *Proceedings of the 34th International Workshop on Statistical Modelling. Volume II.* Guimarães, Portugal.

Rodríguez-Álvarez, M.X., Boer, M.P., van Eeuwijk, F.A., and Eilers, P.H.C. (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spatial Statistics*, **23**, 52 – 71.

Design of truncated repetitive sampling plan for Poisson count data using expected sampling risks

Carlos J. Pérez-González[1], Arturo J. Fernández,[1]

[1] Universidad de La Laguna, San Cristóbal de La Laguna, Spain

E-mail for correspondence: cpgonzal@ull.es

Abstract: Single and repetitive sampling plans represent conventional methods used in inspection the quality of lots or batches of products. Truncated repetitive inspection presented in this paper allows the practitioners to significantly reduce the required sampling effort from the lot. In this scheme, the lots can be reinspected, at most, a prefixed number of times when their acceptance or rejection cannot be concluded from the original inspection. We develop the design of truncated repetitive sampling plans based on defect count data and using expected sampling risks. The Poisson distribution is assumed for the number of defects found in the sample and a gamma prior model on the unknown defect rate is considered. The optimal truncated repetitive sampling plans are obtained by solving several nonlinear programming problems. The results show that optimal truncated plans are better than the conventional single and repetitive schemes in reducing the average sample number of the inspection.

Keywords: Quality control; Expected producer and consumer risks; Poisson distribution; Gamma prior distribution

1 Introduction

This work presents the design of a truncated repetitive sampling plan for lot acceptance when defect counts are Poisson distributed and a prior model on the defect rate is considered; see Pérez-González et al. (2020). In this scheme, that was introduced by Pérez-González et al. (2019), the lots that are not accepted or rejected can be reinspected, at most, a certain number of times that is defined as the truncation parameter of the inspection scheme. The gamma distribution is used to include previous data about the unknown defect rate into the decision process of the inspection. Truncated

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

inspection plans outperforms the single sampling plans as well as the standard repetitive plans proposed by Sherman (1965) in terms of the average sample number (ASN) and allow us to save inspection costs.

2 Truncated repetitive inspection plans

Consider that the quality variable of interest in a manufacturing process of a particular product is the number of defects C of an item. Assuming that C follows a Poisson distribution with unknown defect rate per unit $\lambda > 0$, the producer considers that lots are satisfactory if $\lambda \leq \lambda_0$, where λ_0 is the acceptable defect rate, whereas the inadmisible lots for the consumer correspond to $\lambda \geq \lambda_1$, where λ_1 ($> \lambda_0$) is the rejectable defect rate.
Given the integer decision limits r and s, where $0 \leq r \leq s$, the truncated repetitive sampling plan (n, r, s, t) may be described as follows

- *Step 1.* Initialize the number of the inspection sampling stage $k = 1$.
- *Step 2.* Select the kth random sample of n independent units from the lot or batch and compute the total number $D_k = \sum_{j=1}^{n} C_{jk}$, where C_{jk} is the number of nonconformities of the jth item in this kth sample, with $j = 1, .., n$.
- *Step 3.* If $k < t$, then the lot is accepted when $D_k \leq r$ and rejected when $D_k > s$. Otherwise, a decision cannot be made and Step 2 is repeated with $k = k + 1$ if $k < t$. If $k = t$, then the lot is accepted when $D_t \leq r$ and rejected, otherwise.

The operating characteristic (OC) function for a truncated repetitive plan (n, r, s, t) is denoted by $A_t(\lambda) \equiv A_t(\lambda; n, r, s)$ for any defect rate $\lambda > 0$ as the probability of lot acceptance. Likewise, the ASN function is defined as the expected number of sample units that are inspected per batch until reaching the decision of lot acceptance or rejection. This function for the truncated repetitive plan (n, r, s, t) will be denoted by $ASN_t(\lambda) \equiv ASN_t(\lambda; n, r, s)$ for $\lambda > 0$.
We also assume a gamma prior distribution of λ with parameters a (shape) and b (scale). Then, the prior density function (pdf) is denoted as $h(\lambda) \equiv h(\cdot; a, b)$, for $\lambda > 0$, whereas the prior cumulative distribution function (cdf) is $H(\lambda) \equiv H(\cdot; a, b)$.

2.1 Expected sampling risks

When the prior information about the incoming defect rate is assumed, the use of expected risks can be considered in designing acceptance sampling plans. Given a truncated repetitive plan (n, r, s, t), the expected producer's

risk (EPR) is defined as the probability of rejecting a satisfactory lot and is given by

$$\begin{aligned} EPR(\lambda_0; n, r, s, t) &= E_h[1 - A_t(\lambda; n, r, s) \mid \lambda \leq \lambda_0] \\ &= \int_0^{\lambda_0} \{1 - A_t(\lambda; n, r, s)\} h(\lambda) d\lambda / H(\lambda_0), \end{aligned}$$

whereas the expected consumer's risk (ECR) is the probability of accepting a lot that is unsatisfactory and can be expressed as

$$\begin{aligned} ECR(\lambda_1; n, r, s, t) &= E_h[A_t(\lambda; n, r, s) \mid \lambda \geq \lambda_1] \\ &= \int_{\lambda_1}^{\infty} A_t(\lambda; n, r, s) h(\lambda) d\lambda / \{1 - H(\lambda_1)\}. \end{aligned}$$

3 Optimal designs of truncated repetitive plans

Sampling inspection plans must protect to manufacturers against rejecting good lots as well as to customers against accepting bad lots. Therefore, the maximum expected producer's and consumer's risks at the acceptable and rejectable defect rates, λ_0 and λ_1, need to be specified. The best t-truncated repetitive sampling plan (n, r, s, t) can be determined by solving the constrained optimization problem

$$\begin{array}{ll} \text{Minimize} & E[ASN_t(\lambda; n, r, s)] \\ \text{Subject to} & EPR(\lambda_0; n, r, s, t) \leq \alpha_0, \\ & ECR(\lambda_1; n, r, s, t) \leq \alpha_1, \\ & n, t \in \mathbb{N}, r, s \in \mathbb{N}_0, \\ & 0 \leq r \leq s, \end{array} \tag{1}$$

where $EASN$ denotes the sampling inspection effort of the truncated repetitive plan that can be defined as

$$\begin{aligned} EASN_t \equiv EASN_t(n, r, s) &= E[ASN_t(\lambda; n, r, s)] \\ &= \textstyle\int_0^{\infty} ASN_t(\lambda; n, r, s) h(\lambda) d\lambda. \end{aligned} \tag{2}$$

Table 1 shows, for selected values of λ_0, λ_1 and t, the best t-truncated repetitive sampling plans with $EASN$ given by $N_t = E[ASN_t(\lambda; n_t, r_t, s_t)]$ when the prior mean is given by $\mu_\lambda = \lambda_0, \lambda_1$ and the variance is $\sigma_\lambda^2 = (\lambda_1 - \lambda_0)^2$. The single and standard repetitive plans are also computed and presented in the table for comparison. We observe that the $EASN$ reduces when t increases although we can appreciate that there are truncated plans for $t \leq 6$ with lower $EASN$ than the standard repetitive plans.

4 Conclusion

According to previous results, the expected sampling risks of the truncated plans provide the practitioners more precise estimates of the current producer's and consumer's risks. Likewise, the required number of lot reinspections are quite small and the proposed plans can be more appropriate for

TABLE 1. Single ($t = 1$), t-truncated and standard ($t = \infty$) repetitive plans with minimum $EASN$ when $\alpha_0 = 0.05$, $\alpha_1 = 0.10$ and λ follows a gamma distribution with mean $\mu_\lambda = \lambda_0, \lambda_1$ and $\sigma^2_\lambda = (\lambda_1 - \lambda_0)^2$.

			$\mu_\lambda = \lambda_0$				$\mu_\lambda = \lambda_1$			
λ_0	λ_1	t	n_t	r_t	s_t	$EASN_t$	n_t	r_t	s_t	$EASN_t$
0.2	0.3	1	55	13	13	55.00	70	18	18	70.00
		3	32	6	9	41.34	42	9	12	54.11
		6	22	3	7	39.37	26	4	9	54.95
		∞	29	5	8	40.25	33	6	10	53.56
0.4	0.5	1	103	46	46	103.00	117	54	54	117.00
		3	58	23	29	77.96	65	27	33	88.88
		6	42	15	22	72.92	46	17	25	88.80
		∞	40	14	21	72.91	44	16	24	89.71
0.6	0.7	1	154	100	100	154.00	164	109	109	164.00
		3	82	49	58	112.77	84	51	62	129.13
		6	65	37	47	109.11	70	41	52	126.75
		∞	41	20	33	132.55	41	20	34	161.39

testing expensive and high quality products, whereas single and repetitive schemes always increase the economical and time costs.

Acknowledgments: This work has been partially supported by the Spanish Ministerio de Ciencia e Innovación (MICINN) under the grant PID2019-110442GB-I00.

References

Pérez-González, C.J., Fernández, A.J., Kohansal, A., and Asgharzadeh, A. (2019). Optimal truncated repetitive lot inspection with defect rates. *Applied Mathematical Modelling*, **75**, 223 – 235.

Pérez-González, C.J., Fernández, A.J., and Kohansal, A. (2020). Efficient truncated repetitive lot inspection using Poisson defect counts and prior information. *European Journal of Operations Research*, in press.

Sherman, R. E. (1965). Design and evaluation of a repetitive group sampling plan. *Technometrics*, **7**, 11 – 21.

Hidden Markov Models Incorporating Covariates for Daily Rainfall Time Series

Nadarajah Ramesh[1], Gayatri Rode[1]

[1] University of Greenwich, London, UK

E-mail for correspondence: n.i.ramesh@gre.ac.uk

Abstract: Hidden Markov models provide a rich class of stochastic models that are very useful in hydrological studies. This paper describes a class of hidden Markov models that incorporate covariates in their state distributions to model daily rainfall time series. Greater emphasis is placed on finding a model that can reproduce the second-order properties of the observed rainfall sequences. We present the construction of the likelihood function incorporating time-dependent atmospheric covariates in rainfall distributions. The performance of the model is assessed using daily rainfall data from Leicester, East Midlands region of England.

Keywords: Hidden Markov models; Rainfall; Maximum likelihood; Second-order properties.

1 Introduction

Long term precipitation data is a key input variable in hydrological studies that aim to understand environmental and ecological systems and quantify uncertainty. Stochastic models enable us to study the characteristics of the rainfall process and to generate long sequence of precipitation. Hidden Markov model (HMM) contributes to the development of a rich class of stochastic models that are useful in environmental applications. The HMMs have been used in rainfall modelling by many authors, following earlier work by Zucchini and Guttorp (1991). Hughes et al. (1999) considered a non-homogeneous HMM to model precipitation occurrences. Ramesh and Onof (2014) explored ways of introducing additional dependence in HMMs to model regional rainfall. In this paper, we describe a class of HMMs that incorporate atmospheric covariates in their state distributions to model daily rainfall time series.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

1.1 Model Formulation

Let $\{X(t)\}$ be an irreducible finite state Markov chain with states space $S = \{1, 2, \ldots, m\}$. Suppose that $y_1, y_2, \ldots, y_n$ is the observed sequence of daily rainfall volumes at a location. The distribution of rainfall on any given day is dependent on the state of the underlying Markov chain on that day. Let $f_j\,(j = 1, 2, .., m)$ be the distributions of daily rainfall $Y(t)$ corresponding to state j of the Markov chain. The HMM is characterized by its transition probability matrix $(\Lambda_{m\times m})$ and the diagonal matrix F of state distributions. In our application, we use a 3-state Markov chain $(m = 3)$ with no rainfall in state 1, moderate and heavy rain in states 2 and 3, respectively. In addition, based on empirical evidence, the state distributions f_j in the two rainy states $(j = 2, 3)$ are taken as exponential with parameter, λ_j. This is the standard HMM with exponential state distributions. In an attempt to allow the local climate variables to influence the daily rainfall, we incorporate atmospheric covariates in our model and express the exponential rate parameter λ_j, as a function of the covariates. We used three time varying covariates in this application and they are temperature (U), sea level pressure (V) and relative humidity (W). The parameters of the rainfall distribution in state j at time t is defined as

$$\lambda_{tj} = e^{\beta_{0j}+\beta_1 U_t+\beta_2 V_t+\beta_3 W_t} \tag{1}$$

where $j = 2, ..., m$ are the rainy states of the Markov chain.
To allow the model to capture the dependence relationship more strongly, we use the moving average of the covariates. Hence, the parameters of the state dependent distributions are taken to be functions of a three-day moving average of the covariates. We define a new variable $U_t^{(3)}$ at time t as the moving average of U_t over the past three days as given below

$$U_t^{(3)} = (U_{t-2} + U_{t-1} + U_t)/3. \tag{2}$$

The moving averages $V_t^{(3)}$ and $W_t^{(3)}$ are defined similarly. The state dependent distribution parameter for state j at time t is now defined as

$$\lambda_{tj} = e^{\beta_{0j}+\beta_1 U_t^{(3)}+\beta_2 V_t^{(3)}+\beta_3 W_t^{(3)}}. \tag{3}$$

Let π be the stationary distribution of the Markov chain, $\mathbf{1}$ be a unit vector of ones and Z_t be a vector containing the 3-day moving averages of U, V and W. The likelihood function of this model with covariates is given by

$$L(y_1, y_2, , y_n|\Lambda, F, z_1, z_2, .., z_n) = \pi\prod_{t=1}^{n}[\Lambda F(y_t|z_t)]\mathbf{1}'. \tag{4}$$

The state distribution matrix F in the above equation is defined as

$$F(y_t|z_t)_{m\times m} = \mathrm{Diag}(f_1(y_t|z_t), f_2(y_t|z_t), ..., f_m(y_t|z_t)) \tag{5}$$

where $f_j(y_t|z_t)$ is the state dependent density function dependent on the three-day moving average of covariates. The parameter estimates of the likelihood function (4) are obtained by employing the maximum likelihood estimation using standard routines in R studio.

1.2 Data Analysis

The proposed exponential HMMs are fitted to the winter season daily rainfall data of length 36 years from Leicester, England. A three state traditional hidden Markov model is used as the baseline model (M1), which is then compared with the models incorporating a combination of atmospheric covariates. We studied eight different models using one, two or all three covariates with an emphasis on reproducing the empirical statistics as closely as possible. Almost every covariate model outperformed the baseline model M1 and captured the second-order properties well with an odd few exceptions. We present the results of the best two models along with that of the reference model M1. Table 1 gives a summary of results for the three models. The model M2 incorporates sea level pressure in the state distributions and M3 incorporates all three covariates. The values of the negative loglikelihood, AIC and BIC for the three models are displayed. The results show that the model M3 performs better than the models M1 and M2.

TABLE 1. Summary of likelihood based results for the three models.

Model	M1	M2	M3
Covariates	None	SLP	TEM, SLP, HUM
Negative loglikelihood	6091.80	6083.755	**6068.241**
AIC	12199.7	12185.51	**12158.48**
BIC	12211.8	12199.12	**12175.11**

Figure 1 displays the results of the simulation study undertaken to assess the performance of the models. Simulation intervals (in blue), based on 100 replications, and the empirical statistics (red crosses) are displayed. The plots show that the mean and variance of the process are accurately reproduced by all three models. Models M2 and M3 have notable improvement over M1, in capturing the autocorrelation of the process. The wet spell distribution plot (top right) is seen to accurately trace the empirical distribution with a slight variation around the duration of 5 to 10 days.

2 Conclusions

The results obtained in our analysis show that the model incorporating all three covariates outperformed the other models studied in modelling daily

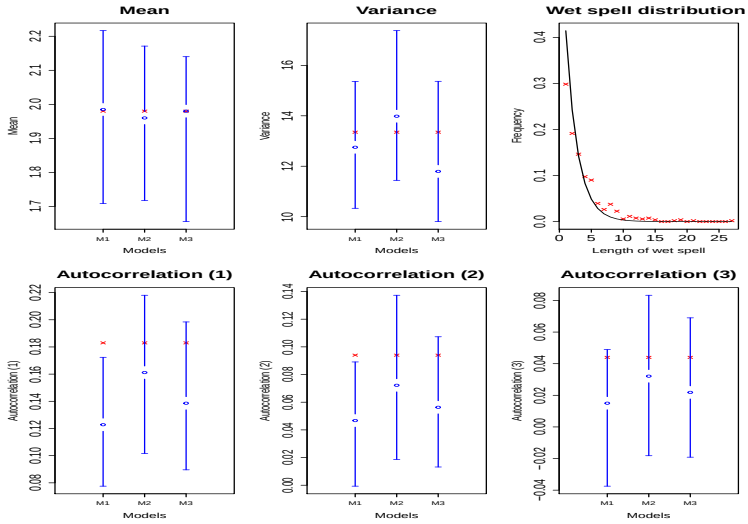


FIGURE 1. Results of simulation study. Simulated intervals are in blue and empirical statistics are in red. Top panels: mean, variance and wet spell distribution. Bottom panel: Autocorrelations at lag 1, lag 2 and lag 3.

rainfall time series. The simulation study provides further evidence for the model performance. Models incorporating atmospheric covarites capture the dependence structure present in the daily rainfall data better than the basic model. Our future work intends to explore models incorporating covariates in the Markov chain parameters of HMM.

Acknowledgments: This work was part-funded by a Vice-Chancellor's scholarship from the University of Greenwich (Ref. No: VCS-ACH-01-17).

References

Hughes, J.P., Guttorp, P. and Charles, S.P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence, *Applied Statistics*, **48**(Part 1), 15-30.

Ramesh, N.I. and Onof, C. (2014). A class of hidden Markov models for regional average rainfall. *Hydrological Sciences Journal*, Vol. **59** (9), 1704-1717.

Zucchini, W., MacDonald, I. L. and Langrock, R. (2016). *Hidden Markov models for time series:an introduction using R*. Chapman and Hall/CRC.

Spatial Clustering via the Cross Entropy Method

Nishanthi Raveendran [1], Georgy Sofronov[1], David Bulger[1]

[1] Department of Mathematics and Statistics, Macquarie University, Sydney, Australia

E-mail for correspondence: nishanthi.raveendran@students.mq.edu.au

Abstract: Spatial clustering is an important component of spatial data analysis which aims to identify the number of clusters and their boundaries. Applications include epidemiology, criminology and many others. In this study, we focus on identifying homogeneous clusters in binary data, which indicate the presence or absence of a certain plant species observed over a two-dimensional lattice. To solve this clustering problem, we propose to combine the Cross Entropy method with Voronoi tessellation to estimate the boundaries of such domains. Our results illustrate that the proposed algorithm is effective in identifying homogeneous clusters in spatial binary data.

Keywords: Spatial clustering; Binary data; Cross entropy; Voronoi tessellation.

1 Introduction

Spatial clustering is one of the main techniques for spatial data mining and analysis. Spatial clustering aims to partition spatial data into a series of meaningful subclasses, called spatial clusters, such that data points in the same cluster are near to each other, but far from those in different clusters. Currently, spatial clustering is widely applied in the field of spatial data analysis, such as spatial epidemiology, land use detection, crime hotspot analysis, population genetics, ecology and many other fields. In the last decade, many clustering algorithms have been developed, ranging from hierarchical methods such as bottom-up (or agglomerative) methods and top-down (or divisive) methods, to optimization methods such as the k-means algorithm. In this study we propose to apply multiple change-point detection methodology, commonly used to detect changes and their locations in time series data, to spatial clustering problems. We focus on binary

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

data, which are commonly involved in various areas such as economics, social sciences, image analysis and epidemiology. Also, such data frequently occur in environmental and ecological research, for instance indicating the presence of an invasive plant species, or when the data happen to fall into one of two categories, say, two types of soil. To solve this clustering problem, we present an effective algorithm based on the Cross Entropy method, an evolutionary stochastic optimization technique, and Voronoi tesselation, to identify homogeneous clusters and their boundaries.

2 Model

There are several methods for modeling spatially correlated presence or absence data. Among them, the autologistic model [Besag, 1972] is a popular tool. Letting Z be the random field of interest, where $Z_i \in \{0, 1\}$ represents the observation at the ith lattice point for $i = 1, \ldots, n$, the full conditional distributions for this model are given by

$$\ln \frac{P(Z_i = 1)}{P(Z_i = 0)} = \boldsymbol{X_i}\boldsymbol{\beta} + \sum_{j \neq i} \eta_{ij} Z_j,$$

where $\boldsymbol{X_i}$ is the ith row of the design matrix, $\boldsymbol{\beta}$ are the regression parameters, and $\boldsymbol{\eta} = \{\eta_{ij}\}$ are dependence parameters such that $\eta_{ij} \neq 0$ iff Z_i and Z_j are lattice neighbors. The summation is the autocovariate, which models the dependence between Z_i and the remainder of the field, denoted Z_{-i}. In this study we consider only models for which $\eta_{ij} = \eta 1_{\{i \sim j\}}$ (where $1_{\{i \sim j\}}$ denotes the indicator function and $\sim$ denotes the neighbour relation). We assume pairwise-only dependencies. The joint distribution is given by

$$\pi(\boldsymbol{Z} \mid \boldsymbol{\theta}) = c(\boldsymbol{\theta})^{-1} \exp\Big(\sum_i Z_i \boldsymbol{X_i}\boldsymbol{\beta} + \frac{\eta}{2} \sum_{i,j} 1_{\{i \sim j\}} Z_i Z_j \Big),$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \eta)'$ and $c(\boldsymbol{\theta})$ is an intractale normalizing function which makes computation challenging for both ML and Bayesian inference. For more details, see [Hughes, 2011].

3 Methodology

In this study we use Voronoi tessleation, which partitions a plane into polygons based on proximity to a given set of points. It has been extensively used in clustering algorithms, especially to define neighbors in point pattern analysis. In this study we obtain the Voronoi tesselation for a given set of cluster points (which represents the number of clusters to be obtained). Each polygon is considered as a cluster. The data points or members in each cluster or polygon are "similar" in some sense and cross-cluster members are "dissimilar" in a corresponding sense.

The clustering problem can be considered as a combinatorial optimization problem. The CE method [Rubinstein and Kroese, 2004] is a leading evolutionary computing technique using a stochastic framework to solve both estimation and optimization problems. It has also been a successful methodology in multiple change-point problems; for example, see [Evans et al, 2011] and [Priyadarshana and Sofronov, 2015]. We use the CE method to estimate the locations of cluster points. In this study we use the "CEoptim" [Benham et al, 2015] and "deldir" [Turner, 2020] R packages for calculations.

4 Results and Conclusion

In this section, we discuss a general example with artificially generated data to illustrate the usefulness of the proposed algorithm. We generate a 30×30 matrix of independent Bernoulli random variables with four homogeneous clusters with the parameters (0.3,0.8,0.7,0.3). At this stage, we fix the number of clusters and apply our method to the above example to find exactly four clusters. Figure 1 shows the true profile of the data and the obtained clusters from the proposed CE algorithm.

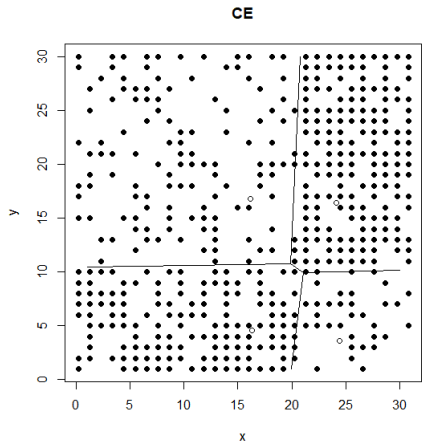


FIGURE 1. Clusters as determined by the CE algorithm.

Figure 2 represents the parameter values for the true profile (left) and the obtained clusters from the proposed CE algorithm (right). The obtained clusters are in excellent agreement with the true profile; the proposed CE algorithm produced only a very small difference between the estimate and the true distribution. We conclude that our algorithm can perform well in identifying homogeneous clusters in spatial binary data.

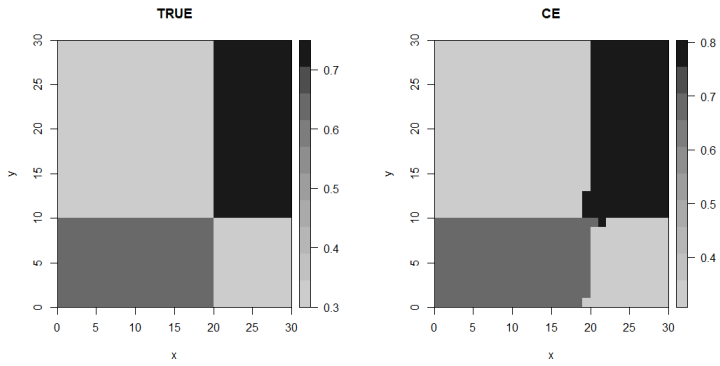


FIGURE 2. (L) true profile and (R) clusters as determined by the CE algorithm.

References

Ahuja, N. (1982). *Dot pattern processing using Voronoi neighborhoods.*. IEEE Transactions on Pattern Analysis and Machine Intelligence,**3**,336 – 343.

Besag, J. (1974). *Spatial interaction and the statistical analysis of lattice systems (with discussion).* Journal of the Royal Statistical Society, Series B: Methodological,**36**,192 – 236.

Benham, T., Duan, Q., Kroese, D. P., and Liquet, B. (2015). *CEoptim: cross-entropy R package for optimization.* arXiv preprint arXiv:1503.01842.

Evans, G.E, Sofronov, G.Y, Keith, J.M, Kroese D.P. (2011). *Estimating change-points in biological sequences via the Cross-Entropy method.*. Annals of Operations Research,**189(1)**,155 – 165.

Hughes J, Haran M, Caragea P.C. (2011). *Autologistic models for binary data on a lattice.* Environmetrics,**22(7)**,857 – 871.

Priyadarshana, W.J.R.M, and Sofronov, G. (2015). *Multiple Break-Points Detection in Array CGH Data via the Cross-Entropy Method.* IEEE/ACM Transactions on Computational Biology and Bioinformatics,,**12(2)**,487 – 498.

Rubinstein, R. and Kroese, D. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning.* New York: Springer-Verlag.

Turner, R. (2020). *deldir: Delaunay triangulation and Dirichlet (Voronoi) tessellation*, R package version 0.1-25.

Complex covariance structure: optimal sampling for an efficient estimation

Juan M. Rodríguez-Díaz[1]

[1] University of Salamanca, Spain

E-mail for correspondence: juanmrod@usal.es

Abstract: In most scientific disciplines models are proposed in order to describe different phenomena. In these models, the behavior of one or more variables is observed, trying to link these responses with other factors or covariates that may (at least partially) explain the former ones. An usual assumption for these observations is that they are independent, and many procedures have been developed for all kind of studies when assuming uncorrelated observations. However, it is clear that this assumption cannot be maintained for many real problems; several covariance structures can arise, and even appear combined, increasing the complexity of the models. Different situations will be examined, and some solutions for obtaining the 'best' designs for estimation of the parameters will be proposed employing optimal experimental design techniques.

Keywords: Covariance matrix; Multiresponse Models; Optimal Design of Experiments.

1 Covariance Structure and Optimal Design of Experiments

Let us initially assume the one-response linear model $y = \mathbf{f}(x)^T\boldsymbol{\beta} + u$, where $\boldsymbol{\beta}$ is the parameter vector of size m, u is the error term, and $\mathbf{f}(x) = (f_1(x), \ldots, f_m(x))^T$, with the $f_i(x)$ linearly independent in the experimental domain $\mathcal{X}$. An exact design ξ is a collection of points $\{x_1, \ldots, x_n\}$ of the independent variable, which represents the experimental conditions, with x_i in $\mathcal{X}$. In matrix notation it can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + u \quad ,$$

where $\mathbf{Y} = \{y_1, \ldots, y_n\}^T$ is the observations vector, $\mathbf{U} = \{u_1, \ldots, u_n\}^T$ the error terms, and $\mathbf{X} = (\mathbf{f}(x_1), \ldots, \mathbf{f}(x_n))^T$ the design matrix. For normally distributed random errors $u \equiv \mathcal{N}(0, \sigma^2)$ the Least Squares Estimators

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

(LSE) of the model parameters coincide with the Maximum Likelihood Estimators and are given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, with $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ and $\mathbf{X}^T\mathbf{X}$ known as the *Information Matrix* of the design ξ. But when a covariance structure is present the Generalized Least Squares (GLS) approach should be used. In this case the GLSE are $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, where $\boldsymbol{\Sigma} = Var(\mathbf{Y})$, and $Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}$. Now the information matrix of ξ will be

$$\mathbf{M}(\xi) = \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} \quad .$$

The standard method to obtain optimal designs requires to know the analytical expression of the model and to compute the derivatives with respect to the parameters in order to work with the linearized model. When no analytical expression of the model can be obtained, some methods for computing these derivatives can be employed, as described in Rodríguez-Díaz and Sánchez-León (2014).

When dealing with correlated observations the size of the design, n, should be fixed in advance. In many experiments it has no sense to take more than one observation to the same experimental unit at the same design point x, thus in the following it will be assumed that $x_i \neq x_j$ for all i, j. Usually, the aim is to find the points $\{x_1, x_2, ...\}$ where to take observations in order to get the best estimates of the parameters of the model, that is, the estimation with minimum variance, providing an *optimal design* for the model. The inverse of the information matrix is proportional to the covariance matrix (the generalized variance) of the parameter estimators) of the model; therefore the aim is usually to minimize (a convex function of) $\mathbf{M}^{-1}(\xi)$. However, there is not an only way of minimizing a matrix, giving rise to different *criterion functions*. A particular criterion function should be chosen depending on the objectives of the practitioners, for instance getting the best estimators of the parameters (one, some of them or all of them), or minimizing the variance of the predicted response. The most used criterion is D-optimality, which focuses on the determinant of the information matrix. A design ξ is D-optimal if maximizes this determinant, what is equivalent to minimize that of the covariance matrix. A-optimality pays attention to the trace of the covariance matrix, thus an A-optimal design minimizes the average of the variances of the estimators of model parameters. When the information matrix depends on unknown parameters, nominal values are needed for them and thus the obtained designs will be *locally optimal*, that is, they are good for (or close) those nominal values used in the computation. Fedorov and Hackel (1992), Pukelsheim (1986) or Atkinson et al. (2007) are classic references on optimal design of experiments.

In many studies, different kind of responses (say k of them) are measured, getting into the field of *multiresponse models*. These models have been studied from the point of view of optimality from different perspectives, considering in general correlation between different variables observed on

the same point, say $\mathbf{y}(x) = (y_1(x), \ldots, y_k(x))^T$ (that from now on will be denoted as one *sample*), but always assuming that the measures taken at different points, $\mathbf{y}(x)$ and $\mathbf{y}(x')$, were independent. However, this assumption may be unrealistic in some situations; for instance when the interest is in analyzing the evolution of a set of characteristics (variables) observed in a specific experimental unit at different time moments it seems clear that, apart from a *static* or *intra* covariance structure between differen type of observations taken at the same time, a *longitudinal* or *inter* correlation between the same type of measures obtained at different times should be taken into account (Rodríguez-Díaz and Sánchez-León, 2019a, 2019b).

2 Observing multiple subjects

To date, the double covariance structure has been considered only for studies carried out over one experimental unit, for which several variables were measured at different times (Rodríguez-Díaz and Sánchez-León, 2019a, 2019b). In the present work the design variable will be time as well, but now N subjects are supposed to be observed at different temporal points $t_1, \ldots, t_n$, which will be the design ξ. The design points t_i can denote any convenient temporal unit. It will be assumed that for each t_i in ξ the values of several characteristics $Y_1, \ldots, Y_k$ will be obtained for all of the subjects, and the aim will be to choose the 'best' design, the one giving the greatest information about the models describing the evolution of the response variables.
The same covariance structure can be assumed for every subject. Furthermore, it is quite usual to assume as well that the N subjects are independent. Two scenarios will be considered:

1. Different models of the variables for each subject (N k models)
2. The model of each variable is valid for all the subjects (k models)

With the above assumptions, some results will be obtained for Models (1) and (2):

- The D-optimal designs for the individual models of each variable in each subject are as well D-optimal for Model 1
- For this model, the parameters of the individual models can be estimated independently and do not depend on the intracovariance. However, the variance of the set of the parameter estimators for each subject do depend on it
- The D-optimal designs for the individual models of each variable in each subject are as well D-optimal for Model 2

- For Model 2 the estimation of the parameters of each response are the average of the corresponding estimations for each subject, and do not depend on the intracovariance, but their covariance matrix does depend on it

References

Atkinson, A.C., Donev A.N. and Tobias R.D. (2007). *Optimum experimental designs, with SAS*. Oxford University Press.

Fedorov, V.V., and Hackl P. (1997). *Model-oriented design of experiments*. New York: Springer-Verlag.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. Philadelphia, PA: SIAM.

Rodríguez-Díaz, J.M., and Sánchez-León G. (2014). Design optimality for models defined by a system of ordinary differential equations. *Biometrical Journal* 56 (5): 886–900.

Rodríguez-Díaz, J.M., and Sánchez-León G. (2019a). Efficient parameter estimation in mulltiresponse models measuring radioactivity retention. *Radiation and Environmental Biophysics* 58 (2): 167–182.

Rodríguez-Díaz, J.M., and Sánchez-León G. (2019b). Optimal designs for multiresponse models with double covariance structure. *Chemo. Intel. Lab. Syst.* 189: 1–7.

Bayesian modelling of complex functional forms

Bijit Roy[1], Emmanuel Lesaffre[1]

[1] Katholieke Universiteit Leuven, Belgium

E-mail for correspondence: bijit.roy@kuleuven.be

Abstract: We present a Bayesian approach for modelling quantities which are complex non-linear functions of other well modeled quantities. We apply our approach to model the body-mass index of infants. Our method benefits from recycling model fits to get posterior samples of the function, and use this for inferential purposes.

Keywords: Joint modelling; Growth curve; Bayesian analysis.

1 Introduction

We are often required to model complex (non-linear) functions of other quantities which are by themselves easy to model. For example height and weight of infants are easily modelled by growth curves, but BMI is a non-linear and a non-increasing function of age. A typical frequentist approach is to use the delta method to get error bounds for BMI using weight and height. However the delta method underestimates the standard error of the constructed function when the underlying quantities are not normally distributed (LePage and Billard, 1992), and it requires large sample sizes (Oehlert, 1992). We provide a Bayesian approach to address this problem. This involves first using a joint model for the multiple responses, and then drawing posterior predictive samples for the quantity we are interested in, and constructing credible intervals for inferential purposes.

2 Motivation: Modelling BMI using growth curves

It was of interest to evaluate potential differences in growth between formula-fed and breastfed infants, particularly BMI development. We will

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

focus on a subset of 80 subjects with two levels of demographic grouping (G1 and G2) and two levels of treatment (Formula and Breastfed). Weight and height measurements were taken at different time points till 2 years. A low sample size was intentionally chosen to demonstrate the applicability of our proposed method in a small data situation.

For modelling height and weight of infants several commonly used growth curve models exist, and each of these models have a particular age range where they perform best (Chirwa *et al*, 2014). Multilevel modelling is used to account for the individual level effects.

BMI has a deterministic relationship with weight and height, thus a joint bi-variate growth model for weight and height will allow us to construct a model for BMI. Due to the complex nonlinear relationship between BMI, weight, and height, it is not straightforward to get error estimates for BMI from a bi-variate weight and height model using standard frequentist techniques. A Bayesian joint model allows us to easily get posterior samples for BMI, construct credible intervals, and study the effect of interventions.

3 Modelling BMI from a joint weight-height model

We can write a bi-variate joint model as $(y_{1i}, y_{2i})^{\mathrm{T}} \sim BVN(\boldsymbol{\mu_i}, \Sigma_e)$, where $\boldsymbol{\mu_i} = (\beta_1^{\mathrm{T}} x_{1i} + z_{1i}^{\mathrm{T}} u_1, \beta_2^{\mathrm{T}} x_{2i} + z_{2i}^{\mathrm{T}} u_2)^{\mathrm{T}}$. Here β_1 and β_2 are the fixed effects for the two responses, and z_{1i} and z_{2i} are the corresponding random effects, with $(z_{1i}, z_{2i})^{\mathrm{T}} \sim MVN(\mathbf{0}, \Sigma_u)$. The benefit of this joint model in a Bayesian framework is that we can easily get credible intervals for any derived quantity at individual levels e.g. $g_i = f(y_{1i}, y_{2i})$ using a posterior predictive approach. Moreover by creating pseudo observations and averaging over the different population subgroups will allow us to easily obtain marginal effects of the different treatments on the different subgroups as well as credible intervals to judge the efficacy of the different treatments.

We have chosen the first order Berkey-Reed model (Berkey and Reed, 1987) to model the height and weight, as it has been shown to perform reasonably well during the ages 0 to 2 years. Along with the fixed effects corresponding to the intercept, time, $\log(time)$ and $1/time$, we also add the corresponding random effects to account for individual differences. In order to control for demographic grouping we added fixed effects for the groups G1 and G2, as well as interactions between the group variables and $time$, $\log(time)$ and $1/time$. The starting time for formula feeding varies in the non breast fed group, and children are breast fed till then. This is accounted for in the model by having fixed effects for the Formula-Group interactions with $time$, $\log(time)$ and $1/time$ beginning at treatment start time. This leads to a large model, with 28 fixed effects and 8 random effects.

The main challenge faced was the large model size. We needed to estimate 36 parameters for the random effects covariance matrix, as well as 3 for the error covariance matrix in addition to the 28 fixed effects. This leads to computational instability specially with small sample sizes. To overcome this

we used shrinkage priors. Shrinkage priors can be thought of as a Bayesian counterpart of penalized regression which can be used to avoid computational instabilities and over fitting in case of large number of predictors. We used Horseshoe priors (Carvalho *et al*, 2010). on the fixed effects and LKJ prior (Lewandowski *et al*, 2009) on the random effects correlation matrix, which leads to better performance (regarding convergence and model fit) in the case of low sample sizes.

4 Results

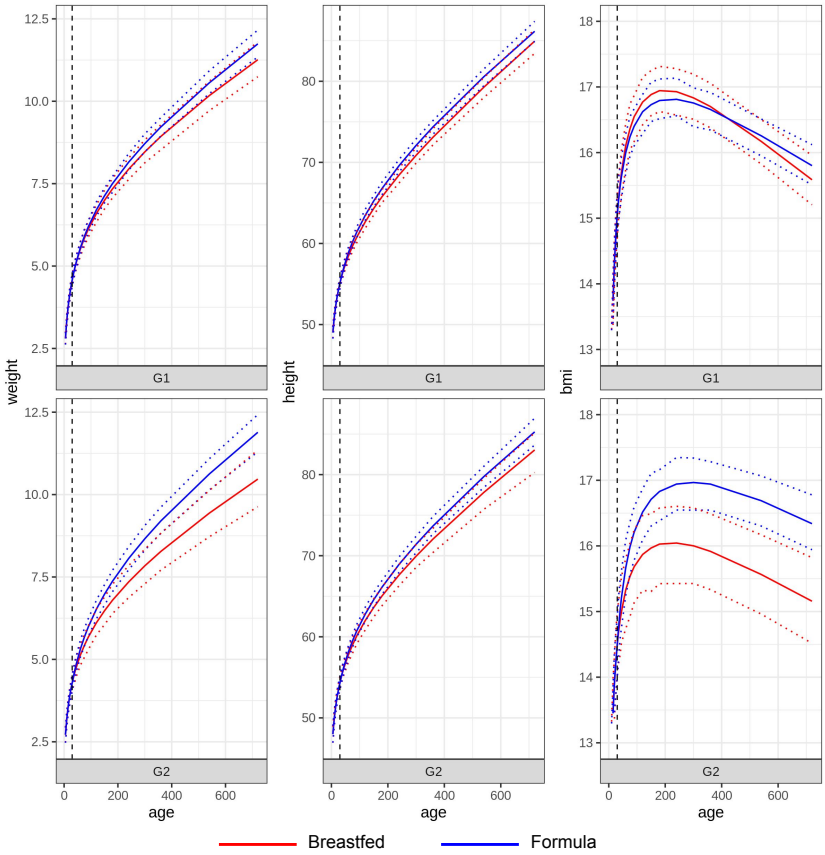


FIGURE 1. Marginal evolution plots for weight, height and BMI.

The joint random effects growth curve model does an excellent job of predicting individual trajectories for weight height and BMI. Figure 1 shows the marginal evolution plots for weight, height, and BMI for the two different subgroups and the two treatment levels as well as the associated

probability bounds. This allows us to draw conclusions on the efficacy of various treatments for different subgroups. We notice that for subgroup G1 there is no significant difference between the two treatment levels. On the other hand in subgroup G2 the treatments have a significant effect on the marginal profile of weight and BMI. We can see that the breastfed group has lower weight and BMI than the formula group.

5 Conclusion

In this article we have shown how we can use a Bayesian approach to simplify the modelling of variables which are functions of other well-modeled variables. We used existing well developed growth curve models for height and weight, combined them into a bivariate joint growth curve model, and used it to model the derived quantity BMI. This approach can be used for other derived quantities also. The benefit of using a Bayesian approach is using posterior predictive distributions, we can easily get individual as well as marginal predictions (and error bounds) for the derived quantities, without any extra modelling effort. The major challenge faced with this approach is computational instability due large number of parameters in the joint model (especially when dealing with mixed effects models with a large number of random effects). We explored the use of shrinkage priors to reduce the computational instability caused by large number of variables in the joint model. In this study we intentionally chose a small sample size to show the possibility of using this approach to fit a large joint model even to small data sets.

References

Berkey, C.S. and Reed, R.B. (1987). A model for describing normal and abnormal growth in early childhood. *Human Biology*, **59**, 973 – 987.

Carvalho, C.M., Polson, N.G., and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465 – 480

Chirwa, E.D. *et al.* (2014). Multi-level modelling of longitudinal child growth data from the Birth-to-Twenty Cohort: a comparison of growth models. *Annals of Human Biology*, **41 (2)**, 168 – 179.

LePage, R., and Billard, L. (1992). *Exploring the Limits of Bootstrap*. John Wiley and Sons.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, **100(9)**, 1989 – 2001.

Oehlert, G., W. (1992). A note on the delta method. *The American Statistician*, **46(1)**, 27 – 29.

Spatial Bayesian geo-additive modelling and prediction soil texture mapping in the Basque Country

Miguel Rua del Barrio[1,2], Joaquín Martínez Minaya[1], Lore Zumeta Olaskoaga[1], Ainara Artetxe[3], Nahia Gartzia-Bengoetxea[3], Ander Arias González[3] and Dae-Jin Lee[1]

[1] BCAM–Basque Center for Applied Mathematics, Bilbao, Bizkaia, Spain
[2] Universitat de València, Valencia, Spain
[3] Neiker–Basque Institute for Agricultural Research and Development

E-mail for correspondence: jomartinez@bcamath.org

Abstract: High-resolution soil maps are important for land use planning, agriculture crop production, forest management, hydrological analysis and environmental protection. In this work, we consider the analysis of soil texture samples in the Basque Country (i.e. the relative proportions of sand, silt, or clay in soil) and use covariate information to predict a high-resolution soil map. We propose the use of geo-additive models for modelling and predicting the spatial distribution of soil texture in a Bayesian framework for compositional data.

Keywords: spatial soil mapping; soil texture; compositional data; additive-log ratio transformation; Dirichlet regression

1 Motivation

Understanding the spatial distribution and variability of soil texture is essential for land use planning and other activities related to agricultural management and environmental protection (e.g.: prevent soil degradation, preserve soil functions and remediate degraded soil). This work is motivated by the "Land Use and Cover Area frame Statistical survey" (LUCAS) project aimed at the collecting harmonised data about the state of land use/cover over the extent of European Union. The work by Ballabio *et al.* (2016) mapped soil properties at a continental scale over the geographical extent of Europe. In this work, we are interested in mapping soil texture

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

distribution at a finer scale, based on 6736 soil samples (at 30 cm depth) surveyed in the Basque Country between 2009-2018. In addition to the soil texture samples, we considered the covariates information at a finer scale such as climatic variables (e.g. average precipitation, min/max temperature), the DEM (digital elevation model) and categorical information such as the geological information (e.g. lithology), or the land usage (pastures, extensive crops, vineyards, etc.).

Soil texture data are usually considered as compositional data (i.e. as relative proportion of different particles smaller than 2mm of sand (0.05-0.2 mm), silt (0.002-0.05 mm) and clay (0-0.002 mm), so the sum of the three components always equals 100) and hence considering each one of the textures separately would result in inconsistent results, like sum values above 100. Aitchison (1986) suggested that compositional variables should be transformed into log ratios. Given a vector of compositional data of D elements $\boldsymbol{x} = (x_1, x_2, ..., x_D)'$, such that $x_d > 0$ and $\sum_d^D x_d = 1$ for $d = 1, ..., D$. The additive log-ratio (ALR) transformation, defines a new vector $\boldsymbol{y} = ALR(\boldsymbol{x}) = (\ln(x_1/x_D), ..., \ln(x_{D-1}/x_D))'$, where x_D is the last component playing the role of the common divisor and $\boldsymbol{y} \in \mathbb{R}^{D-1}$. The drawback of ALR is that the components are treated asymmetrically and the interpretation of the results, may depend on the choice of the common divisor.

2 Bayesian Spatial Dirichlet regression

An alternative to the Aitchison approximation to model compositional data is to assume that the response variable follows a Dirichlet distribution. The Dirichlet distribution is the generalization of the widely known beta distribution, and it is defined by the following probability density:

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{d=1}^{D} y_d^{\alpha_d - 1}, \tag{1}$$

being $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)$ the vector of shape parameters for each category, $\alpha_d > 0 \ \forall d$, $y_d \in (0, 1)$, $\sum_{d=1}^{D} y_d = 1$, and $\mathrm{B}(\boldsymbol{\alpha})$ is the multinomial beta function, which serves as the normalizing constant. The sum of all α's, i.e. $\alpha_0 = \sum_{d=1}^{D} \alpha_d$, is usually interpreted as a precision parameter. The beta distribution is the particular case when $D = 2$. Hence, let $\boldsymbol{y} \sim \mathcal{D}(\boldsymbol{\alpha})$ denote a Dirichlet-distributed random variable. The expected values are $\mathrm{E}(y_d) = \alpha_d/\alpha_0$, the variances are $\mathrm{Var}(y_d) = [\alpha_d(\alpha_0 - \alpha_d)]/[\alpha_0^2(\alpha_0 + 1)]$ and the covariances are $\mathrm{Cov}(y_d, y_{d'}) = (-\alpha_d \alpha_{d'})/[\alpha_0^2(\alpha_0 + 1)]$.

Let η_{di} be the linear predictor for the i^{th} observation in the d^{th} category. Note that in the case of the Dirichlet regression, the logarithm of the shape parameters is employed as a linear predictor. A general formulation for the

geo-additive model has the form:

$$\boldsymbol{\eta}_{di} = \boldsymbol{x}_i'\beta + f(\boldsymbol{s}_i) + \sum_{j=1}^{J} f_j(\boldsymbol{z}_i) \tag{2}$$

where $\boldsymbol{x}_i'\beta$ are the linear effects (e.g.: land usage, lythology), $f(\boldsymbol{s}_i)$ is the spatial component of the geographical coordinates s_i (longitude and latitude) and $f_j(\boldsymbol{\nu}_i)$ are non-linear terms (e.g.: elevation, average precipitation, and min/max temperature). For the model in Eq. (2), we consider a Bayesian approach using Markov chain Monte Carlo simulation techniques to get the posterior distributions of the parameters of interest, and the predictions. In particular, we used the implementation in the `R` package `bamlss` (Umlauf *et al.*, 2019) where $f(s_i)$ and $\sum_{j=1}^{J} f_j(\boldsymbol{z}_i)$ are modelled by tensor product smooths and additive terms of penalized splines (Eilers *et al.*, 2015).

3 Application

We compared two approaches using `bamlss` as an unified framework: i) the Dirichlet geo-additive model (DGM) in Eq. (2) and ii) the model formulation in Eq. (2) with a multivariate normal geo-additive model (MGM) based on the additive log-ratio transformation by Aitchison (1986) for the $D-1$ elements of the soil texture data (i.e. $D = 3$) playing with the common divisors (this leads to 3 different bivariate response models, i.e. x_D = {sand, silt, clay}). Figure 1 shows the predicted soil texture map in the Basque Country at a fine-scale using the soil texture classification by the USDA (United States Department of Agriculture, see Soil Survey Staff, 1993), the map is the predicted mean obtained from the Dirichlet model (predicted maps based on the multivariate normal geo-additive models gave similar results and are not shown).

In terms of model comparisons, we considered goodness-of-fit measures to assess the adequacy of the fitted models. While Aitchinson (1986) suggested working with log-ratios of the compositional data to be able to apply the traditional multivariate techniques, for the Dirichlet model, we need a measure to evaluate the explained variation of our model as the usual R^2, is not an accurate measure for compositional data. Hijazi (2015) proposed R^2-type measures based on model likelihoods, total variability and sums of squares. These measures were computed for the proposed Dirichlet and Multivariate normal geo-additive models providing similar performances.

4 Conclusions

In this work, we have compared different models for the analysis of compositional soil texture data. We considered the `bamlss` methodology as a

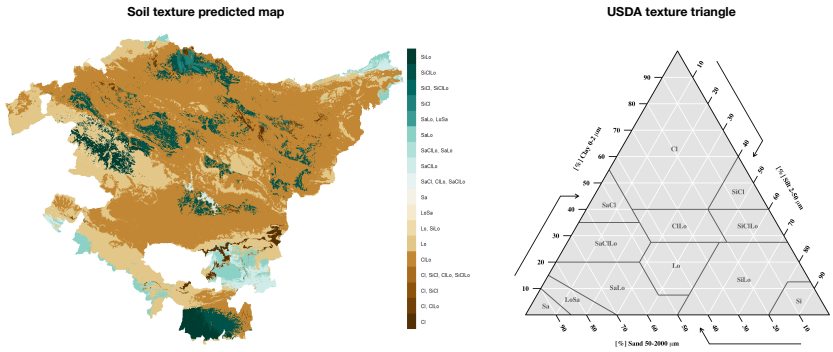


FIGURE 1. Predicted soil texture map (left) with USDA classification of soil textures (right).

unified framework for prediction in a Bayesian geo-additive framework due to the flexibility in incorporating linear and non-linear effects using penalized splines. The predicted maps at a fine-scale provide valuable maps for water management, hydrology, and particularly for the agricultural and forestry sectors in the Basque Country.

References

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

Ballabio, C., Panagos, P., Monatanarella, L. (2016). Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma*, 261, 110 – 123.

Eilers, P.H.C., Marx, B.D. and Durbán, M. (2015). Twenty years of P-splines. *SORT*, 39(2):149 – 186.

Hijazi, R.H. (2015). Residuals and diagnostics in Dirichlet regression. *Tech report of United Arab Emirates University, Department of Statistics.*

Soil Survey Division Staff (1993). *Soil survey manual*. Soil Conservation Service. United States Department of Agriculture. Vol. 3, pp. 63 – 65.

Umlauf, N., Klein, N. and Zeiles, A. (2018). BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond). *Journal of Computational and Graphical Statistics*, Vol.27, issue 3, 612 – 627.

A semi-latent class model for estimating the time of differentiation of cognitive decline between cases and controls

Corentin Segalas[1,2], Hélène Jacqmin-Gadda[2]

[1] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, U.K.
[2] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, Bordeaux, France

E-mail for correspondence: corentin.segalas@lshtm.ac.uk

Abstract: In this work, we propose a semi-latent class random changepoint mixed model that allows the estimation of the time of differentiation between cognitive decline of future demented and normal subjects from a nested-case-control study. Cases are assumed to have a random changepoint trajectory while controls can have either a linear trajectory or a random changepoint trajectory where the class membership follows a logistic model. The log-likelihood of the model is derived and can be optimized using a Levenberg-Marquardt algorithm with Gaussian quadrature for numerical integration. The model is estimated on the Paquid cohort of elderly with very long follow-up (25 years) to estimate the delay between the beginning of the decline of a test of verbal fluency and the onset of dementia.

Keywords: Dementia; Mixed model; Random changepoint.

1 Introduction

Dementia is a syndrome that affects the cognitive abilities of a subject. The pre-diagnosis phase last around fifteen years and during this phase the cognitive decline trajectories are non-linear and heterogeneous (Amieva et al, 2014). Longitudinal data is available in cohorts where the cognitive decline is measured by collecting psychometric scores over time.
To study the cognitive decline in the pre-diagnosis phase of dementia, random changepoint mixed model have been proposed in the literature (van den Hout *et al.*, 2011). By fitting a smooth linear-linear trajectory and

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

including random effects, they model both the non-linearity and heterogeneity of cognitive decline.
When looking retrospectively at cases only and using time to dementia as the timescale, the estimated mean changepoint is the mean delay between the acceleration of cognitive decline and dementia diagnosis. However, with such model, the estimated changepoint tend to identify a late acceleration of cognitive decline, happening only a few years before diagnosis (Segalas *et al.*, 2020). One might rather be more interested in the time at which the cognitive decline trajectory of a demented subject begins to differ from the cognitive decline trajectory of a non-demented subject.
For this objective, cases and non-cases need to be modeled together. We propose to model cognitive evolution using a two-class model with a linear trend for one class and the same linear trend up to a certain date where the decline accelerates for a second class. From this date and up to the diagnosis, trajectories are nonlinear with, possibly, a late acceleration just before diagnosis. In such a model, the changepoint would identify the mean time at which, case trajectory begins to differ from linear cognitive decline in normal ageing.

2 The semi-latent class random changepoint model

We consider a nested case-control study where incident cases of dementia diagnosed during the follow-up of a cohort are matched to controls according to *a priori* defined characteristics with the condition that controls are observed and free of dementia at the visit of diagnosis of the matching case. The delay for a control is the delay to diagnosis of the matching case. We note δ_i the case indicator, 1 for cases and 0 for controls. We denote $Y_i(t_{ij})$ the value of marker Y for subject i at time t_{ij} with $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. The two-class model is written

$$Y_i(t_{ij}) = \beta_{0i} + \beta_{1i} t_{ij} + c_i \beta_{2i} f(t_{ij} - \tau_i, \eta) + \varepsilon_{ij} \tag{1}$$

where c_i is an indicator that equals 1 for subjects with a random changepoint trajectory and 0 otherwise, f is a function based on I-spline that represents the difference from the linear trajectory after the time of differentiation τ_i and depends upon parameters η. We assume that $\beta_{ki} = \beta_k + b_{ki}$ where $b_i = (b_{0i}, b_{1i}, b_{2i})^{\mathrm{T}} \sim \mathcal{N}(0, B)$ with B a positive matrix and that $\tau_i = \mu_\tau + \sigma_\tau \tilde{\tau}_i$ where $\tilde{\tau}_i \sim \mathcal{N}(0, 1)$ is independent from b_i. The residual errors ε_i are assumed to follow a centered Gaussian distribution with diagonal variance matrix $\sigma_\varepsilon \mathbb{I}_{n_i}$ and are assumed independent from all the random effects.
In this model, t_{ij} denotes the delay as defined is our nested case-control study design. β_0 is the mean value of the marker for subjects in the linear class at the time of the case diagnosis, β_1 is the mean slope of the cognitive

decline during the normal cognitive ageing phase. For subjects whose trajectory presents a random changepoint, f models smoothly the difference between this normal cognitive ageing and a pathological cognitive decline while β_2 measures its mean intensity.
In the nested case-control design, some subjects are controls because they are free of dementia at a certain date even though they might develop dementia at a later visit. Therefore, we can not realistically assume that all controls are on the linear class and we chose to model the probability for a control of being in the changepoint class by a logistic model

$$\pi_i = \mathbb{P}(c_i = 1|X_i, \delta_i) = \left(\frac{\exp(\eta^\top X_i)}{1+\exp(\eta^\top X_i)}\right)^{1-\delta_i}$$

that can depend upon some covariates X_i.

3 Estimation

The log-likelihood of model (1) is written

$$\ell_N(Y;\theta) = \ell^0_{N_0}(Y;\theta) + \ell^1_{N_1}(Y;\theta) \tag{2}$$

where θ is the vector of all parameters from the model, Y the complete data from the nested case-control study, N_0 and N_1 denotes respectively the number of controls and cases such as $N = N_1 + N_2$ and where

$$\ell^0_{N_0}(Y;\theta) = \sum_{i=1}^{N_0} \log[(1-\pi_i)f(Y_i|c_i = 0,\theta) + \pi_i f(Y_i|c_i = 1,\theta)],$$

$$\ell^1_{N_1}(Y;\theta) = \sum_{i=1}^{N_1} \log f(Y_i|c_i = 1;\theta)$$

are the contributions of the controls and the cases respectively to the log-likelihood. $f(Y_i|c_i = 0,\theta)$ is the individual contribution of a subject to the likelihood of the linear class and follows a multivariate Gaussian density with mean 0 and variance $Z_{0i}B_0Z_{0i}^{\mathrm{T}}+\sigma^2_\varepsilon\mathbb{I}_{n_i}$ where Z_{0i} is a $n_i\times 2$ matrix with rows $(1, t_{ij})_{j=1,\ldots,n_i}$ and B_0 a 2×2 definite positive matrix, variance of the random effects $(b_{0i}, b_{1i})^{\mathrm{T}}$. $f(Y_i|c_i = 1,\theta) = \int f(Y_i|c_i = 1,\tilde{\tau}_i;\theta)f(\tilde{\tau}_i)\mathrm{d}\tilde{\tau}_i$ is the individual contribution of a subject to the likelihood of the changepoint class as defined by the random changepoint model (1).
The complete log-likelihood (2) can then be estimated using Gauss quadrature for the numerical integration and Levenberg-Marquardt algorithm (Marquardt, 1963) for the optimization procedure. The estimation procedure will be validated in a simulation study.

4 Application

The method will be applied to data from the french cohort Paquid (Letenneur *et al.*, 2014). Our objective is to evaluate the time of differentiation between cognitive trajectories of future demented and normal subjects as measured by the Isaacs Set Test which assess verbal fluency.
We built a nested case-control study from the 901 incident cases of dementia from Paquid. For each of these cases, we matched one control with the same age (±2 years), same educational level, same sex and with the condition that the control has to be observed non demented at the visit of diagnosis of the case.
We estimated for the Isaacs Set Test a simplified model where $c_i = \delta_i$, *i.e.* that assumes a linear trajectory for all controls and a changepoint trajectory for all cases. The mean estimated time of differentiation between cases and controls was estimated at around -11.1 years before diagnosis with a 95% confidence interval of $[-12.5; -9.7]$. We will compare these estimates to those of the semi-latent class model (1).

References

Amieva, H. *et al.* (2014). Compensatory mechanisms in higher-educated subjects with Alzheimers disease: a study of 20 years of cognitive decline. *Brain: A Journal of Neurology* **137(4)**, 1167 - 1175.

Letenneur, L. *et al.* (1994). Incidence of dementia and Alzheimers disease in elderly community residents of south-western France. *International Journal of Epidemiology* **23(6)**, 1256 - 1261.

Marquardt, D. W. (1963). An Algorithm for Least Square Estimation of Non-Linear Parameters. *SIAM Journal on Applied Mathematics* **11**, 431 - 441.

Segalas, C. *et al.* (2020). A curvilinear bivariate random changepoint model to assess temporal order of markers. *Statistical Methods in Medical Research*

van den Hout, A. *et al.* (2011). Smooth random change point models. *Statistics in Medicine* **30(6)**, 599 - 610.

A flexible marginal rate model for recurrent events with a zero-recurrence proportion

Ivo Sousa-Ferreira[1], Cristina Rocha[1], Ana Maria Abreu[2]

[1] Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal and CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal

[2] Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal and Centro de Investigação em Matemática e Aplicações, Portugal

E-mail for correspondence: ivo.ferreira@staff.uma.pt

Abstract: A new marginal rate model for gap times between recurrent events is proposed, which is derived from a non-homogeneous Poisson process (NHPP). Since the distribution of the gap times often requires flexible shapes of the rate function, the approach taken here is to model the baseline log-cumulative rate function as a restricted cubic spline function of log time. Moreover, a proportion of subjects that will never experience any recurrence is incorporated. The proposed model allows covariates in both the latency and incidence components. An application to a real data set is also provided for illustrative purposes.

Keywords: Gap times; Non-homogeneous Poisson process; Recurrent events; Restricted cubic spline; Zero-recurrence.

1 Introduction

Recurrent events data arise frequently in medical studies where each subject may experience a particular event repeatedly over time. In this work, we follow the approach of Zhao and Zhou (2012) to model the gap times between recurrent events, considering that the recurrence process is derived from a NHPP for which the gap times are generally not independent. However, we assume a completely parametric baseline rate function in which the covariates have a multiplicative effect. The main challenge here is to select the most appropriate baseline form. Motivated by Royston and Parmar (2002), we propose to use restricted cubic splines to capture, in a flexible way, how the rate evolves over time.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Model formulation

Suppose that there are n independent subjects in study and that each one can experience a maximum of K_i $(i = 1, \ldots, n)$ recurrences of an event. For the ith subject, let T_{ik} be the calendar time related with the kth event $(k = 1, \ldots, K_i)$ and $Y_{ik} = T_{ik} - T_{i,k-1}$ be the gap time between two consecutive events, where $0 \equiv T_{i0} < T_{i1} < \ldots < T_{iK_i}$. Based on Zhao and Zhou (2012), the recurrence process is assumed to be a NHPP with independent increments. Then, we consider a multiplicative model in which the marginal rate function of the recurrence process is given by

$$h(y|t_{i,k-1}, \boldsymbol{z}_{ik}) = h_0(y + t_{i,k-1}) \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ik}), \tag{1}$$

where $h_0(\cdot) > 0$ is a baseline rate function, $\boldsymbol{z}_{ik}$ is the covariate vector of subject i and $\boldsymbol{\beta}$ is the corresponding regression coefficients vector. Following the approach of Royston and Parmar (2002), we propose to model the log-cumulative baseline rate function as a restricted cubic spline function of log time, which provides analytically tractable expressions. From (1), the cumulative rate function of the recurrence process is

$$\begin{aligned} H(y|t_{i,k-1}, \boldsymbol{z}_{ik}) &= \Big[\exp\big(\log H_0(y + t_{i,k-1}) \big) - \exp\big(\log H_0(t_{i,k-1}) \big) \Big] \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ik}) \\ &= \Big[\exp\Big(s\big(\log(y + t_{i,k-1}); \boldsymbol{\gamma} \big) \Big) - \exp\Big(s\big(\log t_{i,k-1}; \boldsymbol{\gamma} \big) \Big) \Big] \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ik}), \end{aligned}$$

where $\log H_0(y + t_{i,k-1}) = \log \int_0^y h_0(u + t_{i,k-1}) du$ is the log-cumulative baseline rate function. For pre-specified m distinct internal knots $r_1 < \ldots < r_m$ with $r_{\min} < r_1$ and $r_{\max} > r_m$ boundary knots, the restricted cubic spline function of $x = \log t$ may be written as

$$s(x; \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x + \gamma_2 v_l(x) + \ldots + \gamma_{m+1} v_m(x),$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{m+1})'$ is the parameters vector, $v_l(x)$ is the lth basis function. The complexity of the curve is regulated by the number of degrees of freedom (d.f.), given by d.f.$= m+1$. Conventionally, when d.f.$= 1$ it means that no internal knots are specified, and so $s(x; \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 x$. This particular case gives rise to the Weibull marginal rate model. In turn, when $\gamma_1 = 1$ it leads to the classical HPP that has constant rate.

In some scenarios, it might exist a tangible proportion of the population under study that becomes recurrence free. Let Y_{i1} denote the first gap time for the ith subject in the population. Therefore, we can observe two cases: i) if $K_i > 1$, subject i experiences at least one recurrence, so he is a recurrent subject; and ii) if $K_i = 1$, subject i may either be a recurrent subject with probability π or a zero-recurrence subject with probability $1 - \pi$. Then, Y_{i1} has a survival function given by

$$P(Y_{i1} > y) = 1 - \pi + \pi P(Y_{i1} > y | T_{i0} = 0),$$

where $P(Y_{i1} > y | T_{i0} = 0)$ is the (proper) survival function of the first gap time. A natural extension is to assume that covariates influence the proportion of recurrent subjects via a logistic regression model.

The inferential procedure is based on the maximum likelihood method, assuming a non-informative right-censoring mechanism. For each subject i, we define $\delta_i = I(K_i > 1)$ and $K_i^* = \max(K_i - 1, 1)$. The likelihood function is expressed as

$$\mathcal{L} = \prod_{i=1}^{n} \left\{ \pi_i \prod_{k=1}^{K_i^*} f(y|t_{i,k-1}, \boldsymbol{z}_{ik}) \right\}^{\delta_i} \left\{ 1 - \pi_i + \pi_i P(Y_{i1} > y | T_{i0} = 0) \right\}^{1-\delta_i},$$

where $f(y|t_{i,k-1}, \boldsymbol{z}_{ik})$ is the probability density function. The computational implementation of the maximum likelihood method was performed in `R` software (R Core Team, 2020), using the usual optimization procedures.

3 An application to re-hospitalization data

A data set on re-hospitalization of patients diagnosed with colorectal cancer is analysed. The data are the gap times (in days) of successive re-hospitalizations of 403 patients after removing their tumours. There is a total of 861 re-hospitalizations, ranging from 1 to 22, with mean 2.3 and median 1.0. About 199 patients (49.4%) had no recurrence at all. Some covariates in the data set are: chemotherapy; gender; Dukes' stage; and Charlson comorbidity index. The data are available in the `R` library `frailtypack`.
To choose the proper number of d.f., in the models without covariates and zero-recurrence proportion, we use the Akaike (AIC) and Bayesian (BIC) information criteria. So, models with 1 to 4 d.f. were fitted. The AIC values (6883.1, 6871.3, 6872.4 and 6872.9) and BIC values (6892.6, 6885.5, 6891.4 and 6896.7) indicate that 2 d.f. is the most adequate choice. Then, the proposed model was applied and the results are summarized in Table 1.
In our model, the reference group consists of male patients who did not receive chemotherapy, with Dukes' stage A−B and Charlson index 0. For this group, the zero-recurrence proportion is $1 - (1/\{1 + \exp(-0.180)\}) = 0.455$. The chemotherapy coefficient estimates are negative in both components, with a non-significant effect on the time to readmission. In the rate

TABLE 1. Parameter estimates of the flexible marginal rate with zero-recurrence proportion for the re-hospitalization data.

Rate component (latency)				Logistic component (incidence)			
Parameters	Estimate	SE	p-value	Parameters	Estimate	SE	p-value
γ_0	−5.868	0.642	−				
γ_1	1.019	0.170	−				
γ_2	0.001	0.005	−	Intercept	0.180	0.268	0.500
Chemo	−0.048	0.115	0.677	Chemo	−0.341	0.266	0.200
Gender (female)	−0.444	0.111	<0.001	Gender (female)	−0.163	0.248	0.511
Dukes' stage				Dukes' stage			
C	0.198	0.126	0.116	C	0.329	0.261	0.207
D	0.641	0.147	<0.001	D	1.979	0.593	<0.001
Charlson index							
1 − 2	0.357	0.208	0.087				
≥ 3	0.528	0.117	<0.001				

Male: reference for gender; A−B: reference for Dukes' stage; 0: reference for Charlson index.

component, recurrent females have a significantly lower risk of readmission compared with recurrent males. The other two important risk factors are the Dukes' stage D and Charlson index ≥ 3. In the logistic component, only the Dukes' stage D has a significantly increasing effect, which means that patients in this stage have a lower chance of being recurrence free.

4 Concluding remarks

In this paper, we propose a new flexible parametric model derived from a NHPP to analyse gap times between recurrent events. The model is formulated considering each gap time conditional on the prior recurrence time, in such a way that the relationship between successive gap times is no longer a problem. Furthermore, it is characterized by a fully parametric baseline rate function, based on restricted cubic splines. The proposed model also incorporates a proportion of zero-recurrence subjects to conveniently take into account the existence of subjects that will never experience any recurrence. Although it is not shown here, the Cox-Snell residual plots were used to informally evaluate the models goodness-of-fit, allowing to confirm that the assumptions underlying the final model are plausible to analyse the re-hospitalization data.
Finally, we aim to extend our model in order to deal with the unobserved heterogeneity across subjects, including a random effect term and thus obtaining a frailty model.

Acknowledgments: I. Sousa-Ferreira is grateful to the *Universidade de Lisboa* for his PhD scholarship. The research was partially sponsored by portuguese funds through *FCT – Fundação para a Ciência e a Tecnologia*, projects UIDB/00006/2020 (*Centro de Estatística e Aplicações*) and UIDB/04674/2020 (CIMA – Center for Research in Mathematics and Applications, from the Statistics, Stochastic Processes and Applications group).

References

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Royston, P., and Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, **21**(15), 2175 – 2197.

Zhao, X., and Zhou, X. (2012). Modeling gap times between recurrent events by marginal rate function. *Computational Statistics & Data Analysis*, **56**(2), 370 – 383.

Estimation of the Transition Probabilities condition on repeated measures in Multi-state models

Gustavo Soutinho[1], Luís Meira-Machado[2], Pedro Oliveira[1]

[1] EPIUnit, ICBADS, University of Porto, Portugal.
[2] Department of Mathematics and Centre of Molecular and Environmental Biology (CBMA), University of Minho, Portugal.

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

Abstract: The topic of joint modeling of longitudinal and survival data has received remarkable attention in recent years. In cancer studies for example, these models can be used to assess the impact that a longitudinal marker has on the time to death or relapse. Analyzes of such studies, in which individuals may experience several events, can be successfully performed by multi-state models. The goal of this work is to introduce feasible estimation methods for the transition probabilities conditionally on covariates observed with repeated measures through the use of the landmark methodology and the adaptation of existing methods for joint modeling of longitudinal and survival data. Results of the simulation studies confirm the superiority of the proposed estimator when compared to methods that do not take in consideration the effect of the covariate on the estimated transition probabilities or do not assume all the existence of repeated measures (Breslow estimator).

Keywords: Joint modeling, Markov assumption, Multi-state models, Transition probabilities.

1 Introduction

Multi-state model is a model for a time continuous stochastic process which can be used to describe complex event history data with several events (Meira-Machado and Sestelo, 2019). In medical science studies beyond the times-to-event a main goal is to identify the impact of a set of repeated measures as a time-dependent covariate on the transition among states. In order to produce valid inferences in these cases a joint modeling analysis of longitudinal and multiple survival outcomes are required (Rizopoulos,

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2012). The final model is built using two sub-models; a longitudinal sub-model (such as a linear mixed effects model) and a time-to-event sub-model (such as a proportional hazards model) for each transition intensity which are linked through an association structure quantifying the relationship between the outcomes of interest. The background concepts related to the extension of the joint modeling to multi-state models can be found in Ferrer et al (2016). The aim of this paper is propose a feasible estimation method for the transition probabilities conditionally on covariates observed with repeated measures. To this end we will use the subsampling approach, also termed as landmarking, proposed by de Uña-Álvarez and Meira-Machado (2015), combined with methods proposed by Rizopoulos (2012). The landmark methods considers subsamples of individuals of the data that belong in a given state at a pre-specified time point and gives rise to consistent estimators regardless the Markov assumption.

1.1 Joint multi-state model specification

The joint modeling approach for multi-state models can be described by a linear mixed effect model and a survival sub-model for each transition. The longitudinal sub-model follows the gaussian assumptions and the observed measure Y_{ij} at time t_{ij} is given by $Y_{ij} = X_i\left(t_{ij}\right)^T \beta + Z_i\left(t_{ij}\right)^T b_i + \varepsilon_{ij}$, where $X_i\left(t_{ij}\right)$ and $Z_i\left(t_{ij}\right)$ represent the vectors of time-dependent covariates of the individual and b_i is the vector of random effects with $b_i \sim N(0, \Sigma)$. The β parameter is a fixed vector and $\varepsilon_i \sim N\left(0, \sigma^2 I_{n_i}\right)$ where n_i is the number of longitudinal measures by individual (Ferrer, 2016).
The time-to-event outcome at time t from state h to state k, with $h, k \in S$ the finite state space, is modeled by a proportional hazards sub-model which takes the following form $\lambda^i_{hk}(t|b_i) = \lambda_{hk,0}(t)\exp\left\{X^{ST}_{hk,i}\gamma_{hk} + W_{hk,i}(b_i,t)\eta_{hk}\right\}$, where $\lambda_{hk,0}(.)$ is a parametric baseline intensity (with weibull, exponential or piecewise constant distributions, for instance). The baseline covariates are denoted by w_i with coefficients γ_{hk}. The multivariate function $W_{hk,i}(b_i, t)$ defines the dependence structure between the longitudinal and multi-state process and represents the true and unobserved value of the longitudinal outcome for patient i at time t. The association between the longitudinal and the times-to-event for each transition is given by η_{hk}.

1.2 Estimation and Dynamic predictions of the transition probabilities

In this study the maximum likelihood estimation for joint models will be used to estimate the parameters of the joint multi-state model under the landmark approach described in Uña-Álvarez and Meira-Machado (2015). The maximization of the log-likelihood function will be done using an EM algorithm coupled with a quasi-Newton algorithm in case of slow convergence. As referred above the aim of this paper is to estimate the conditional transition probabilities $p_{hj}(s, t \mid Y)$ where Y denotes a covariate

with longitudinal measures (as tumor markers measured at different moments) $\widetilde{y}_i(v) = \{y_i(u), 0 \leq u \leq v\}$. For each individual the transition probability is estimated and is assumed that the patient has survived up to the last time point s (Rizopoulos, 2012)

2 Simulation study

The longitudinal and multi-state data were generated through a joint modeling with 1000 replicates with 400 individuals given by $Y_{ij} = \beta_0 + \beta_1 \times t_{ij} + b_{i0} + b_{i1} \times t_{ij} + \varepsilon_{ij}$ and $\lambda^i_{hk}(t|b_i) = \lambda_{hk,0}(t) \exp\{\gamma_{hk} + W_{hk,i}(b_i, t)\, \eta_{hk}\}$, where $h \in \{0,1\}$, $k \in \{0,1,2\}$ and $b_i \sim N\left((0,0)^T, \begin{pmatrix} 20 & 0.2 \\ 0.2 & 0.02 \end{pmatrix}\right)$. The longitudinal times, initially were the same for each individual, given by $t_{ij} = 0.33, 0.66, \cdots, 16.50$ and the $\varepsilon_i \sim N(0, 18)$. The parametric baseline intensities were obtained from exponential distributions with rate parameters 3, 1.7 and 0.5. We took the value 2 for the γ_{hk} and for η_{hk} we took the values -0.7, -0.7 and -0.6 for the transitions $0 \to 1$, $0 \to 2$ and $1 \to 2$, respectively. The vector of true transition times, $T_i^* = \left(T^*_{i,01}, T^*_{i,02}, T^*_{i,12}\right)$, were generated following the procedures described in Beyersmann et al. (2011). By comparing T_i^* and C_i, the vector of times $T_i = \min(T_i^*, C_i)$, where C_i denotes the censoring times, which characterizes the multi-state process, was deduced. The longitudinal measurements, generated from the linear mixed sub-model, were truncated at T_{i1} the first observed time of the multi-state process.

2.1 Results

The transition probabilities for the Landmark approach (`LM`), Breslow's method (`BRES`) and Joint Modeling-Landmark estimator (`JMLM`) were obtained through Monte Carlo simulation with 1000 replicas with 400 individuals. For each replica, eight individuals were retained with the purpose to identify the influence of the longitudinal marker on the estimation of the transition probabilities (decreasing, constant, increasing and random values of the marker). The results reveal the `JMLM` estimator has a better performance for all $p_{00}(8, t \mid Y)$ independently of the longitudinal marker trend. In fact the boxplots of `BRES` and `LM` estimators show a systematic bias and consequently appear to be inadequate to identify the evolution of the repeated marker of the individuals (Figure 1). The variability of the `JMLM` estimates increase as the difference between t and $s = 8$ is greater but even though the proposed `JMLM` still produces estimates with less bias in accordance with the ratio between the mean square errors (MSEs) for the transition probabilities $\hat{p}_{00}(8, t)$ and $\hat{p}_{11}(8, t)$ with $t = \{10, 12, 14, 16, 18\}$.

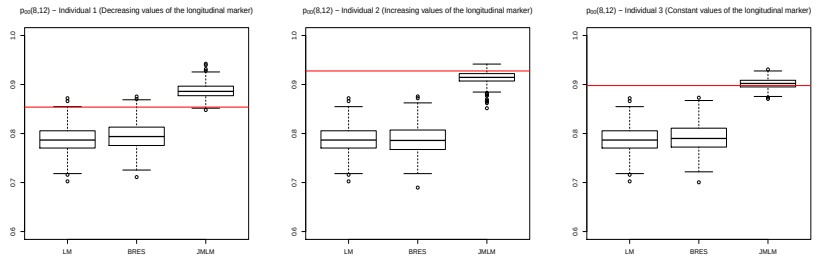


Figure 1: Boxplots of the M = 1000 estimates of the transition probabilities.

By comparing the values of $p_{00}(8, 12 \mid Y)$ and $p_{00}(8, 18 \mid Y)$, between individual 1 and individual 2, we may conclude that a increasing trend on the longitudinal marker means higher true value. From the results is also possible to observe the ability of the `JMLM` to reflect the evolution of the longitudinal measures of the marker. In fact, for instance, considering $\hat{p}_{00}\,(8, 12)$, for individual 2 with an increasing trend of the longitudinal marker, as the Breslow estimator takes into account the higher value the transition probabilities decrease comparing to the `LM` estimator. However the effect of the previous repeated measures have as consequence the increase of the `JMLM` estimation, following the true values.

3 Conclusions

Results obtained from simulation studies and in the real data application confirmed the good performance of the `JMLM` estimator, providing accurate estimated transition probabilities. The proposed method also demonstrated to have more sensibility to reflect the evolution of the longitudinal measures when comparing to the Breslow's based method which only makes use of a single value of the covariate.

References

Beyersmann, J., Allignol, A. and Schumacher, M (2011). *Competing risks and multistate models with R*. Springer.

de Uña-Álvarez, J., Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. *Biometrics*, **71**, 141 – 150.

Ferrer, L., Rondeau, V., Dignam, J.J., Pickles, T., Jacqmin-Gadda, H, and Proust-Lima, C. (2016). Joint modelling of longitudinal and multistate processes: application to clinical progressions in prostate cancer. *Statistics in Medicine*, **35(22)**: 3933 – 3948.

Meira-Machado, L. and Sestelo, M. (2019). Estimation in the progressive illness-death model: A nonexhaustive review. *Biometrical Journal*, **61(2)**, 245 – 263.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Robust statistical boosting with quantile-based loss functions

Jan Speller[1], Christian Staerk[1], Andreas Mayr[1]

[1] Department of Medical Biometrics, Informatics and Epidemiology, University of Bonn, Germany

E-mail for correspondence: speller@imbie.uni-bonn.de

Abstract: We investigate robust loss functions in statistical boosting, which is particularly suitable for high-dimensional data situations. To achieve robustness against outliers in the outcome variable we consider different robust losses. The stepwise boosting algorithm implicitly reweights the residuals in each iteration with the gradient of the loss function. For composite losses, e.g. the Huber and Bisquare loss, there is a cut-off value to choose. For this purpose, a fixed quantile for the amount of outliers is used that adapts this value in each iteration to the size of this residuals. As an application we investigate the performance of the boosting methods for various amounts of outliers in a high-dimensional riboflavin data set.

Keywords: Bisquare Loss; Gradient Boosting; Huber Loss; Robust Regression.

1 Introduction

Modern tools for data analysis are becoming more frequently confronted with large and complex data sets. Real data sets often contain outliers or are corrupted in some way. For the case of linear regression we propose a robust and data-adapted use of statistical boosting. The boosting algorithm uses ideas from machine learning and iteratively updates the estimated coefficients for a model using a component-wise gradient method of steepest decent (for a non-technical introduction of boosting see Mayr and Hofner, 2018). An empirical risk function is minimized and the algorithm is stopped after finitely many iterations. In each iteration, the residuals of the current fit to the outcome are re-evaluated by the gradient of the loss function. Also, the choice of the stopping iteration is an important tuning parameter for the predictive performance (Hofner et al., 2012).

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

Additionally, the design of the algorithm leads to automatic variable selection that can be particularly helpful for high-dimensional data sets. We want to take advantage of the modular structure of the algorithm and test different robust loss functions with respect to outlier corrupted data.
For composite loss functions we present a quantile-based approach, which adapts the cut-off value in each iteration to the given fit and can also be adapted to the corruptness of the data. This quantile valuation is already used for TreeBoost (Friedman, 2001) in M-regression.
We consider the Huber and Bisquare loss in this context in Section 2, and evaluate their robustness in a high-dimensional data application for biomarkers in Section 3.

2 Methods

To fit a linear regression model for p parameters and given observations $((\mathbf{x}_1, y_1), \dots (\mathbf{x}_n, y_1))$, we can estimate the regression coefficient vector $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ for p explanatory variables by minimizing the empirical risk function $\mathcal{R} := \sum_{i=1}^{n} \rho(y_i - f(\mathbf{x}_i))$ with $f(\mathbf{x}_i) := \boldsymbol{\beta}\mathbf{x}_i^{\mathrm{T}}$ and loss function ρ.
Starting with an offset value estimate $\hat{f}^{[0]}$, the further iteration m is iteratively calculated by evaluating the negative gradient $-\frac{\partial \rho}{\partial f}$ of ρ at the current residuals such that we get the negative gradient vector of the m-th iteration:

$$\mathbf{u}^{[m]} := \left(-\frac{\partial}{\partial f} \rho \left(y_i - \hat{f}^{[m-1]}(\mathbf{x}_i) \right) \right)_{i=1,\dots,n}$$

We update the current fit by $\hat{f}^{[m]} = \hat{f}^{[m-1]} + \nu \mathbf{u}^{[m]}$ with fixed learning rate $0 < \nu \leq 1$ as long as m reaches the stopping iteration m_{stop}.
The weighting of the residuals by the loss function has a major influence on the regression model. Typical mean regression has for example the L_2 loss $\rho(x) = x^2$, but this loss function is very sensitive to outliers in the data, so that more robust approaches can be helpful.
A much noted loss has been investigated by Huber (Huber, 1964) as a convex mixture between the L_2 and the L_1 loss (absolute value):

$$\rho_H(x) := \begin{cases} \frac{x^2}{2}, & |x| \leq k \\ k(|x| - \frac{k}{2}), & |x| > k \end{cases}$$

Tukey's Bisquare loss is commonly used as an example of a non-convex, smooth and robust weighting given by

$$\rho_B(x) := \begin{cases} 1 - (1 - (\frac{x}{k})^2)^3, & |x| \leq k \\ 1, & |x| > k. \end{cases}$$

For such loss functions (see Figure 1) the choice of the cut point k determines how the residuals are reweighted in each iteration. If we would choose

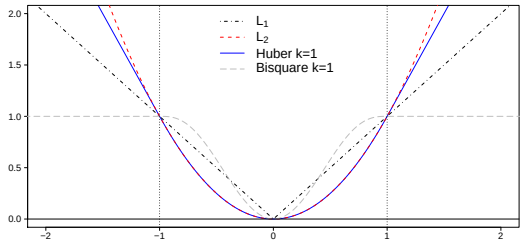


FIGURE 1. The L_1, L_2, Huber and Bisquare losses with cut-off value $k = 1$.

a constant k for all iterations, then, due to the approximation of the fit to the desired solution, from a specific iteration on, almost all residuals would be evaluated only with the inner weights of the loss function. If k is initially chosen too large, it could even happen that in no iteration the more robust, outer part of the loss function takes effect.
To ensure weighting in each iteration m for all observations in the sense of the loss function, we use a nonparametric quantile-based cut-off value

$$k^{[m]} := \text{quantile}_\tau \left(|y_i - \hat{f}^{[m-1]}(\mathbf{x}_i)|, i = 1, \ldots, n \right),$$

which applies in every step the same fraction τ of observations with the inner and $1 - \tau$ with the outer weight.
If outcome data are corrupted by a certain amount ε, we can directly incorporate this in our model, whereby we use the direct translation for Tukey's Bisquare loss ($\tau = 1 - \varepsilon$). For the Huber loss we choose τ based on an efficiency criterion depending on ε (Huber, 1981).

3 Riboflavin analysis

The open access, high dimensional data set with riboflavin production as continuous response variable contains $n = 71$ observations with $p = 4088$ gene expressions as covariates. In 100 runs, 50 observations were randomly selected as training data set and a fixed percentage ($\varepsilon = 0, 6, 10, 20\%$) of them was replaced by outliers. In more detail, four standard deviations of the response variable were randomly added or subtracted to the original response. For optimal stopping a 25-fold bootstrap approach was used that automatically determines m_{stop}. In each run we evaluated the performance of the model fits for the different loss functions on the remaining unmodified, independent test data through the mean squared error of prediction (MSEP).
As expected, in Figure 2 it can be observed that the MSEP for higher amounts of outliers increases for all considered loss functions. The L_2 loss shows an increasingly sensitive behaviour for larger amounts of outliers. In contrast, the robust losses are adaptive to stronger contamination of the

analysed data by outliers. Especially the Bisquare loss shows a favourable predictive performance for higher amounts of corrupted data.

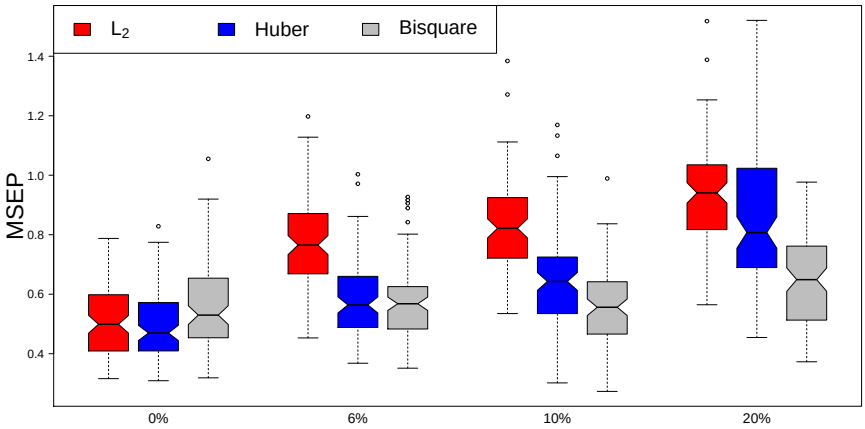


FIGURE 2. The MSEP for different amounts ($\varepsilon = 0, 6, 10, 20\%$) of outliers for the L_2 and quantile-based Huber and Bisquare losses.

4 Conclusion

With appropriate prior knowledge regarding the amount of outliers in the outcome distribution, boosting with the quantile-based Huber and Bisquare losses performs quite robust in our riboflavin analysis in comparison to the conventional L_2 loss. In addition to methods for outlier detection, the sensitivity of the choice of the quantile has to be tested in further investigations. In the optimal case a robust analysis can be guaranteed even if the researcher has no prior knowledge about the expected amount of corrupted data.

References

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, **29**(05), 1189 – 1232.

Hofner, B., Mayr, A. and Schmid, M. (2012). The Importance of Knowing When to Stop. *Methods of Information in Medicine*, **51**(02), 178 – 186.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**(1), 73 – 101.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley & Sons.

Mayr, A. and Hofner, B. (2018). Boosting for statistical modelling: A non-technical introduction. *Statistical Modelling*, **18**, 365 – 384.

Flexible amputation models for investigating missing data

Christian Staerk[1], Linda Müller[2], Andreas Mayr[1]

[1] Department of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Germany
[2] Koblenz University of Applied Sciences, Germany

E-mail for correspondence: christian.staerk@imbie.uni-bonn.de

Abstract: We propose flexible missing data ("amputation") models based on beta distributed missing probabilities, which are particularly suited for investigating different missing mechanisms. In the proposed models the marginal distribution of these probabilities can be directly specified so that deviations from the "missing completely at random" (MCAR) mechanism can be controlled. We illustrate the flexibility of the models when applied on a diabetes data set, where the results of a Bayesian multiple imputation method and a complete case analysis are compared with respect to the analysis of the full data set.

Keywords: Biomedical data; Missing data; Multiple imputation; Simulations.

1 Introduction

The appropriate analysis of data with missing values is crucial, particularly in biomedical applications where several participants may drop out early from a clinical study. Recent works have compared the performance of multiple imputation (MI) techniques with a complete case analysis (CCA) and investigated how the results are affected by the proportion of missing cases (Madley-Dowd et al., 2019) and the type of missing mechanism (Hughes et al., 2019), distinguishing between data missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Another important factor that may not have received much attention is the severity of MAR or MNAR deviations from the MCAR case, for which both MI and CCA yield unbiased results. With common "amputation" methods for generating missing data based on logistic regression it is very difficult to control the marginal distributions of the missing probabilities, which determine how large the MAR or MNAR mechanisms depart from MCAR.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

In this work we propose flexible amputation models for generating missing data, where the marginal distributions of missing probabilities can be arbitrarily specified and fully controlled by the researcher. As an important example we consider beta distributed missing probabilities. In contrast to alternative models based on logistic regression, model parameters do not need to be adapted using numerical methods in order to achieve the desired missing rate. Furthermore, the proposed models do not assume any particular distribution for the variables on which the missingness depends and are also well-suited for skewed and non-normally distributed variables. We illustrate the flexibility of the proposed models and compare them to logistic regression models used in the function `ampute` (Schouten et al., 2018) of the R-package `mice` (van Buuren and Groothuis-Oudshoorn, 2011). Based on a completely observed diabetes data set we generate missing values according to the different amputation models and evaluate the results of MI and CCA in relation to the analysis of the full data set.

2 Missing data models

We consider missing data models where the probability $\pi(Y)$ of missingness depends on a random variable Y. This includes the case of MAR, where another variable X is missing with probability $\pi(Y)$ depending on Y, and the case of MNAR, where Y itself is missing with probability $\pi(Y)$. Let F^Y denote the cumulative distribution function (cdf) of Y. Here we assume that F^Y is continuous, but the methodology can be extended to the discrete case. Let G denote the cdf of the targeted distribution for the missing probabilities $\pi(Y)$ and let G^{-1} denote its quantile function. We propose the following missing data models, which are generalizations of previously considered logistic regression models (e.g. Schouten et al., 2018):

$$\textit{right: } \pi(Y) = G^{-1}(F^Y(Y)) \qquad \textit{tail: } \pi(Y) = G^{-1}(2\,|F^Y(Y) - 0.5|)$$
$$\textit{left: } \pi(Y) = G^{-1}(1 - F^Y(Y)) \qquad \textit{mid: } \pi(Y) = G^{-1}(1 - 2\,|F^Y(Y) - 0.5|)$$

These four models differ in the way the missing probability $\pi(Y)$ depends on Y. For the model *right*, $\pi(Y)$ increases in Y, while for the model *left*, $\pi(Y)$ decreases in Y. For the model *tail*, $\pi(Y)$ first decreases in Y up to the median of the distribution of Y and then increases in Y again, implying that missing probabilities are largest in the tails of the distribution. Similarly, the model *mid* yields the largest missing probabilities around the median of the distribution of Y. Note that $F^Y(Y)$, $1 - F^Y(Y)$, $2\,|F^Y(Y) - 0.5|$ and $1 - 2\,|F^Y(Y) - 0.5|$ all follow a uniform distribution $U(0, 1)$. Thus, by the inverse transform method for each model we have $\pi(Y) \sim G$, i.e. the distribution of missing probabilities is given by G and the mean of G equals the expected number of missing cases.

The distribution G can be arbitrarily specified. Here, we specifically employ a beta distribution $\pi(Y) \sim \text{Beta}(\mu, \tau)$ with mean $\mu \in (0, 1)$ and precision

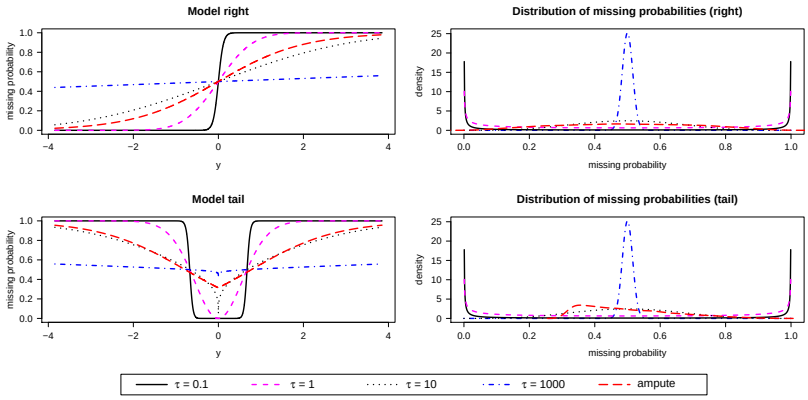


FIGURE 1. Models *right* and *tail* with $\mu = 0.5$, $\tau \in \{0.1, 1, 10, 1000\}$ and corresponding logistic regression models from `ampute` for $Y \sim N(0, 1)$.

parameter $\tau > 0$ controlling the variance via $\mathrm{Var}(\pi(Y)) = \frac{\mu(1-\mu)}{\tau+1}$. Figure 1 illustrates models *right* and *tail* with $\mu = 0.5$ and different choices of τ for normally distributed $Y \sim N(0, 1)$. The mean μ specifies the average number of missing cases, while τ controls how strong the missingness depends on the variable Y: if τ is small, the variance of $\pi(Y)$ is large and thus the missingness strongly depends on Y. The limiting case $\tau \to \infty$ corresponds to MCAR, where $\pi(Y) \equiv \mu$ does not depend on Y.
The cdf F^Y of Y is often explicitly known in simulation studies where Y is generated according to a chosen distribution. However, in real data applications the "true" cdf F^Y is typically not available and has to be estimated based on a finite sample. If $y_1, \ldots, y_n$ is the observed sample of Y, we use (a shifted version of) the empirical cdf of Y in order to estimate F^Y, i.e. $\hat{F}_n^Y(t) = \frac{1}{n}\sum_{i=1}^{n} 1\{y_i \leq t\} - \frac{1}{2n}$, for $t \in \mathbb{R}$.

3 Diabetes data example

As an illustrative example we consider the publicly available diabetes data (Efron et al., 2004) which consists of ten baseline variables $X_1, \ldots, X_{10}$ and a numeric outcome Y measuring disease progression after one year for $n = 442$ diabetes patients. In a linear regression model $Y = \beta_0 + \sum_{j=1}^{10} \beta_j X_j + \epsilon$ we focus on estimating the effect β_3 of body mass index (BMI) at baseline (X_3) on the outcome Y. The least squares estimated effect based on the full data set is given by $\hat{\beta}_3 = 5.60$ with $(4.19; 7.01)$ as 95% confidence interval. We evaluate MI and CCA in relation to the full data analysis by simulating $10,000$ missing data sets based on the model *tail* with beta distributed missing probabilities with missing rate $\mu = 0.15$ and varying precision τ and based on the corresponding logistic model from `ampute`. The Bayesian normal model of `mice` is used for MI with $m = 5$ imputations.

TABLE 1. Results of CCA and MI for diabetes data with MAR in BMI (X_3) depending on Y and MNAR in Y. Amputation is based on model *tail* with missing rate $\mu = 0.15$ and varying precision τ as well as on corresponding `ampute` model.

	$\tau = 0.1$		$\tau = 1$		$\tau = 10$		$\tau = 1000$		`ampute`	
MAR X_3	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
CCA	-1.60	1.61	-1.42	1.44	-0.56	0.67	-0.05	0.32	-0.42	0.55
MI	-0.47	0.53	-0.50	0.61	-0.18	0.40	-0.02	0.31	-0.15	0.38
MNAR Y	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
CCA	-1.60	1.61	-1.42	1.44	-0.56	0.67	-0.05	0.32	-0.42	0.55
MI	-1.60	1.61	-1.42	1.45	-0.56	0.68	-0.05	0.35	-0.42	0.57

Table 1 shows the results in terms of bias and root mean squared error (RMSE) for estimating β_3 in relation to $\hat{\beta}_3$ from the full data analysis. In case of MAR in X_3 depending on Y, absolute bias and RMSE tend to be larger for small τ. MI yields lower absolute bias than CCA in case of MAR, but interestingly there is still some bias for small τ. Possible reasons for this may be misspecifications of the analysis model for Y and of the imputation model for X_3. In case of MNAR in Y, both CCA and MI yield similar and increasingly biased results for smaller values of τ. Note that the case $\tau = 1000$ is similar to MCAR where CCA and MI are unbiased. Although results for the function `ampute` hint in the same directions, the induced missing mechanism is rather "close" to MCAR in this case, so that general effects of missingness may be underestimated. Furthermore, missing probabilities from `ampute` are relatively small in the left tale of the right-skewed distribution of Y, while the proposed models provide full control of the missing probabilities even for skewed variables.

References

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32(2)**, 407 – 499.

Hughes, R.A., Heron, J., Sterne, J.A., and Tilling, K. (2019). Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology*, **48(4)**, 1294 – 1304.

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, **110**, 63 – 73.

Schouten, R.M., Lugtig, P., and Vink, G. (2018). Generating missing values for simulation purposes: A multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, **88(15)**, 2909 – 2930.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, **45(3)**, 1 – 67. URL https://www.jstatsoft.org/v45/i03/.

Neural network classification of movement patterns in a virtual reality experiment

Frederike Vogel[1], Nils Vahle[2], Jan Gertheiss[1], Martin Tomasik[2]

[1] Department of Economics and Social Sciences, Helmut Schmidt University, Hamburg, Germany
[2] Department of Psychology and Psychotherapy, University of Witten-Herdecke, Germany

E-mail for correspondence: vogelf@hsu-hh.de

Abstract: There is solid empirical evidence that the stereotypes people hold about age and aging are applied in a self-reflexive way once they get old themselves. In this study, we used virtual reality technology to trigger this effect in young people and collected their body movement patterns. The present paper shows computational evidence for changes in head and arm movements as a function of the experimental condition.

Keywords: Age Stereotypes; Supervised Learning; Virtual Reality

1 Introduction

Self-reflexive age stereotypes can cause heavy effects on the health of an aging person, see e.g., Stewart et al. (2012). Since activating said stereotypes in experimental settings has shown to be difficult, see e.g., Rivers and Sherman (2018), we incorporated virtual reality in our own study to create a strongly immersive environment, see Vahle and Tomasik (2020) for details. Corresponding data were gathered in the following way: $n = 72$ students (age 20-35) were randomly assigned with either a younger (i.e., age-congruent control condition) or an older (i.e., age-incongruent experimental condition) virtual avatar asked to perform simple movement tasks such as raising their arms or inspecting their hands. During these performances that lasted a total time of 8.17 minutes, three-dimensional coordinates and rotations of three points in space (head and two hands) were tracked (referred to as channels from now on) with a resolution of 10Hz

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

(i.e., we have a total number of time points of 4970). Figure 1 shows an exemplary course of the experiment for head movements.

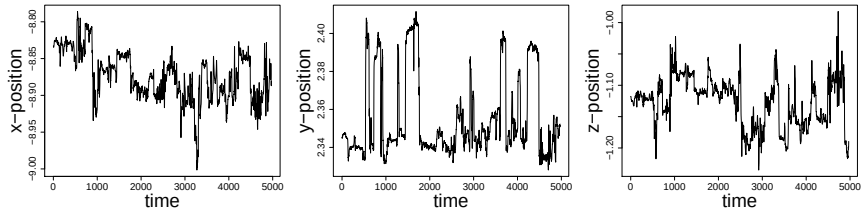


FIGURE 1. Example of head movement patterns in three axes.

To reduce dimensionality and identify the main directions of variability, we used principal component analysis (PCA). An example finding is given in Figure 2. As we can see, densities of first PC scores of the head's y-rotation

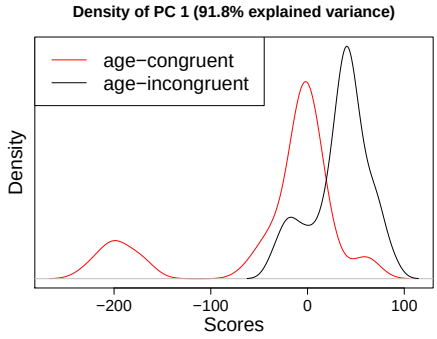


FIGURE 2. Densities of first PC scores for y-rotation of the head.

differ in modality. Such preliminary findings indicate discrepancies between different groups of the experiment. Our ambition now is to determine techniques to separate these groups distinctly and give predictions about group affiliation. We use supervised learning with neural networks as they have proven to deliver promising results for human activity recognition tasks, see e.g., Herath et al. (2017).

2 Methodology and Model Description

2.1 Feedforward Neural Network (FFN)

A feedforward neural network $F^\theta : \mathbb{R}^d \to (0,1)^2$ with two hidden layers is defined as a composition of affine functions and non-linear activation functions, i.e.

$$F^\theta = \psi \circ a_3^\theta \circ \sigma_{q_2} \circ a_2^\theta \circ \sigma_{q_1} \circ a_1^\theta$$

where d denotes the input dimension and q_1, q_2 the number of neurons in the hidden layers which are characterized by the affine functions

$$a_1^\theta : \mathbb{R}^d \to \mathbb{R}^{q_1}, \qquad a_2^\theta : \mathbb{R}^{q_1} \to \mathbb{R}^{q_2}, \qquad a_3^\theta : \mathbb{R}^{q_2} \to \mathbb{R}^2$$

defined by

$$a_i^\theta(x) := A_i x + b_i, \qquad i = 1, 2, 3,$$

whose components $A_1 \in \mathbb{R}^{q_1 \times d}$, $A_2 \in \mathbb{R}^{q_1 \times q_2}$, $A_3 \in \mathbb{R}^{q_2 \times 2}$ (the weight matrices), $b_1 \in \mathbb{R}^{q_1}$, $b_2 \in \mathbb{R}^{q_2}$, $b_3 \in \mathbb{R}^2$ (the bias vectors) compose the parameter $\theta \in \mathbb{R}^{2+(1+d+q_2)q_1+3q_2}$. For the activation functions, we define $\sigma_j : \mathbb{R}^j \to \mathbb{R}^j, j \in \{q_1, q_2\}$ and $\psi : \mathbb{R}^2 \to \{y \in \mathbb{R}^2 | y_1, y_2 \geq 0, y_1 + y_2 = 1\}$ as ReLU and softmax function, respectively, i.e. $\sigma_j(x_1, \ldots, x_j) = (x_1^+, \ldots, x_j^+)$ and $\psi(x)_j = e^{x_j}/(e^{x_1} + e^{x_2})$, $j = 1, 2$. In our application, we use FFN for an input consisting of basic features (mean, standard deviation, minimum and maximum of three-axes-positions and rotations, resulting in a total of 24 input variables for each body part) drawn from the whole courses of the experiment. All weights and biases are initialized via a Xavier uniform initializer.

2.2 Convolutional Neural Network (CNN)

In our setting, convolutional layers perform one-dimensional discrete convolution for each channel $c \in C$ separately leading to an output of

$$(K * S)(i) = \sum_{c \in C} \sum_{1 \leq u \leq F} K(u, c) S(i + u, c), \qquad 1 \leq i \leq m - F + 1,$$

for a kernel $K \in \mathbb{R}^{F \times |C|}$ with filter length $F \in \mathbb{N}$, and a time segment $S \in \mathbb{R}^{m \times |C|}$. A bias vector may be added to the result as well as an activation function may be applied afterward. This procedure can be repeated for several filters resulting in an output $O \in \mathbb{R}^{(m-F+1) \times n_F}$ where n_F denotes the number of filters. In a next step, we use max-pooling which slides a window of size F_M over the data and summarizes them by taking the maximum eventuating in an output of the form

$$M(i, j) = \max_{k = F_M \cdot (i-1), \ldots, F_M \cdot i - 1} |O(k, j)|$$

for $1 \leq i \leq (m - F + 1)/F_M$, $1 \leq j \leq n_F$. In our case, we use CNN to classify not the whole course of the experiment itself, but only parts of it. Therefore, we insert time segments consisting of $m = 70$ time points respecting all six channels (i.e., $C = \{1, \ldots, 6\}$). We build our network by inducing one convolutional layer with $n_{F_1} = 70$ kernels of filter length $F_1 = 35$ and constant bias. Max-pooling with a window size of $F_M = 3$ and another convolutional layer with $n_{F_2} = 70$ kernels of filter length $F_2 = 2$ are instantiated subsequently. All kernels are initialized by a truncated normal distribution. The output is then flattened and inserted into a feedforward layer with 100 neurons and tanh-activation. Eventually, the softmax function is applied.

3 Results

For classification (age-congruent control group vs. age- incongruent experimental group), we construct neural networks as described above using `Tensorflow 2.0`. For both FFN and CNN, sparse categorical crossentropy is minimized using an Adam-Optimizer. For comparison, a logit model (see `classif.glm` method in the `R` package `fda.usc`) is also applied to basic feature input (abbr. logitBF). Furthermore, we applied the above CNN (abbr. CNN1) again as well as the logit model (abbr. logit1) on time segments allowing just one channel (y-rotation) for 70 time points per segment. All runs have been executed for the head, left and right hand separately. Training of models was conducted 100 times on randomly chosen 80 % of the input data before evaluating the classification accuracy on the remaining data. Results can be abstracted from Table 1.

TABLE 1. Classification accuracy for different models in percent.

Part of Body	FFN	CNN6	CNN1	logitBF	logit1
Head	92	99.7	81.32	61.2	72.96
Right Hand	75	98	55.5	45.73	53.78
Left Hand	84	97	49.87	45.8	53.29

As we see, classification based on basic features works significantly better using FFN compared to using the logit model. When it comes to prediction based on direct inserting of time passages, our CNN provides great accuracy given that all channels are considered. Both the CNN and the logit model, however, have problems classifying the data correctly if just one channel is used, especially when inserting hand movement patterns.

References

Herath, S., Harandi, M. and Porikli, F. (2017). Going deeper into action recognition: a survey. *Image and Vision Computing*, **60**, 4 – 21.

Rivers, A.M. and Sherman, J. (2018). Experimental design and the reliability of priming effects: reconsidering the "train wreck". *PsyArxiv*.

Stewart, T.L., Chipperfield, J.G., Perry, R.P. and Weiner, B. (2012). Attributing illness to "old age": consequences of a self-directed stereotype for health and mortality. *Psychology & Health*, **27**, 881 – 897.

Vahle, N. and Tomasik, M. J. (2020). *Changes in memory and physical performance during an aging experience in virtual reality.* Manuscript in preparation.

The (in)stability of Bayesian model selection criteria in disease mapping

Maren Vranckx[1], Thomas Neyens[1], Christel Faes[1]

[1] Hasselt University, Data Science Institute (DSI), The Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt, Belgium

E-mail for correspondence: maren.vranckx@uhasselt.be

Abstract: Several model comparison techniques exist to select a best model from a set of candidate models. This study explores the performance of model comparison statistics among several Bayesian software packages that are often used for spatially discrete disease modelling: the deviance information criterion (DIC), the Watanabe-Akaike information criterion (WAIC) and the log marginal predictive likelihood (LMPL). We focus on the software packages CARBayes, OpenBUGS, NIMBLE and Stan, in which we fit Poisson models to disease incidence outcomes with intrinsic conditional autoregressive, convolution conditional autoregressive and log-normal error terms. From data studies, we learn important disparities in model selection. Based on these conclusions, we provide recommendations on the optimal use of model comparison statistics for all kind of applications.

Keywords: Disease mapping; Software packages; DIC; WAIC; LMPL.

1 Introduction

Most comparisons between Bayesian software tools focus on parameter estimation and prediction (Vranckx et al. 2019). Model selection, which aims at appointing a representative model from a set of candidate models, given the data, is usually not considered. Bayesian model selection itself is a widely debated topic. Many approaches have been proposed over time, such as the deviance information criterion (DIC), the Watanabe-Akaike information criterion (WAIC) and the log marginal predictive likelihood (LMPL).
The focus of this manuscript is to explore the available model selection tools in different Bayesian software packages for spatial disease mapping, together with their practical advantages and disadvantages. It is not our intention to compare the model selection criteria, but to look at the stability of these criteria.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 General methodology

2.1 Models

Disease mapping is used to estimate an unknown relative disease risk R_k for an area k of a spatially discrete study region. For each area k, we have the number of (newly diagnosed) disease or mortality cases (Y_k) and the expected number of cases (E_k). The general model formulation is

$$\begin{aligned} Y_k|E_k, R_k &\sim Poisson(E_k R_k), \quad k \in \{1, ..., n\}, \\ ln(R_k) &= \mu + \mathbf{x_k^T}\beta + \phi_k \ , \end{aligned} \tag{1}$$

where β denote the regression parameters and μ the intercept term.
Depending on the prior distribution of the random effects ϕ, different models can be constructed. When no spatially structured association is assumed, a Poisson-lognormal model can be used. On the other hand, spatial models induce spatial association in the model. Popular spatial models are the intrinsic and convolution models (both Besag et al. 1991).

2.2 Model comparison techniques

Several tools are available to select the best model from a set of models. A commonly used model comparison statistic is the DIC (Spiegelhalter et al. 2002). The DIC balances between goodness of fit and model complexity. As a goodness-of-fit measure, it uses the posterior mean of the deviance ($E_{\theta|\mathbf{y}}(D)$). The deviance D is defined as $-2 \log p(\mathbf{y} \mid \theta)$, where $\log p(\mathbf{y} \mid \theta)$ is the log-likelihood of the data $\mathbf{y}$ given the parameter θ. To compensate for the complexity of the model, the $\mathrm{p_{DIC}}$ is calculated as

$$\mathrm{p_{DIC}} = E_{\theta|\mathbf{y}}(D) - D(E_{\theta|\mathbf{y}}\theta)\,. \tag{2}$$

The DIC is then defined as

$$\mathrm{DIC} = E_{\theta|\mathbf{y}}(D) + \ \mathrm{p_{DIC}}\,. \tag{3}$$

More recently, Watanabe (2010) introduced the WAIC. WAIC uses a measure for accurate predictions, which is also compensated with a so-called effective number of parameters due to the double use of the data. It is defined as

$$\mathrm{WAIC} = -2\ \mathrm{lppd} + 2\ \mathrm{p_{WAIC}} \tag{4}$$

where $\mathrm{lppd} = \sum_{k=1}^{n} \log p(y_k \mid \mathbf{y})$ is the log of the joint posterior predictive distribution for all units $k = 1, ..., n$ and $\mathrm{p_{WAIC}} = \sum_{k=1}^{n} \mathrm{var}_{\theta|\mathbf{y}} [\log p(y_k \mid \theta)]$ is the penalization term.
Other model selection techniques are based on cross-validation, which aim to investigate prediction accuracy. Leave-one-out cross validation takes a

single observation from the data for validating and uses the other data points for fitting. Based on this idea, the conditional predictive ordinate (CPO_k) for one unit k can be defined as

$$\text{CPO}_k = p(y_k \mid \mathbf{y_{-k}}) \tag{5}$$

where $\mathbf{y_{-k}}$ represents the data without observation y_k. As overall measure for model selection, the LMPL (Geisser and Eddy 1979) can be used, defined as

$$\text{LMPL} = \sum_{k=1}^{n} \log(\text{cpo}_k)\,. \tag{6}$$

The model with the lowest DIC, the lowest WAIC and the highest LMPL is preferred. However, there is no general threshold value to give a positive support to a model.

3 Data analysis

3.1 Data description

The data analysis is based on the asthma dataset of Georgia (USA) publicly available from the OASIS online system of the Georgia Division of Public Health (https://oasis.state.ga.us/). It represents the counts of newly diagnosed asthma cases in 2005.

3.2 Results

Table 1 shows that the model comparison estimates calculated by the software packages differ considerably among the different software packages. CARBayes prefers the intrinsic model, while R2OpenBUGS the convolution model.
Using a unifying calculation method, Figure 1 shows the trace plot of the DIC for the intrinsic model and the packages CARBayes, R2OpenBUGS. For CARBayes, model comparison statistics often do not converge at all due to a large difference in estimates between different MCMC chains. Therefore, care is needed in model choice based on these results.

4 Conclusion

Looking at the estimates resulting from the different software packages, we noticed that using different packages with their own specific calculation method, can lead to different model preference. This difference is partially due to different calculation method. Therefore, users of different software packages should be aware that model comparison statistics are not comparable over the different packages. However, when using the same calculation method, differences can still occur due to among other things software specific posterior samples. Moreover, convergence of the parameters does not necessarily mean convergence of the model comparison statistics.

TABLE 1. A summary of the model comparison statistics resulting from CARBayes and R20penBUGS after 20 000 iterations and a burn-in of 10 000.

	CARBayes		R2OpenBUGS	
	DIC	p_{DIC}	DIC	p_{DIC}
Intrinsic	1234.749	153.048	1006	-10.08
Convolution	1386.896	206.600	931	-83.94
Poisson-lognormal	1280.178	169.513	1156	135.6

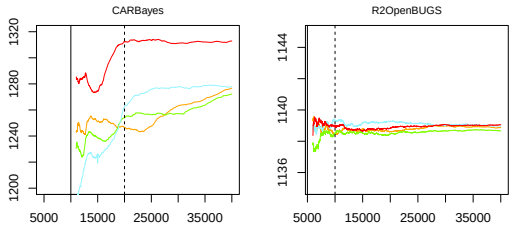


FIGURE 1. Trace plots of the DIC for the intrinsic model and for CARBayes, R2OpenBUGS. On the x-axis, the number of iterations are indicated. The different colors indicates different MCMC chains. The vertical lines indicates the needed burn-in for convergence of the parameters and sufficient iterations to calculate the parameter estimates.

References

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.*, **43**, 1 – 20

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *J Am Stat Assoc*, **74**, 153 – 160.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Stat. Soc., Ser. B*, **64**, 583 – 639.

Vranckx, M., Neyens, T. and Faes, C. (2019). Comparison of different software implementations for spatial disease mapping. *Spat. Spatio-temporal Epidemiol.*, **31**

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *J. Mach. Learn. Res.*, **11**, 3571 – 3594.

Visualization techniques for semiparametric APC analysis – Using Generalized Additive Models to examine touristic travel distances

Maximilian Weigert[1], Alexander Bauer[1], Johanna Gernert[2], Marion Karl[3], Helmut Küchenhoff[1], Jürgen Schmude[2]

[1] Statistical Consulting Unit StaBLab, LMU Munich, Germany
[2] Department of Geography, LMU Munich, Germany
[3] UQ Business School, University of Queensland, St Lucia, Australia

E-mail for correspondence: maximilian.weigert@stat.uni-muenchen.de

Abstract: Examination of age, period and cohort effects is a crucial aspect in many long-term studies. In this work, we extend a holistic APC analysis framework by introducing innovative visualization techniques facilitating the intuitive interpretation of complex temporal structures. Our concepts are motivated by the representation of age, period and cohort in Lexis diagrams. We introduce ridgeline matrices, a two-dimensional extension of ridgeline plots, to commonly visualize distributions for age groups, periods and cohorts. The established APC concept of generalized additive models is used to circumvent the identification problem by fitting a bivariate tensorproduct spline between age and period. We outline our concepts by analyzing altering travel distances of German tourists.

Keywords: APC analysis; generalized additive models; graphical visualization; ridgeline plots; tourism research

1 Introduction

Analyzing temporal developments in specific population groups is a common goal in social sciences. Based on long-term panel or repeated cross-sectional data time-related effects are separated into age, period and cohort effects where a cohort usually represents individuals with common birth years. The estimation of such APC (age-period-cohort) models is faced with an identification problem as each component can be expressed as a linear combination of the others, e.g. age = period − cohort. Over the last decades different approaches for dealing with this identification problem have been

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

developed including constrained generalized linear, Bayesian hierarchical and spline-based models (see Yang and Land, 2013, for an overview of APC methodology). Our focus is on semiparametric spline-based models where an interaction surface between age and period represents all three types of effects. We extend this holistic framework by (a) introducing the ridgeline matrix as a descriptive visualization tool and by (b) refining graphical representations of estimated effect structures.

2 Data

Our study is based on an annual representative cross-sectional survey among yearly approximately 7 500 people in Germany (FUR, 2019). Survey data are available from 1971 to 2018 and comprise travel behavior in the (short-term) past and the main dimensions of travel decision making along with socio-demographic information about travelers.

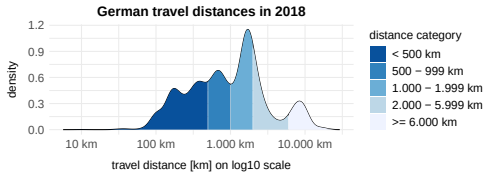


FIGURE 1. Travel distance curve of German tourists in 2018 on log10 scale

We analyze travel distances as one of the key aspects of destination choice (Yang et al., 2018). Age and cohort of a traveler and the travel period are used as proxies for internal and external effects, respectively. Figure 1 examplarily shows a *travel distance curve*, i.e. the distribution of distances for travels in 2018 on a logarithmic scale. In the modeling process, distance categories are examined rather than raw distances since the latter are to some extent arbitrary as exact travel destinations are mostly unknown.

3 Methods

The key idea of our framework relies on Lexis diagrams, i.e. two-dimensional diagrams common in medical APC applications where age groups, periods and cohorts are usually depicted along the x-axis, the y-axis and diagonals, respectively (see e.g. Carstensen, 2007). We introduce *ridgeline matrices* as a novel technique for descriptively visualizing APC structures. Ridgeline matrices are a two-dimensional extension of ridgeline plots (Wilke, 2018), an established tool to display densities against a secondary variable, and comprise a layout with age groups along the horizontal and periods along the vertical axis. Accordingly, diagonals represent specific cohorts.
Our modeling approach is based on generalized additive regression models with penalized splines (see Wood, 2017) and builds upon the work of

Clements et al. (2005) who considered a bivariate tensorproduct spline between age and period for APC modeling. The resulting model addresses the identification problem by implicitly regarding the cohort effect as a statistical interaction between age and period, represented by the diagonal of the estimated tensorproduct surface. For exponential family responses with expected value μ and link function $g(\cdot)$ we use the model structure

$$g(\mu) = \beta_0 + f(\texttt{age}_i, \texttt{period}_i), \qquad i = 1, \ldots, n$$

where β_0 denotes the intercept and $f(\cdot, \cdot)$ is a tensorproduct of the two marginal spline bases. Graphical visualization of effect structures is based on heatmaps of the estimated surface and the extraction of individual effects for age, period and cohort from the fitted model. Additional to the estimation of temporal structures, the modeling framework allows for an integration of additional covariates on individual or aggregated level.

4 Results

Figure 2 exemplarily shows a ridgeline matrix visualizing travel distance curves for travelers aged 20, 30 and 40 over five decades. It illustrates that younger age groups, newer periods and later-born cohorts are associated with longer travel distances.

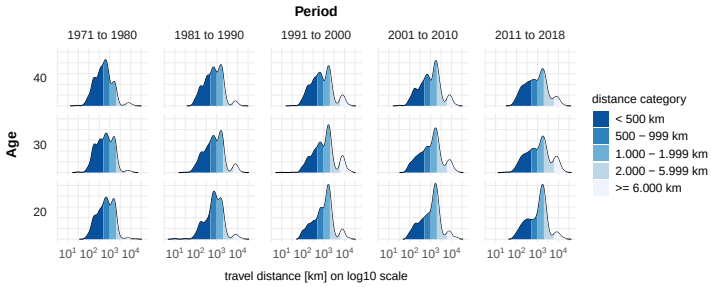


FIGURE 2. Ridgeline matrix depicting the development of travel distances (displayed on log10 scale) for travelers aged 20, 30 and 40. Cohorts born between 1951 to 1960 and 1971 to 1980 are highlighted blue and green, respectively.

Each threshold between distance categories is modeled individually leading to four additive logistic regression models. Figure 3 displays the estimated mean APC effects for all thresholds. The overall age effects show a distinct bimodal pattern with a maximum among 20 to 35 year-olds and a lower peak within the age group of 45 to 55. Period and cohort effects reveal an increasing chance for longer distances across all thresholds. The elaboration of further visualizations of the estimated surface and model uncertainties is currently underway. Model performance was evaluated based on the AUC criterion. All models reached values between 0.62 and 0.68 on test data.

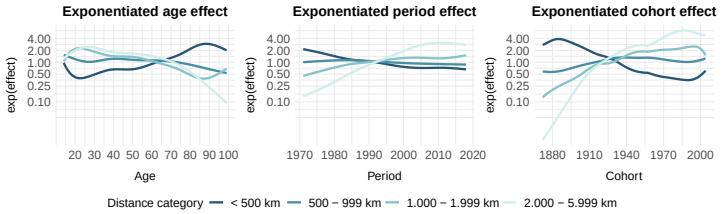


FIGURE 3. Exponentiated mean age (left), period (middle) and cohort (right) effects for all modeled thresholds on log10 scale.

5 Conclusion

The presented APC framework comprises an intuitive spline-based modeling approach and innovative multidimensional visualizations. In the end, it will offer guidance on visualization techniques for both general description and evaluation of model estimates and uncertainties. We showcased our concepts with an application in tourism research. However, it can easily be adapted to similar research settings in other fields.

References

Carstensen B. (2007). Age-period-cohort models for the Lexis diagram. *Statistics in medicine*, **26**, 3018 – 3045.

Clements M.S., Armstrong B.K. and Moolgavkar F.H. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics*, **6(4)**, 576 – 589.

FUR (2019). Reiseanalyse 2019 [Travel demand analysis Germany]. For-schungsgemeinschaft Urlaub und Reisen e.V. Last retreived: 30/03/2020. `https://reiseanalyse.de/how-is-the-survey-conducted/`.

Wilke C.O. (2018). *ggridges: Ridgeline Plots in 'ggplot2'*. R package version 0.5.2. `https://cran.r-project.org/package=ggridges`.

Wood S.N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd edition. Chapman and Hall/CRC.

Yang Y. and Land K.C. (2013). *Age-period-cohort analysis: New models, methods, and empirical applications*, Chapman and Hall/CRC.

Yang Y., Liu H., Li X.R. and Harrill R. (2018). A shrinking world for tourists? Examining the changing role of distance factors in understanding destination choices *Journal of Business Research*, **92**, 350 – 359.

Bias induced during the estimation of quality-adjusted life-years

Alexandra Welsh[1], Deborah A. Costain[1], Andrew C. Titman[1]

[1] Dept. of Mathematics and Statistics, Lancaster University, UK

E-mail for correspondence: a.k.welsh@lancaster.ac.uk

Abstract: Quality-adjusted life-years (QALYs) are a summary measure used to evaluate the effectiveness of medical treatments in terms of both quality and length of life. One method used to estimate QALYs is the area under the time-utility curve (AUC). However, this approach may induce bias, due to its inability to capture the dependency between the quality of life measures and the survival time. A simulation study is conducted to assess the bias induced when estimating QALYs using the AUC method, using data including censored individuals and missing responses.

Keywords: Quality-adjusted life-years; Joint longitudinal-survival models.

1 Introduction

In order to inform healthcare resource decisions, the financial cost and the health outcomes associated with any given treatment must be evaluated. The health outcomes used in economic analyses should incorporate the impact of the treatment on both the length of life and health-related quality of life (HQoL). The quality-adjusted life-year (QALY) is one such summary measure; one QALY is equivalent to one year of life in perfect health.
Instruments such as the EQ-5D questionnaire (EuroQol Group, 1990) can be used to obtain utility values for HQoL states. Utility values indicate the desirability of the state and are usually between 0 = Death and 1 = Perfect health. QALYs are calculated as the length of life weighted by the relevant longitudinal utility scores.
An area under the curve (AUC) method can be used to estimate QALYs, where linear interpolation of the longitudinal HQoL data points is used to establish the health utility over time, and the value 0 is taken after the time of death. However, summary measures such as the AUC may

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

result in biased estimates, especially in the presence of missing data (Bell et al., 2014). The AUC method does not take into account the dependence between the HQoL observations and the survival process, which may be one cause of bias.

We aim to determine under which circumstances the AUC approach for estimating QALYs can be biased. Building upon the work conducted by Bell et al. (2014), we also consider the effect of the method used to estimate HQoL at the time of death on the bias of the QALY estimate. A simulation study is conducted in which the longitudinal HQoL data trends, the presence of censoring, and the missing data patterns are varied.

2 Simulation Study

2.1 Methodology

Firstly, dependent longitudinal HQoL and survival data are generated for each subject. For each set of parameters chosen, 1000 replicates are drawn. The data is generated for $n = 100$ total subjects per iteration, with longitudinal response times denoted by t_{ij}, where $i = 1, \ldots, n$ and $j = 0, 1, \ldots, 10$ are the subject and time indices, respectively.

The longitudinal response for subject i at time j is generated such that

$$Y_{ij} \;=\; (\beta_0 + \nu_{0i}) + (\beta_1 + \nu_{1i})t + \epsilon_{ij}\,, \tag{1}$$

where $\boldsymbol{\nu}$ are subject-specific random effects, and $\epsilon_{ij} \sim \mathrm{N}(0, 0.01^2)$ are independent error terms. The intercept, $\beta_0 = 0.8$, remains constant for the longitudinal data patterns studied. The fixed effect coefficient of time, β_1, is selected from the set $\{-0.05, 0\}$. The random effects are taken to be either random intercept (RI), with $\nu_{0i} \sim \mathrm{N}(0, \sigma_1^2)$, or random intercept and random slope (RIRS), with $\boldsymbol{\nu_i} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$. The standard deviations of the random intercept and random slope are given by $\sigma_1 = 0.05$ and $\sigma_2 = 0.01$, respectively, with correlation parameter defined as $\rho = 0.2$. In order to reflect the structure of HQoL data, longitudinal data points are truncated at a maximum of 1.

The hazard function for subject i is given by

$$h_i(t\,;\, \boldsymbol{x_i},\, \boldsymbol{\nu}_i) \;=\; h_0(t) \exp(\gamma_1(\beta_0 + \nu_{0i}) + \gamma_2(\beta_1 + \nu_{1i})t)\,, \tag{2}$$

where $\boldsymbol{\gamma}$ determines the degrees of association between the longitudinal and survival processes. This model is based upon the general methodology introduced by Wulfsohn and Tsiatis (1997). The baseline hazard function, $h_0(t)$, is taken to follow that of a Weibull$(1.2, 12)$ distribution. The association parameters are equal, with $\gamma_1 = \gamma_2 = 0.2$. Survival times are truncated at $t = 10$.

Censoring of the survival times is also considered. For scenarios including censoring, 50% of subjects have a censored survival time, C, with $C \sim$ Uniform$(0, 10)$.

The impact of missing HQoL data is also considered. Missingness of a given response is generated using a Bernoulli(0.2) distribution, leading to a missing completely at random (MCAR) response pattern. Imputation through last observation carried forward (LOCF), and no imputation are considered to handle missing responses.

We estimate the QALYs gained through use of the methods developed by Glasziou et al. (1998). The mean QALY restricted to time L is defined as

$$\text{QALY}_L = \int_0^L P(t)Q(t)\,\mathrm{d}t\,,$$

where $P(t)$ is the proportion of subjects alive at time t, and $Q(t)$ is the mean HQoL of those subjects at time t. The function $P(t)$ is estimated using the Kaplan-Meier estimator; $Q(t)$ is estimated through interpolation of the HQoL for each individual i, which are then combined to yield a mean function for the entire group.

A longitudinal response is thus required at all observation times, and distinct censoring and survival times for all individuals, in order to estimate $Q(t)$. For those individuals who experienced the event, three methods are considered to estimate the HQoL at their recorded survival time: LOCF, extrapolation based on a linear regression model, and linear interpolation between the last observation and 0, as this is the value taken after death. Each censored subject requires a response at time $t = 10$; this response was estimated for the individual using either LOCF, or extrapolation based on a linear mixed effects model. From this point, responses could be linearly interpolated for each subject at all times necessary.

2.2 Study Results

In order to deduce the best form of the QALY estimator for each model, the study was completed using four scenarios of increasing complexity. The scenarios are denoted as follows: R used complete data with no censoring, RC used complete data including censored individuals, RM used MCAR data with no censoring, and RCM used MCAR data including censoring.

The bias and mean squared error (MSE) for the best RI and RIRS models in each scenario are shown in Table 1. Both LOCF and extrapolation are appropriate choices to estimate responses at death times in scenario R; interpolation to 0 is significantly inferior in all scenarios. In scenario RC, models can use any combination of LOCF and extrapolation to estimate responses at death times and $t = 10$ for censored subjects without a significant impact on the level of bias induced. When MCAR data are considered, in scenarios RM and RCM, no imputation produces significantly superior results to imputation of the missing data through LOCF. Conclusions are the same for both RI and RIRS methods throughout, although when MCAR data is considered, RI models are likely to underestimate, and RIRS models likely to overestimate, QALYs, respectively.

	Method	Model			
		RI		RIRS	
		Bias (SD)	**MSE**	**Bias (SD)**	**MSE**
R	LOCF	-0.047 (0.253)	0.066	-0.008 (0.194)	0.038
	Ext.	-0.049 (0.253)	0.066	-0.014 (0.194)	0.038
RC	LOCF/LOCF	-0.041 (0.282)	0.081	-0.021 (0.206)	0.043
	LOCF/Ext.	-0.041 (0.282)	0.081	-0.025 (0.206)	0.043
	Ext./LOCF	-0.043 (0.282)	0.081	-0.027 (0.207)	0.043
	Ext./Ext.	-0.043 (0.282)	0.081	-0.031 (0.207)	0.044
RM	LOCF	-0.047 (0.253)	0.066	0.032 (0.195)	0.039
	Ext.	-0.050 (0.253)	0.066	0.024 (0.196)	0.039
RCM	LOCF/LOCF	-0.041 (0.282)	0.081	0.023 (0.208)	0.044
	LOCF/Ext.	-0.041 (0.282)	0.081	0.006 (0.208)	0.043
	Ext./LOCF	-0.044 (0.283)	0.082	0.015 (0.209)	0.044
	Ext./Ext.	-0.043 (0.283)	0.082	0.001 (0.209)	0.044

TABLE 1. The model error for each of the best models of the simulation study. The methods, LOCF and extrapolation (ext.), are described in Subsection 2.1, and are given in order of death, censoring (if applicable).

3 Discussion and Future Work

Several extensions to the current simulation study have been considered. Inclusion of simulated covariates in the longitudinal and the survival models is one such proposal. More complex HQoL patterns, such as those which change gradient over time, would also be an appropriate development for the study. Another consideration is to include missing at random (MAR) or missing not at random (MNAR) missing data patterns for the longitudinal responses. Finally, in order to better extrapolate responses at $t = 10$ for censored subjects, a possibility would be to use the best linear unbiased predictor (BLUP) to make use of the subject's own random effects.

One approach proposed as an alternative to AUC for the estimation of QALYs is the use of joint longitudinal-survival modeling (Rizopoulos, 2012). By fitting a joint model to the longitudinal and survival data, the QALYs can be estimated by integrating the fitted model over the survival times. Li et al. (2013) have proposed a joint model, which makes use of a 'reverse' time scale, applying it to HQoL data in order to estimate QALYs. In our future work, we aim to investigate the potential benefits of using joint modeling approaches, rather than the AUC method, to estimate QALYs from dependent longitudinal HQoL and survival data.

Acknowledgments: Special thanks to the Economic and Social Research

Council (ESRC) for their financial support of this project [ES/P000665/1].

References

Bell, M. L., King, M. T., and Fairclough, D. L. (2014). Bias in area under the curve for longitudinal clinical trials with missing patient reported outcome data: Summary measures versus summary statistics. *SAGE Open*, **4**(2), 2158244014534858.

EuroQol Group (1990). EuroQol - a new facility for the measurement of health-related quality of life. *Health Policy*, **16**, 199 – 208.

Glasziou, P. P., Cole, B. F., Gelber, R. D., Hilden, J., and Simes, R. J. (1998). Quality adjusted survival analysis with repeated quality of life measures *Statistics in Medicine*, **17**, 1215 – 1229.

Li, Z., Tosteson, T.D., Bakitas, M.A. (2013). Joint modeling quality of life and survival using a terminal decline model in palliative care studies. *Statistics in Medicine*, **32**, 1394 – 1406.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data with Applications in R*. CRC Press.

Wulfsohn, M. S., and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**(1), 330 – 339.

Hierarchical species distribution modelling across high dimensional nested spatial scales

Craig Wilkie[1], Jafet Belmont[1], Claire Miller[1], Marian Scott[1], Tom August[2], Philip Taylor[3]

[1] University of Glasgow, Glasgow, UK
[2] UK Centre for Ecology and Hydrology, Wallingford, UK
[3] UK Centre for Ecology and Hydrology, Edinburgh, UK

E-mail for correspondence: craig.wilkie@glasgow.ac.uk

Abstract: We propose a two-stage modelling approach to evaluate how a large suite of environmental metrics available over nested spatial scales shape species distributions. We focus on dragonfly communities, where the data consist of partially observed presence records, making identifying the ecological processes driving the true species distribution/occupancy patterns difficult.

Keywords: Detectability; Dragonflies; Occupancy; High Dimensionality

1 Introduction

Understanding how species distributions are affected by environmental changes is of major interest in many ecological studies. However, describing such processes is no easy task due to the sources of uncertainty that occur at different spatial and temporal scales and that are induced by imperfect detectability.

We propose a two-stage statistical modelling framework for analysing how environmental metrics describing freshwater connectivity interact with land-use change to affect species distributions, while accounting for imperfect detectability of the species. Specifically, we look at UK dragonfly species richness, since their species presence records are only partially observed due to imperfect detection.

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

2 Data sets

Data were provided by Hydroscape (web: hydroscapeblog.wordpress.com), a project investigating how anthropogenic stressors and connectivity interact to influence biodiversity and ecosystem function in UK freshwaters. Dragonfly occupancy records (for over 4000 × 1km grid cells, matched to lakes) from 2000 to 2016 were taken for 41 non-invasive species from the National Biodiversity Network, Biological Records Centre and British Dragonfly Society repositories. Species-specific covariates that may affect species detection probability were taken from Powney et al. (2014). Anthropogenic stressors (% agricultural and urban land use) and connectivity metrics (e.g. perimeter, number of lakes, river length) were calculated on 7 spatial scales surrounding each lake to capture the impact of different types of connectivity on freshwater ecosystems (Taylor et al., in prep.).

3 Statistical Methods

Species richness (the total number of species occupying a grid cell) can be underestimated when the probability of detecting the different species is less than 1. The species occupancy is potentially affected by many environmental variables over nested spatial scales. We take a 2-stage approach, estimating detectability in stage 1 and identifying and modelling the effects of the covariates on the adjusted species richness in stage 2.

3.1 Stage 1: estimating occupancy, accounting for detectability

First, we analysed the observed occupancy of dragonfly species in each grid cell by fitting a species-specific multispecies occupancy model (eqn. 1) using the species-specific covariates from equation 2 (Kéry and Royle, 2008).

$$\begin{aligned} z_{ij} &\sim \text{Bernoulli}(\psi_i) && \text{State process} \\ \sum_{K_j} y_{ij\cdot} | z_{ij} &\sim \text{Binomial}(K_j, p_i z_{ij}) && \text{Aggregated observation process} \\ \text{logit}(\psi_i) \sim \text{N}(\mu_{\psi_i}, \sigma^2_{\psi_i}); \text{logit}(p_i) &\sim \text{N}(\mu_{p_i}, \sigma^2_{p_i}) && \text{Species heterogeneity model} \end{aligned} \tag{1}$$

$$\mu_{p_i} = \alpha_0 + \sum_{m=1}^{M} \alpha_m (m\text{th species-specific parameter})_i \tag{2}$$

where $y_{ij\cdot}$ is the number of times species i was detected in grid cell j across K visits, p_i is the detection probability for the ith species, z_{ij} is the latent variable for true species occupancy and ψ_i is the occupancy probability. Grid cell-level species richness is computed as a derived quantity of the predicted occupancy, as $S_j = \sum_i z_{ij}$. Noninformative priors were specified to run the Gibbs sampler in R through JAGS.

3.2 Stage 2: understanding the effects of the covariates on species richness

Second, we evaluated the effect of grid cell-level covariates on species richness, using two sub-steps:
(a) ***Random forests*** (Strobl et al., 2009; accounting for high correlations and differing scales) to identify the explanatory variables that are "important" to the response, with a reduced set selected using prediction MSE.
(b) The reduced set of potential explanatory variables are considered in a ***generalised additive model (GAM)***, allowing for smooth, nonlinear relationships, with interactions modelled using tensor products. Stage 1 uncertainties are included through inverse-variance weighting, via the gamm function in the mgcv package in R.

4 Results

The dragonfly occupancy and detection probabilities varied widely within the community, as shown in Figure 1. The estimated detection probability is below 50% for most species, showing the importance of accounting for uncertainty in observed species occupancy.

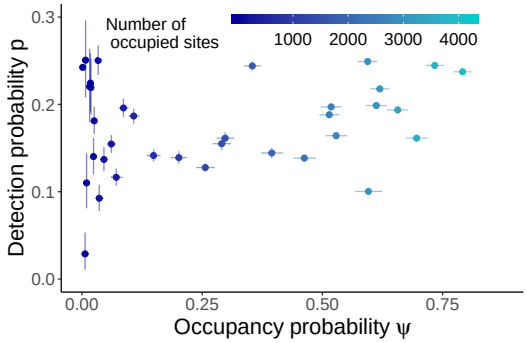


FIGURE 1. Estimated species occupancy and detection probabilites with number of sites each of the species is estimated to be present.

For stage 2, we provide an example for moderate alkalinity, deep lakes. Random forests identified the 6 most relevant potential explanatory variables to dragonfly richness from a dataset of 144 potential explanatory variables. Of these, two variables represented the same parameter (% agriculture) at two spatial scales. Only the most important of these two variables was retained. The 5 remaining potential explanatory variables were considered in a quasipoisson-response GAM, incorporating the inverse of the stage 1 prediction variance as weights. Figure 2 shows the smooths for the resulting model. Log(catchment mean rainfall) has a positive coefficient (1.18). Square root of 500m buffer Strahler 2 length per ha is a connectivity variable with a generally positive effect, but logit(% agriculture in catchment)

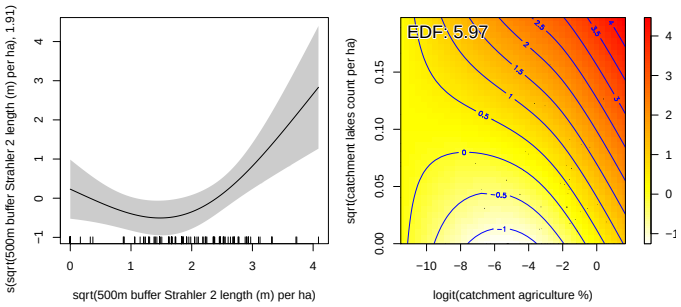


FIGURE 2. Fitted smooths for moderate alkalinity, deep lakes.

interacts with a connectivity variable (square root of catchment lake count per ha), suggesting that the effects of connectivity on dragonfly species richness vary with increasing stress caused by nearby agriculture. The model explains approximately 28% of variance in estimated species richness, appearing to be a moderately good fit to the data.

5 Discussion and conclusions

This two-stage approach presents a computationally efficient method for dimension reduction of the nested spatial covariate space to model species richness in the presence of imperfect detection.

Acknowledgments: JB's work is financially supported by the Mexican Council of Science and Technology (CONACyT) under scholarship number 494334. All others worked as part of NERC Hydroscape (NE/N005740/1). Steve Brooks (NHM) provided additional advice on dragonfly ecology.

References

Kéry, M. and Royle, J. (2008). Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys. *Journal of Applied Ecology*, **45**, 589 – 598.

Powney, G., Brooks, S., Barwell, L., Bowles, P., Fitt, R., Pavitt, A., Spriggs, R., and Isaac, N. (2014). Morphological and geographical traits of the British Odonata. *Biodiversity Data Journal*, **2**, e1041.

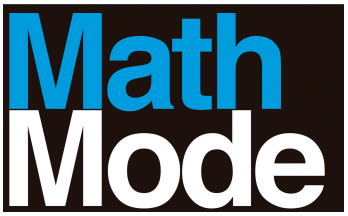
Strobl, C., Hothorn, T., and Zeileis, A. (2009). Party on! A new, conditional variable importance measure for random forests available in the party package. *The R Journal*, **1**, 14 – 17.

Taylor, P., Carvalho, L., Law, A., Baker, A., Chapman, D., Willby, N., Wilkie, C. (in preparation). The connectivity of UK freshwaters.

Author Index

35th IWSM 2020 Sponsors

We are very grateful to the following organisations for sponsoring 35th IWSM 2020.



Zabalduz

Jardunaldi, kongresu, sinposio, hitzaldi eta omenaldien argitalpenak

Publicaciones de jornadas, congresos, simposiums, conferencias y homenajes

INFORMAZIOA ETA ESKARIAK • INFORMACIÓN Y PEDIDOS

UPV/EHUko Argitalpen Zerbitzua • Servicio Editorial de la UPV/EHU
argitaletxea@ehu.eus • editorial@ehu.eus
1397 Posta Kutxatila - 48080 Bilbo • Apartado 1397 - 48080 Bilbao
Tfn.: 94 601 2227 • www.ehu.eus/argitalpenak

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea