Contents

Hein Putter, Cristian Spitoni, Andrew Titman: Multi-state modeling: beware the Markov assumption! ................................. 1
Catherine Matias: A semiparametric extension of the stochastic block model for longitudinal networks .............................. 7
Leonhard Held: Challenging Statistical Significance: A Case for Reverse-Bayes .......................................................... 8
Iain Currie: GLAM: The IWSM Story ........................................ 17
John Aston: Applied Statistics for Covariance Functions ............. 29
Timo Adam, Andreas Mayr, Thomas Kneib, Roland Langrock: Statistical boosting for Markov-switching distributional regression models ................................................................. 30
Clair Barnes, Richard E. Chandler: New approaches to post-processing of multi-model ensemble weather forecasts .......... 36
Pariya Behrouzi, Fentaw Abegaz, Ernst C. Wit: Dynamic Chain Graph Models for Ordinal Time Series Data ............... 40
Moritz Berger, Michael Wagner, Matthias Schmid: Modelling Ratio Outcomes with Gamma Distributed Components ....... 44
Mauro Bernardi, Paola Stolfi: Modelling Cryptocurrencies using Robust Time Varying Undirected Graphs ..................... 50
Aoibheann Brady, Julian Faraway, Ilaria Prosdocimi: Attribution of large-scale drivers of peak river flows in Ireland ..... 54
Roberto Colombi, Sabrina Giordano: Modelling uncertainty and inaccuracy in ordinal data ............................................. 59
Nelu Dan, Ruggero Bellio, Thomas Reutterer: A Note on Latent Rating Regression for Aspect Analysis of User-Generated Content .................................................................................. 63
Emiko Dupont, Ilaria Prosdocimi, Daniel Simpson, Nicole Augustin: Statistical modelling of time dependence in peak river flows .................................................................................. 69
Paul H. C. Eilers: Goodbye moments, hello expectiles .......... 78
Amira Elayouty, Marian Scott, Claire Miller, Susan Waldron: Functional principal component analysis for non-stationary dynamic time series .............................................................................. 84
Federica Nicolussi, Manuela Cazzaro: Chain Regression Graph Models for Context-Specific independencies .............. 90
Anna France, Deborah A. Costain, Andrew C. Titman: Handling Time Dependent Covariate Effects in Multiple Imputation: Application to Stroke in a Cox Modelling Framework .......... 96
MARCO GERACI: Additive quantile mixed models with an application to longitudinal accelerometer measurements ........................................ 102
MENGYI GONG, CLAIRE MILLER, MARIAN SCOTT: Spatio-temporal modelling of remote-sensing lake surface water temperature data 106
STEPHEN D. GREGORY: Could bigger be better? Longer Atlantic salmon smolts seem more likely to return as adults...................... 112
LEONARDO GRILLI, MARIA FRANCESCA MARINO, OMAR PACCAGNELLA, CARLA RAMPICHINI: Multilevel modelling with level 2 missing covariates: the relationship between student ratings and teacher beliefs and practices............................ 118
ANDREAS GROLL, TREvor HASTIE, THOMAS KNEIB, GERHARD TUTZ: Boosting Methods for Effects Selection in Cox Frailty Models ................................................................. 122
GILLIAN HELLER, WENLI HU, BRETT SWANSON: Inflated Beta Regression Modelling for Speech Intelligibility Testing ............... 128
JENNIFER HELLIER, RICHARD EMSLEY, ANDREW PICKLES: Estimating dose-response for time-to-remission with instrumental variable adjustment ..................................................... 134
MAIKE HOHBERG, PETER PÜTZ, THOMAS KNEIB: Treatment effects beyond the mean using distributional regression ...................... 139
SEN HU, ADRIAN O’HAGAN, T. BRENDAN MURPHY: Mixture of Experts Models for Joint Insurance Claims Modelling with Multivariate Gamma Distributions ........................................... 145
SARAH KARASEK, HERWIG FRIEDL, PETER SCHEIBELHOFER: Statistical Analysis of Semiconductor Images .......................................... 151
NADJA KLEIN, MANUEL CARLAN, THOMAS KNEIB, STEFAN LANG, HELGA WAGNER: Effect Selection in Distributional Regression .... 157
PHILIPPE LAMBERT AND VINCENT BREMHORST: Identification issues and efficient inference in semi-parametric cure survival model with shared covariates ......................................................... 163
MICHAEL LEBACHER, PAUL W. THURNER, GÖRAN KAURMANN: International Arms Transfers: A Dynamic Separable Network Model With Heterogeneity Components .............................. 168
DAE-JIN LEE AND MARÍA DURBÁN: Predicting above-ground tree biomass: a Seemingly Unrelated Penalized Regression approach . 174
DUNCAN LEE: Quantifying the impact of air pollution on respiratory prescription rates in Scotland ...................................................... 179
MAGDALENA LEITNER, HELGA WAGNER: Bayesian Effect Fusion for Categorical Predictors in Logistic Regression .......................... 183
ALTEA LORENZO-ARRIBAS, MARK J. BREWER, ANTONY M. OVERSTALL: Constrained log-likelihood for partial proportional odds models for ordinal response variables ........................................ 189
ROBSON J. M. MACHADO, ARDO VAN DEN HOUT, GIAMPIERO MARIA: Penalised maximum likelihood estimation in multistate models with an application to cardiovascular disease ............ 195
Gilbert MacKenzie, Yasin Al-tawarah, Milisa Blagojevic and Kevin Burke: Non-PH parametric survival modelling  
Luca Maestrini, Matt P. Wand: Variational message passing for skew t regression  
Kenan M. Matawie and Sargon Hasso: Evaluating Statistical Information Retrieval Models with Different Indexing Enhancement Strategies  
Mario Martínez-Araya and R. Murray Lark: Modelling of X-ray density (greyscale) signals from CT-scan images of rooted soil  
Annette Møller, Jan Gertheiss: A classification tree for functional data  
Vicente Núñez-Antón, Ainhoa Oguiza, Jorge Virto: Analysis of Dependent Pseudo-Panel Data: Proposal and Comparison of Three Different Methodological Approaches  
Marius Ötting, Roland Langrock, Christian Deutscher, Vianey Leos-Barajas: The Hot Hand in Professional Darts  
Defen Peng, Kevin Burke, Gilbert MacKenzie: The Weibull MPR model for interval censored survival data  
Jennifer Pohle, Ruth King, Mihaela van der Schaar, Roland Langrock: Coupled Markov-switching regression: inference and a case study using electronic health record data  
Ori Rosen, Michael Bertolacci, Edward Cripps, Sally Cripps: Covariate-Dependent Spectral Estimation of Multiple Nonstationary Time Series  
Bruno Santos, Thomas Kneib: Structured additive multiple-output noncrossing Bayesian quantile regression models  
Gunther Schauberger, Patrick Mair: A Regularization Approach for the Detection of Differential Item Functioning in Generalized Partial Credit Models  
Francesco Sera, Antonio Gasparini: An extended random-effects framework for meta-analysis  
Mahdi Shafiee Kamalabad and Marco Grzegorczyk: Non-homogeneous dynamic Bayesian networks with edge-wise coupled parameters  
Rosaria Simone: Accelerated estimation of mixture models for rating variables  
Andrea Sottosanti, Mauro Bernardi, Alessandra R. Brazzale: Mining signals of astronomical sources via Bayesian non-parametric mixture modelling  
Oliver Stoner: Correcting Under-Recording of Eruptions in Historical Volcano Data  
Ejike R. Ugba, Jan Gertheiss: An Augmented Likelihood Ratio Index for Categorical Response Models
Multi-state modeling: beware the Markov assumption!

Hein Putter\textsuperscript{1}, Cristian Spitoni\textsuperscript{2}, Andrew Titman\textsuperscript{3}

\textsuperscript{1} Leiden University Medical Center, The Netherlands
\textsuperscript{2} University of Utrecht, The Netherlands
\textsuperscript{3} Lancaster University, United Kingdom

E-mail for correspondence: h.putter@lumc.nl

\textbf{Abstract:} There has been a remarkable surge of activity lately on the topic of non-parametric estimation of transition probabilities in non-Markov multi-state models. This work summarizes the most flexible of recent approaches, the landmark Aalen-Johansen estimator, and discusses ongoing work on testing the Markov assumption, inspired by the landmark Aalen-Johansen estimator. The landmark Aalen-Johansen estimator is based on a subsample of the complete data, containing all subjects present in a specific state at a specific time point. Because it is based on reduced data, it is less efficient than the Aalen-Johansen estimator, which might still perform reasonably well in situations where the multi-state model shows only mild deviations from Markovianity. In the final part we will explore ways of combining the efficiency of the Aalen-Johansen approach with the robustness of the landmark Aalen-Johansen approach.

\textbf{Keywords:} Multi-state models; Transition probabilities; Markov assumption.

1 Multi-state models, the Aalen-Johansen estimator and the Markov assumption

Multi-state models are finding increased application in medical research. They allow a detailed view of the disease or recovery process of a patient, and they can be used to obtain prediction probabilities of future events, after a given event history. The relevant quantities for these prediction probabilities in multi-state terminology are the transition probabilities, the probabilities to be in a state $m$ at time $t$, given that the patient is in state $\ell$ at an earlier time $s$. When the multi-state model is Markov, an elegant theory connects the transition intensities of the multi-state model to the

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
transition probabilities, leading to the Aalen-Johansen estimator (Aalen & Johansen 1978).

1.1 Notation and concepts

Let $X(t)$ be a random multi-state process, taking values in the state space $1, \ldots, K$, with $K$ finite. Let $\mathcal{H}_t$ denote the history of the multi-state process up to and including time $t$. The transition hazards are defined by

$$\lambda_{jk}(t \mid \mathcal{H}_{t-}) = \lim_{\Delta t \downarrow 0} \frac{P(X(t + \Delta t) = k \mid X(t) = j, \mathcal{H}_{t-})}{\Delta t}.$$ 

In general (non-Markov) multi-state models, these transition hazards will depend not only on the present state $j$ at time $t$, but also on the further past $\mathcal{H}_{t-}$. When the multi-state process is Markov, the transition hazards simplify to

$$\lambda_{jk}(t) = \lim_{\Delta t \downarrow 0} \frac{P(X(t + \Delta t) = k \mid X(t) = j)}{\Delta t}.$$ 

The transition probabilities are defined by

$$P_{\ell m}(s, t \mid \mathcal{H}_{s-}) = P(X(t) = m \mid X(s) = \ell).$$

For Markov models, the transition probabilities simplify to $P_{\ell m}(s, t) = P(X(t) = m \mid X(s) = \ell)$. The final concept of interest is the state occupation probabilities

$$\pi_m(t) = P(X(t) = m) = \sum_{\ell} \pi_{\ell}(0) P_{\ell m}(0, t),$$

1.2 Data and estimators

The observed data consist of right censored versions of the multi-state process $X_i(\cdot)$. Let $C_i$ be a right censoring time, for $i = 1, \ldots, n$, assume that $X_i(\cdot)$ and $C_i$ are independent and identically distributed, and define the censored counting and at-risk processes

$$N_{ijk}(t) = \#\{u \leq t, X_i(u-) = j, X_i(u) = k, C_i \geq u\},$$

$$Y_{ij}(t) = 1\{X_i(t-) = j, C_i \geq t\},$$

and their sums over the individuals $N_{\bullet jk}(t) = \sum_{i=1}^n N_{ijk}(t)$, and $Y_{\bullet j}(t) = \sum_{i=1}^n Y_{ij}(t)$. Non-parametric estimates of the transition intensities are defined as extensions of the Nelson-Aalen estimators

$$\hat{\Lambda}_{jk}(t) = \sum_{0 < u \leq t} \frac{dN_{\bullet jk}(u)}{Y_{\bullet j}(u)} = \sum_{0 < u \leq t} \Delta \hat{\Lambda}_{jk}(u).$$

To define the Aalen-Johansen (AJ) estimator, it is convenient to gather these transition intensity estimates into $K \times K$ matrices $\hat{\Lambda}(t)$, where the $(j, k)$th element contains $\hat{\Lambda}_{jk}(t)$ (0 if a direct transition from $j$ to $k$ is not possible), and the diagonal elements are defined as $\hat{\Lambda}_{jj}(t) = -\sum_{k \neq j} \hat{\Lambda}_{jk}(t)$. 

Similarly, put the transition probabilities \( P_{\ell m}(s, t) \) into a \( K \times K \) matrix \( P(s, t) \). The Aalen-Johansen estimator of \( P(s, t) \) is then given by
\[
\hat{P}(s, t) = \prod_{s < u \leq t} \left( I + \Delta \hat{\Lambda}(u) \right).
\]

The Aalen-Johansen estimator provides a consistent estimator of the transition probabilities under an independent censoring assumption, and if the multi-state model is Markov.

With \( \hat{\pi}(0) \) the \( 1 \times K \) vector containing the empirical proportions \( \hat{\pi}_k(0) = n^{-1} \sum_{i=1}^n 1 \{ X_i(0) = k \} \), under appropriate conditions \( \hat{P}(t) = \hat{\pi}(0) \hat{P}(0, t) \) provides a consistent estimator of the state occupation probabilities \( P(X(t) = m) \). Although the non-parametric Aalen-Johansen estimator will not in general give consistent estimators of the transition probabilities in non-Markov multi-state models, Datta and Satten (2001) have shown that, even for non-Markov multi-state models, the estimator of state occupation probabilities derived from the non-parametric Aalen-Johansen estimator is consistent.

2 The landmark Aalen-Johansen estimator

The idea behind the landmark Aalen-Johansen estimator (Putter & Spitoni 2016) is quite simple. A transition probability \( P(X(t) = m \mid X(s) = \ell) \) can be seen as a state occupation probability \( P(X(t) = m) \) in a subset of subjects that happen to be in state \( \ell \) at time \( s \). This is not unlike landmarking, where time-dependent covariates are dealt with by selecting subsets of individuals that are at risk at time \( s \). For fixed time \( s \) and state \( \ell \), the estimator is based on re-estimated transition rates obtained from selecting subjects with \( X_i(s) = \ell \). We will use the superscript \( (LM) \) to denote versions of estimators based on a landmark data set which selects subjects who are at state \( \ell \) at time \( s \). The landmark based versions of counting and at risk processes, and of transition rates estimates, are defined as
\[
N_{\bullet,jk}^{(LM)}(t) = \sum_{i=1}^n N_{ij}(t) 1\{X_i(s) = \ell\},
\]
\[
Y_{\bullet,j}^{(LM)}(t) = \sum_{i=1}^n Y_{ij}(t) 1\{X_i(s) = \ell\}, \quad \hat{\Lambda}_{jk}^{(LM)}(t) = \sum_{0 < u \leq t} \frac{dN_{\bullet,jk}(u)}{Y_{\bullet,j}(u)}.
\]

The estimated transition rates are gathered into a \( K \times K \) matrices \( \hat{\Lambda}^{(LM)}(t) \) as before. Finally, the landmark Aalen-Johansen (LMAJ) estimator is given by
\[
\hat{P}^{\text{LMAJ}}_{\ell m}(s, t) = \hat{\pi}^{(LM)}(s) \prod_{s < u \leq t} \left( I + \Delta \hat{\Lambda}^{(LM)}(u) \right),
\]
with \( \hat{\pi}^{(LM)}(s) \) a \( 1 \times K \) vector with \( \hat{\pi}_\ell^{(LM)}(s) = 1 \), and other values equal to 0.
3 Tests for the Markov assumption

The difference between the AJ and LMAJ estimators can be seen by looking at the estimated cumulative hazards. These are shown for a subset of 488 patients of the CSL1 trial, a placebo-controlled randomized clinical trial to evaluate the efficacy of prednisone in patients with histologically verified liver cirrhosis. The endpoint of the trial was overall survival; no significant difference was found between the two arms, but more detailed analysis revealed a number of interactions between (time-dependent) covariates and treatment. One of these time-dependent covariates is the prothrombin index, an indicator of liver functioning, which is dichotomized into normal and low (less than 70% of normal values). Changes between normal and low are assumed to occur at the moment of measurement. Patients with a low level of prothrombin can recover and return to a normal level. Their levels may decrease and increase several times. This can be incorporated by defining normal and low prothrombin levels as states (here 1 and 2, respectively) in a three-state (reversible illness-death) multi-state model with death as third state.

Figure 1 shows the estimated cumulative hazards based on all placebo-treated subjects, and based on the subset of placebo-treated subjects in state 1 at time \( s = 2 \) and in state 2 at time \( s = 2 \), separately. A similar plot can be made for the prednisone group (not shown here). The landmark Aalen-Johansen estimator would use the subset of placebo-treated subjects in state 1 at time \( s = 2 \) for estimating \( P_{1m}(s,t) \) and the subset of placebo-treated subjects in state 2 at time \( s = 2 \) for estimating \( P_{2m}(s,t) \). This comparison between estimated hazards is the basis of a formal test for the Markov assumption. A log-rank test statistic \( Z_{jk}^\ell(s) \) can be constructed for each transition \( j \rightarrow k \), contrasting the transition hazards \( \lambda_{jk}^\ell \) between subjects in state \( \ell \) and not in state \( \ell \) at time \( s \). Each log-rank test statistic \( Z_{jk}^\ell(s) \) has a standard normal distribution, if the Markov assumption holds, but the challenge is to obtain a global test of the Markov assumption. This can be achieved by combining the individual log-rank test statistics over different transitions, as well as over qualitative states \( \ell \) and time \( s \). This is outside the scope of this short paper, however.

4 The hybrid Aalen-Johansen estimator

The LMAJ estimator is robust against violations of the Markov assumption, but since it is based on a subset of subjects in “the right place at the right time”, it is less efficient than the AJ estimator. In Markov models we would like to use the AJ estimator, and in non-Markov models the LMAJ estimator. Figure 1 shows that only transition 1 → 2 is really different between subjects in state 1 and in state 2 at time \( s = 2 \). The Markov test described above formally confirms this. One could then formulate a hybrid
Aalen-Johansen estimator of $P_{\ell m}(s, t)$, by using $\widehat{\Lambda}_{jk}$ for transitions $j \rightarrow k$ that do not differ between subjects in state $\ell$ and not in state $\ell$ at time $s$, and $\widehat{\Lambda}_{jk}^{(LM)}$ for transitions $j \rightarrow k$ that do differ. When these estimated transition hazards are collected in matrix $\widehat{\Lambda}^{(H)}$, the hybrid Aalen-Johansen estimator is then given by

$$\widehat{P}_{\ell m}^{\text{HAJ}}(s, t) = \pi^{(LM)}(s) \prod_{s < u \leq t} \left( I + \Delta \widehat{\Lambda}^{(H)}(u) \right),$$

with $\pi^{(LM)}(s)$ as before. When a global test for the Markov assumption is not rejected, one could use the Aalen-Johansen estimator.

Figure 2 shows the result of this procedure, applied to the CSL data for $P_{12}(s, t)$, $s = 2$, separately for placebo (solid) and prednisone (dashed) patients. The LMAJ estimator suggests a much larger difference between
the two treatments than AJ. Overall, the hybrid Aalen-Johansen lies in between AJ and LMAJ.

![Graphs showing transition probability estimates](image)

**FIGURE 2.** Transition probability estimates of $P_{12}(s,t)$, $s = 2$, from Aalen-Johansen, landmark Aalen-Johansen, and hybrid Aalen-Johansen.

**References**


A semiparametric extension of the stochastic block model for longitudinal networks

Catherine Matias

1 Centre National de la Recherche Scientifique, France.

E-mail for correspondence: Catherine.Matias@math.cnrs.fr

To model recurrent interaction events in continuous time, an extension of the stochastic block model is proposed where every individual belongs to a latent group and interactions between two individuals follow a conditional inhomogeneous Poisson process with intensity driven by the individuals’ latent groups. The model is shown to be identifiable and its estimation is based on a semiparametric variational expectation-maximization algorithm. Two versions of the method are developed, using either a nonparametric histogram approach (with an adaptive choice of the partition size) or kernel intensity estimators. The number of latent groups can be selected by an integrated classification likelihood criterion. Finally, we demonstrate the performance of our procedure on synthetic experiments, analyse two datasets to illustrate the utility of our approach and comment on competing methods.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
Challenging Statistical Significance: A Case for Reverse-Bayes

Leonhard Held

1 Epidemiology, Biostatistics and Prevention Institute (EBPI), University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland

E-mail for correspondence: leonhard.held@uzh.ch

Abstract: Analysis of credibility is a reverse-Bayes technique that has been proposed by Matthews (2001a,b, 2018) to overcome some of the shortcomings of significance tests. A significant result is deemed credible if current knowledge about the effect size is in conflict with a sceptical prior derived from the data that makes the effect non-significant. In this talk I describe this approach and propose to use Bayesian predictive tail probabilities to quantify the evidence for credibility. This gives rise to a $p$-value for credibility, $p_C$, taking into account both the internal and the external evidence for an effect. The assessment of intrinsic credibility is based on the internal data only and leads to a new threshold for ordinary significance that is remarkably close to the recently proposed 0.005 level. Finally, a $p$-value for intrinsic credibility, $p_{IC}$, is proposed that is a simple function of the ordinary $p$-value for significance and has a direct frequentist interpretation in terms of the replication probability that a future study under identical conditions will give an estimated effect in the same direction as the first study.

Keywords: Analysis of Credibility; Confidence Interval; $p$-value; Replication Probability; Significance Test

1 Introduction

“We can make judgments of initial probabilities and infer final ones, or we can equally make judgments of final ones and infer initial ones by Bayes’s theorem in reverse.”

IJ Good (1983)

Standard $p$-values for point null hypotheses still dominate most of the applied literature (Greenland and Poole, 2013), despite the fact that they are commonly misused and misunderstood (Wasserstein and Lazar, 2016;
Matthews et al., 2017). Although being criticised intensively in the literature, the dichotomisation of p-values into “significant” and “non-significant” and thus into results worthy or not of taking seriously is still commonplace in practice.

In a series of papers, Robert Matthews (Matthews, 2001a, b, 2018) has developed the Analysis of Credibility, a specific reverse-Bayes method to assess the credibility of significant findings. Reverse-Bayes approaches allow the extraction of the properties of the prior distribution needed to achieve a certain posterior statement for the data at hand. The idea to use Bayes’s theorem in reverse originates in the work by IJ Good (Good, 1950, 1983) and is increasingly used to assess the plausibility of scientific findings (Greenland, 2006, 2011; Held, 2013; Colquhoun, 2017).

AnCred is based on a conventional confidence interval (and not just a p-value) for an unknown effect size \( \theta \) with lower limit \( L \) and upper limit \( U \), say. Assume that both \( L \) and \( U \) are symmetric around the effect size estimate \( \hat{\theta} \) (assumed to be normally distributed) and that both are either positive or negative, i.e. the effect is significant at significance level \( \alpha \), usually 5%. The normality assumption can be justified as an approximation for continuous, survival, binary and count outcomes, if the sample size is large (Spiegelhalter et al., 2004, Section 2.4).

Matthews (2001a, b) proposed assessing the credibility of a statistically significant finding by computing from the data a sceptical prior distribution for the effect size \( \theta \), normal with mean zero, that – combined with the information given in the confidence interval for \( \theta \) – results in a posterior distribution which is just “non-significant” at level \( \alpha \), in the sense that either the \( \alpha/2 \) or the \( 1-\alpha/2 \) posterior quantile is zero. If previous evidence indicates that plausible values for the parameter \( \theta \) are in conflict with this sceptical prior, the existence of the effect size estimate \( \hat{\theta} \) will be deemed credible.

It can be shown that the limits \( \pm S \) of the corresponding equi-tailed prior credible interval at level \( 1-\alpha \) are

\[
S = \frac{(U - L)^2}{4\sqrt{UL}}, \tag{1}
\]

where \( S \) is called the sceptical limit and the interval \( [-S, S] \) is called the critical prior interval. Note that (1) holds for any value of \( \alpha \), not just for the traditional 5% level. Equivalently, the variance \( \tau^2 \) of the sceptical prior can be expressed as a function of the variance \( \sigma^2 \) (the squared standard error, assumed to be known) of the estimate \( \hat{\theta} \), the corresponding test statistic \( t = \hat{\theta}/\sigma \) and \( z_{\alpha/2} \), the \( 1-\alpha/2 \) quantile of the standard normal distribution:

\[
\tau^2 = \frac{\sigma^2}{t^2/z_{\alpha/2}^2 - 1}, \tag{2}
\]

where \( t^2 > z_{\alpha/2}^2 \) is required for significance at level \( \alpha \). Equation (2) shows
that the prior variance $\tau^2$ can be smaller or larger than $\sigma^2$, depending on the value of $t^2$. If $t^2$ is close to $z_{\alpha/2}^2$ (i.e. the effect is “borderline significant”), a weak sceptical prior with large variance will be sufficient to make the posterior just non-significant. This implies that the significant finding is only credible if prior evidence for a relatively large effect size already exists. If $t^2$ is substantially larger than $z_{\alpha/2}^2$, then the prior variance will be relatively small and a smaller effect size of the original study is sufficient to achieve credibility.

For illustration I revisit a recently published randomized placebo-controlled clinical trial (Hayward et al., 2017) on the efficacy of corticosteroids in the treatment of sore throat (Held, 2017). There were 102/288 events in the intervention group and 75/277 events in the control group for the outcome complete resolution of pain at 48 hours. The number of patients and the number of events are relatively large in both groups, so the normality assumption for the log odds ratio estimate $\hat{\theta}$ is justified. A log odds ratio estimate of $\exp(\hat{\theta}) = 0.39$ can easily be calculated (95% CI from 0.03 to
Analysis of Credibility for the log odds ratio \( \theta \) gives the sceptical limit 0.84. The associated standard deviation \( \tau = 0.43 \) of the sceptical prior is hence considerably larger than the standard error \( \sigma = 0.18 \). Figure 1 displays the sceptical prior together with the confidence interval for the effect size and the associated posterior. Credibility of the Hayward et al. (2017) results can be investigated in the light of previous clinical trials on the same research question. Sadeghirad et al. (2017) performed a systematic review and identified three preceding studies relevant for the Hayward et al. (2017) analysis. A random-effects meta-analysis of the three trials preceding the Hayward (2017) study gives a log odds ratio estimate of \( \hat{\theta}_0 = 1.35 \) (95% CI from 0.48 to 2.22, \( p = 0.0023 \)), also shown in Figure 1. Since the point estimate \( \hat{\theta}_0 \) is larger than the sceptical limit 0.84, the results from the Hayward et al. (2017) study are considered credible (at the 95% level) using the Matthews (2001b) approach. Note that the uncertainty of \( \hat{\theta}_0 \) is here not taken into account, a deficiency that we address later.

Matthews (2018) has recently proposed the concept of intrinsic credibility for “out of the blue” findings without any prior support. He suggests to declare a significant result as intrinsically credible, if the effect size estimate \( \hat{\theta} \) is outside the sceptical prior interval, i.e. \(|\hat{\theta}| > S\). For \( \alpha = 0.05 \), he shows that this check is equivalent to the conventional two-sided \( p \)-value being smaller than 0.0127. Intrinsic credibility is defined differently later, leading to the more stringent threshold 0.0056.

AnCred is an attractive method for evidence quantification in the light of previous studies and has been advocated by key figures in health sciences (Spiegelhalter, 2004; Spiegelhalter et al., 2004; Chalmers and Matthews, 2006; Greenland, 2008; Colquhoun, 2017). AnCred has already been used in a number of clinical studies and meta-analyses, e.g. Barclay et al. (2008); Engelman and Maeyens (2010); Sinclair et al. (2016). The concept of intrinsic credibility can be used as a “double-check” of a significant finding in the absence of prior knowledge. For example, to assess whether the evidence from a non-confirmatory study meets a more stringent hypothetical confirmatory standard, intrinsic credibility may provide a possible alternative to predictive evidence threshold scaling (Neuenschwander et al., 2017).

More recently, so-called \( p \)-values for credibility based on the Box (1980) approach to assess prior-data conflict have been proposed in Held (2017). In contrast to the dichotomous AnCred result (credible yes/no), \( p \)-values for credibility provide a quantitative assessment of the evidence for credibility. Furthermore, they take the uncertainty of the previous estimate \( \hat{\theta}_0 \) into account and have the attractive property that they do not depend on the level \( 1 - \alpha \) of the original confidence interval for the parameter of interest. The \( p \)-value for credibility \( p_C \) is based on the the largest confidence level \( 1 - \alpha^* \), where the stated effect is (just) credible at level \( 1 - \alpha^* \). In analogy
to the well-known duality of confidence intervals and standard p-values, the p-value for credibility $p_C$ is then defined as $p_C = \alpha^*$. To determine the p-value for credibility $p_C$ in the normal setting, let $t_0 = \hat{\theta}_0/\sigma_0$ denote the test statistic of the preceding study and $t = \hat{\theta}/\sigma$ the corresponding test statistic of the current study. Further, let $c = \sigma^2/\sigma_0^2$ denote the ratio of the variances of the current and preceding effect size estimates. The p-value $p_C$ for credibility can then be shown to be the solution of the equation

$$\left(\frac{t_0^2}{\sigma_0^2/2} - 1\right)\left(\frac{t^2}{\sigma^2/2} - 1\right) = c,$$

(3)

see Held (2017, Section 2) for a derivation. Returning to the example shown in Figure 1, the p-value for credibility turns out to be $p_C = 0.045$, indicating evidence for credibility of the results reported in the Hayward et al. (2017) trial in the light of previous studies.

Equation (3) reveals some interesting and attractive properties of p-values for credibility: First, $p_C$ is always larger than both ordinary p-values $p$ and $p_0$, say, from the current and previous studies, respectively, i.e. $p_C > p$ and $p_C > p_0$. If previous evidence for a treatment effect is overwhelming, i.e. $p_0$ is very small, then $p_C$ will be only slightly larger than $p$. However, if $p$ is very small but $p_0$ isn’t, then $p_C$ will be even larger than $p_0$. Secondly, $p_C$ not only depends on $p$ and $p_0$, but also on the variances $\sigma^2$ and $\sigma_0^2$. This is illustrated in Figure 2, where $p = p_0 = 0.01$ in all three plots, so $t^2 = t_0^2$ and therefore $c = \sigma^2/\sigma_0^2 = \hat{\theta}^2/\hat{\theta}_0^2$. In the top panel we have $c = 1$, so $\sigma^2 = \sigma_0^2$ or equivalently $\hat{\theta} = \hat{\theta}_0$, and $p_C = 0.069$. In the middle panel, $c = 4$, so $\hat{\theta} = 2\hat{\theta}_0$. Now we obtain $p_C = 0.14$, so less evidence for credibility because the variance $\sigma^2$ of $\hat{\theta}$ is larger than the variance $\sigma_0^2$ of $\hat{\theta}_0$. Things are reversed in the bottom panel, where now $\sigma^2 = \sigma_0^2/4$ and there is more evidence for credibility ($p_C = 0.035$). This asymmetry in the incorporation of the original and replication study data is natural, placing more weight on current studies with large sample size (i.e. smaller variance) relative to the sample size of preceding studies.

The concept of intrinsic credibility (Matthews, 2018) has recently been combined with the Box (1980) approach to quantify prior-data conflict in Held (2018). Intrinsic credibility at the 5% level can then be shown to be equivalent to the requirement $p < 0.0056$ for the conventional two-sided p-value. This is remarkably close to the recently proposed new threshold of 0.005 for statistical significance (Johnson, 2013; Benjamin et al., 2018), for a recent discussion see Ioannidis (2018). Equivalently, intrinsic credibility can be assessed by a comparison of the sceptical prior variance $\tau^2$ and the data variance $\sigma^2$: if and only if $\tau^2 \leq \sigma^2$, then the result is intrinsically credible. Thus, we see immediately from Figure 1 that the result from the Hayward et al. (2017) trial is not intrinsically credible. A third way to assess intrinsic credibility is based on the so-called credibility ratio $U/L$ of the upper to the lower limit of the confidence interval: if $U/L < 5.828$ then
FIGURE 2. Analysis of Credibility with $p$-value $p_C$ for credibility. Both ordinary $p$-values are $p = p_0 = 0.01$ and $\alpha = 5\%$. In the top plot we have $c = 1$, in the middle $c = 4$ and in the bottom $c = 1/4$. The $p$-value for credibility $p_C$ varies between 0.035 and 0.14.
the result is intrinsically credible (Held, 2018). This way to assess intrinsic credibility is attractive, as it only needs the limits $U$ and $L$ of the confidence intervals. For example, the credibility ratio for the Hayward et al. (2017) log odds ratio estimate is $0.75/0.031 \approx 24$, far too large to achieve intrinsic credibility.

Note that the assessment of intrinsic credibility does not require to change the significance level $\alpha$, so $p$-values between 0.0056 and 0.05 are still “significant”, but not intrinsically credible. Results in this intermediate category have been called “suggestive” by Benjamin et al. (2018).

To quantify the evidence for intrinsic credibility, $p$-values for intrinsic credibility ($p_{IC}$) have been proposed in Held (2017). They are based on the conflict of the data with the sceptical prior derived from the very same data using the Box (1980) approach. The explicit formula

$$p_{IC} = 2 \left[ 1 - \Phi \left( \frac{t}{\sqrt{2}} \right) \right]$$

can be derived from (3) with $p_{IC} = p_C$, $t = t_0$ and $c = 1$. For example, the $p$-value for intrinsic credibility of the Hayward et al. (2017) estimate shown in Figure 1 is $p_{IC} = 0.13$. Thus, ignoring prior knowledge from previous studies, the significant result from Hayward et al. (2017) ($p = 0.033$) lacks evidence for intrinsic credibility.

The $p$-value for intrinsic credibility has an attractive interpretation as a replication probability, directly linking the AnCred with $p_{rep}$, the probability of replicating an effect (Senn, 2002; Killeen, 2005). Specifically, if the Hayward et al. (2017) would have been replicated under identical conditions (with equal sample size), then the probability to obtain an estimate in the same direction as in the original experiment is $p_{rep} = 1 - p_{IC}/2 = 0.93$.

References


GLAM: The IWSM Story

Iain Currie¹

¹ Heriot-Watt University, Edinburgh, UK
E-mail for correspondence: I.D.Currie@hw.ac.uk

Abstract: Data that can be arranged in an array are common in statistics and then models often have a row-column-depth-... structure. In such cases, data sets and models can be large, and can present computational problems even for modern computers, particularly when smoothing is carried out in a Generalized Linear Model (GLM) framework. A Generalized Linear Array Model or GLAM is a fast low-footprint method of computation for such data sets and models. The GLAM approach and algorithms were first described at the IWSM in Florence in 2004. We describe GLAM and then mention some of its many applications since that time. The IWSM conferences have played a key role in GLAM’s development. We tell the story of GLAM through these workshops, starting with the 1999 conference in Graz and finishing in Bristol in 2018.

Keywords: Arrays; Efficient Computation; GLAM; P-splines; Smoothing.

1 The Motivating Example

The forecasting of human mortality rates at individual ages is necessary for the correct pricing of annuity and pension products in the insurance business, the provision of state pensions and, more generally, care for the elderly in an ageing society. This is a hard problem since it is self-evident that the future course of mortality is dependent on many unknown factors, such as medical advances, drug resistance, obesity, and many more besides. Some measure of the reliability of any forecast is certainly required.

To fix ideas, we suppose we have mortality data for ages 50 to 90 and for years 1975 to 2015. These data lie naturally in a matrix with rows indexed by age and columns indexed by year. The data thus comprise two matrices $D = (d_{x,y})$, the matrix of the number of deaths at age $x$ in year $y$, and $E = (e_{x,y})$, the matrix of the total time lived at age $x$ in year $y$. Plots of the observed log hazard rates, $\log(d_{x,y}/e_{x,y})$, show that mortality increases

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
with age and decreases with time, as is very well-known; we try to capture
the underlying 2-dimensional pattern with a general smooth surface.
We use the method of P-splines (Eilers and Marx, 1996). We define B-spline
bases for age: \( B_a = \{ B_{a,1}, \ldots, B_{a,c_a} \} \) and year: \( B_y = \{ B_{y,1}, \ldots, B_{y,c_y} \} \). Let \( x_a = (x_1, \ldots, x_{n_a})^T \) and \( x_y = (y_1, \ldots, y_{n_y})^T \) denote the vectors of
ages and years. Let \( B_a = [B_{a,1}(x_a) : \ldots : B_{a,c_a}(x_a)] \), \( n_a \times c_a \), and \( B_y = [B_{y,1}(x_y) : \ldots : B_{y,c_y}(x_y)] \), \( n_y \times c_y \), be the marginal regression matrices
for age and year. Then

\[
B = B_y \otimes B_a, \quad n_a n_y \times c_a c_y,
\]

is a suitable regression matrix for 2-dimensional modelling of the age-year
log hazard surface; here \( \otimes \) denotes the Kronecker product. The left panel
of Figure 1 illustrates the underlying 2-dimensional basis. As in one dimen-
sion, we can associate each regression coefficient with the summit of its
associated B-spline. The right panel shows the location of both the data
and coefficient grid; GLAM takes advantage of this data and coefficient
structure.
We force smoothness of the fitted surface by smoothing (1) the columns of
the coefficient grid, i.e. smoothing by age, and (2) the rows of the coefficient
grid, i.e. smoothing by year. The 2-dimensional penalty is

\[
P = \lambda_a I_{c_y} \otimes D_a^T D_a + \lambda_y D_y^T D_y \otimes I_{c_a}
\]

where \( D_a \) and \( D_y \) are second order difference matrices, and \( I_{c_a} \) and \( I_{c_y} \)
are identity matrices; see Currie et al. (2004) for further details.
We assume that \( D_{x,y} \sim P(e_{x,y} \mu_{x,y}) \) where \( \mu_{x,y} \) is the hazard rate or force
of mortality at age \( x \) in year \( y \). If there is no penalty we have a generalized
linear model. One of the attractive features of P-splines is the way the penalty $P$ is incorporated into the scoring algorithm for a GLM. We have

$$(B^T\tilde{W}_d B + P)\hat{\theta} = B^T\tilde{W}_d \tilde{z}$$

where $\tilde{W}_d$ is the diagonal matrix of weights and $\tilde{z}$ is the working-variable vector; here $\sim$ and $\hat{}$ represent current and updated estimates respectively. In equation (3) there is no attempt to take advantage of the structure of the data and/or the model. The $n_a \times n_y$ data matrices $D$ and $E$ have been flattened to vectors $d$ and $e$, and the $c_a \times c_y$ coefficient matrix $\Theta$ has also been flattened to a vector $\theta$. In principle, this gives a simple solution but certainly fifteen years ago, when this work was first done, there were two worries: the model matrix was large and computational time was substantial. Importantly, there was little prospect of extending the computation to three or higher dimensions. This was the problem that GLAM set out to address.

2 Array Arithmetic

We consider a sequence of data, model matrices and coefficient structures:

1. One dimension: data vector $y$, $n_1 \times 1$, model matrix $X_1$, $n_1 \times c_1$, coefficient vector $\theta$, $c_1 \times 1$ (ordinary regression).

2. Two dimensions: data matrix $Y$, $n_1 \times n_2$, model matrix $X_2 \otimes X_1$, $n_1 n_2 \times c_1 c_2$, coefficient matrix $\Theta$, $c_1 \times c_2$ (our motivating example).

3. Three dimensions: data array $Y$, $n_1 \times n_2 \times n_3$, model matrix $X_3 \otimes X_2 \otimes X_1$, $n_1 n_2 n_3 \times c_1 c_2 c_3$, coefficient array $\Theta$, $c_1 \times c_2 \times c_3$.

Before we consider the efficient evaluation of the penalized scoring algorithm (3) we look at a simpler problem: the efficient evaluation of the fitted values $X\hat{\theta}$, where $\hat{\theta}$ is an estimate of $\theta$. In one dimension there is no special structure and we proceed as usual. In two dimensions with $X = X_2 \otimes X_1$ we have the beautiful and well-known formula

$$(X_2 \otimes X_1)\hat{\theta}, n_1 n_2 \times 1 \equiv X_1 \hat{\Theta} X_2^T = (X_2 (X_1 \hat{\Theta})^T)^T, n_1 \times n_2, \quad (4)$$

where $\equiv$ indicates that each side contains the same values but not arranged in the same way. Searle (1982, p333) gives this formula but what he doesn’t mention are two very remarkable properties:

1. the right hand side (RHS) operates sequentially on $\hat{\Theta}$ without the storage of the potentially large matrix $X_2 \otimes X_1$,

2. less obviously, the number of multiplications required to evaluate the RHS is substantially smaller than that required to evaluate the LHS.
The extension of the RHS of (4) to three dimensions requires: (1) the “multiplication” of the matrix $X_1$ onto the 3-dimensional coefficient array $\hat{\Theta}$, (2) the “transpose” of the result, and (3) the repetition of (1) and (2) with first $X_2$ and then $X_3$. We must define what we mean by multiplication and transpose in this situation.

We consider the “multiplication” of the matrix $X_1$, $n_1 \times c_1$, onto the array $A$, $c_1 \times c_2 \times c_3$. We think of $A$ as $c_3$ matrices of size $c_1 \times c_2$ and construct the $c_1 \times c_2 c_3$ matrix, $M$, as follows:

We form the ordinary matrix product $X_1 M$ with size $n_1 \times c_2 c_3$. Next we reinstate dimensions two and three of $X_1 M$ to give a $n_1 \times c_2 \times c_3$ array denoted $H(X_1, A)$, the $H$-transform of the array $A$ by the matrix $X_1$.

The second requirement is the definition of the “transpose” of an array. We define the rotation of the array $A$, $c_1 \times c_2 \times c_3$, to be the array $R(A)$, $c_2 \times c_3 \times c_1$, obtained by permuting the indices of $A$.

It is convenient to combine these two operations into a single function: the rotated $H$-transform of the array $A$ by the matrix $X_1$ is

$$\rho(X_1, A) = R\{H(X_1, A)\}. \quad (5)$$

We note that $\rho(X_1, A)$ is $c_2 \times c_3 \times n_1$ so this array is ready to receive, $X_2$, a matrix with $c_2$ columns.

We return to our original problem: the efficient evaluation of $X\hat{\theta}$ in item 3 at the beginning of this section. We have

$$(X_3 \otimes X_2 \otimes X_1)\hat{\theta}, n_1 n_2 n_3 \times 1 \equiv \rho\{X_3, \rho[X_2, \rho(X_1, \hat{\Theta})]\}, n_1 \times n_2 \times n_3. \quad (6)$$

We obtain a nice corollary to (6) if we replace $\hat{\theta}$ with $y$, $\hat{\Theta}$ with $Y$ and $X$ with $X^T$. Then (6) becomes

$$X^T y, c_1 c_2 c_3 \times 1 \equiv \rho\{X_3^T, \rho[X_2^T, \rho(X_1^T, Y)]\}, c_1 \times c_2 \times c_3. \quad (7)$$

If we also assume that $X^TX = I$ and we have independent normal errors with common variance (as is the case in a classical factorial design) then $\hat{\theta} = X^T y$ and (7) is a general (and compact) statement of Yates’s algorithm (Yates, 1937), a popular method of computation in factorial experiments in pre-computer days.
3 The GLAM algorithms

We turn now to the efficient computation of first the scoring algorithm (3) and second the standard errors of fitted values. We need one further definition: the row tensor of a matrix \( X \) with \( c \) columns is

\[
G(X) = (X \otimes 1^T) \ast (1^T \otimes X)
\]

(8)

where \( 1 \) is a vector of 1s of length \( c \) and \( \ast \) indicates element-by-element multiplication. We note that the columns of \( G(X) \) contain all the pairwise products \( x_j \ast x_k \) of the columns of \( X \). We now state the three GLAM algorithms in the 2-dimensional case \( X_1, n_1 \times c_1, X_2, n_2 \times c_2 \) and \( \Theta, c_1 \times c_2 \).

The general formulae are the obvious extensions of these formulae. Let \( X = X_2 \otimes X_1 \).

1. **Linear function:**

\[
X\theta, n_1n_2 \times 1 \equiv \rho[X_2, \rho(X_1, \Theta)], n_1 \times n_2.
\]

(9)

2. **Inner product function:** let \( \tilde{W} \) be the \( n_1 \times n_2 \) matrix of weights, i.e. \( \text{vec}(\tilde{W}) = \text{diag}(\tilde{W}_d) \). We have

\[
X^T \tilde{W}_d X, c_1c_2 \times c_1c_2, \equiv \rho[G(X_2)^T, \rho(G(X_1)^T, \tilde{W})], c_1^2 \times c_2^2.
\]

(10)

3. **Diagonal function:** we have \( \text{Var}(\hat{\theta}) = (X^T \tilde{W}_d X + P)^{-1} = S_m \), say. We seek the diagonal elements of \( \text{Var}(X\theta) = XS_mX^T \). This variance matrix is \( n_1n_2 \times n_1n_2 \) and is potentially huge; the GLAM formula calculates the \( n_1n_2 \) diagonal elements only. We reorganize \( S_m, c_1c_2 \times c_1c_2 \), into the matrix \( S, c_1^2 \times c_2^2 \) (see section 5 of Currie et al. (2006) for details of this reorganization). Then

\[
\text{diag}\{XS_mX^T\}, n_1n_2 \times 1, \equiv \rho[G(X_2), \rho(G(X_1), S)], n_1 \times n_2.
\]

(11)

A formal proof of the equivalences in (9), (10) and (11) is given in section 4 of Currie et al. (2006). Each GLAM algorithm requires element reorganization, notably from the output in array form of the inner product algorithm to the matrix form required by the scoring algorithm, and from the matrix form of \( S_m = \text{Var}(\hat{\theta}) \) to the array form of \( S \). Computationally, reorganization is very efficient and R-code for all three GLAM algorithms is given in section 5 of Currie et al. (2006).

We conclude with four remarks on the algorithms:

1. A nice feature of the RHSs of (9) and (11) is that the output matrices/arrays of fitted values and variances of fitted values match those of the original data matrices/arrays.
2. Each of these formulae processes each dimension in turn, a process we call *marginal processing*; this automatically solves any storage problems.

3. GLAM offers two ways of computing the effective dimension, \( ED \), of the fitted model. We have \( ED = \text{tr}(H) \) where \( H \) is the hat-matrix

\[
H = X(X^T\hat{W}_dX + P)^{-1}X^T\hat{W}_d = Q\hat{W}_d,
\]

say. Now \( \text{tr}(AB) = \text{tr}(BA) \) for conformable matrices so the second GLAM algorithm gives \( ED \) immediately. Rather more elegantly, let \( R \) be the array form of the diagonal elements of \( Q \) given by the third GLAM algorithm, i.e. the RHS of (11), and let \( W \) be the array form of the weights. Then, \( ED \) is the sum of the elements of \( R \ast W \).

4. The computations of \( X\hat{\theta}, X^T\hat{W}_dX \) and \( \text{diag}\left(X\operatorname{Var}(\hat{\theta})X^T\right) \) are apparently very different. Despite this, the GLAM algorithms all have the *same functional form* and depend on the *same two functions*, \( \rho(\cdot) \) and \( G(\cdot) \). This strikes me as very remarkable; at any rate, the algorithms are highly efficient computationally and can lead to improvements in computer time of orders of magnitude, and may even bring some problems within reach. Tables 2 and 3 in the RSS paper provide details of the efficiency gains possible with GLAM.

4 Applications

4.1 Modelling and forecasting mortality

The modelling of 2-dimensional mortality data with GLAM is described in Eilers et al. (2006) and Currie et al. (2006). Forecasting, considered in Currie et al. (2004) in pre-GLAM days, is easily incorporated into the GLAM system. The R-package MortalitySmooth (Camarda, 2012) for modelling and forecasting both 1- and 2-dimensional data makes full use of GLAM.

4.2 An Example in Three Dimensions

The data comprise the number of deaths from a respiratory disease classified by age: 1, . . . , 105, year of death: 1959, . . . , 1998, and month: 1, . . . , 12. This gives a \( 105 \times 40 \times 12 \) data array with 50400 entries. The data are smoothed across each dimension with \( P \)-splines and a \( 15 \times 10 \times 7 \) coefficient array. This is a large problem: the regression matrix alone has over \( 5 \times 10^7 \) elements. This example is described in Currie et al. (2006); the treatment of missing values, the choice of initial estimates and various devices for improving computational time still further are also discussed.
4.3 Scattered Data and Density Estimation

Surprisingly, some examples with scattered data can be fitted into the GLAM scheme. Eilers et al. (2006) described such an example in two dimensions. The method uses a nice trick: we enclose the data in a rectangle which is then divided into a large number of bins. For each bin we record (1) $W$, the number of observations and (2) $S$, the average value of these observations. The error introduced by this discretization is negligible and GLAM can be used. This approach was used in three papers on density estimation at the IWSM in Galway. The method of smoothing parameter selection was: AIC in Eilers and Marx (2006), mixed model in Durbán et al. (2006) and a Bayesian method in Lambert and Eilers (2006). All three papers exploited GLAM.

4.4 Spatio-temporal Smoothing

Dae-Jin Lee and Maríá presented two papers on spatio-temporal smoothing, the first in 2008 in Utrecht and then in 2009 in Ithaca. These papers gave a practical solution to the problem of fitting smooth space-time interaction models. The spatial data were scattered over two dimensions while the temporal data were regularly spaced. We have a 2-dimensional GLAM where space is one GLAM dimension and time is the other. In Utrecht they used a mixed model representation which gave a decomposition of the model terms similar to ANOVA which they christened $P$-spline ANOVA or PS-ANOVA. The following year they introduced the idea of nested bases; here the time part of the space-time interaction uses a basis that is nested in the basis for the main effect of time. The use of GLAM and the nested bases gave substantial improvements in computer time while PS-ANOVA gave a clear interpretation of model terms. See Lee and Durbán (2008, 2009, 2011) and Lee et al. (2013).

4.5 Lee-Carter Model

The Lee-Carter model (Lee and Carter, 1992) is widely used in the modelling and forecasting of mortality (our motivating example). With the notation of section 1 we have \( \log \mu_{x,y} = \alpha_x + \beta_x \kappa_y \), \( x = x_1, \ldots, x_n \), \( y = y_1, \ldots, y_n \). One problem with this model is that irregularities in the estimated $\beta$ terms and to a lesser extent the $\alpha$ terms lead to forecasts with unsatisfactory properties. This problem was initially addressed in Delwarde et al. (2007) and then more generally in Currie (2013). This second paper estimated smoothed and constrained regression coefficients in a GLM setting and made full use of GLAM. Unusually explicit formulae for GLAM’s expression for the inner products were given.
4.6 Additive GLAMs

GLAM is just a regression model so there is no reason why we should not have a model with more than one component. Our 2006 RSS paper presented an additive model with two GLAM components for the analysis of micro-array data. We could also have both GLAM components and regression components, a semi-GLAM. An example of this kind of model occurs in the analysis of variety trials. Fields are generally rectangular so fertility and any row-column effects can be modelled in a GLAM setting while variety effects are modelled in the usual way. In an important paper, Rodríguez-Álvarez et al. (2018) exploited earlier work to the full and used the mixed model decomposition of PS-ANOVA for the analysis of such trials. The work is supported by the R-package SpATS.

4.7 Components of Mortality

We return to our motivating example, the modelling of mortality data. We consider first 1-dimensional data and the smooth modelling of mortality by age for a given year. Let \( B_a, n_a \times c_a \), be the regression matrix of \( B \)-splines with penalty matrix \( D_a^T D_a, c_a \times c_a \), as before. In recent years, smoothing parameter selection has increasingly been through a mixed model representation. One method of achieving this is with the singular value decomposition (SVD). Let \( U_a \Lambda_a U_a^T \) be the SVD of \( D_a^T D_a \). With second order differences there are two zero eigenvalues so

\[
\Lambda_a = \text{blockdiag}\{0_2, \Lambda_{a,s}\}
\]

where \( \Lambda_{a,s} \) contains the \( c_a - 2 \) non-zero eigenvalues. Let \( U_a = [U_{a,n} : U_{a,s}] \) be the partition of \( U_a \) into the eigenvectors corresponding to the zero and non-zero eigenvectors respectively. For simplicity, we will assume a normal approximation and take \( y_{i,j} = \log(d_{i,j}/e_{i,j}) \) to be independent \( \mathcal{N}(0, \sigma^2) \) variables. Let \( y = (y_{1,j}, \ldots, y_{n_a,j})^T \) for some year \( j \). Then the model \( y = B_a \theta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I) \) with penalty matrix \( \lambda D_a^T D_a \) can be transformed into a standard form mixed model \( y = X_a \beta + Z_a \alpha + \epsilon \) where

\[
X_a = [1_a : x_a], \ Z_a = B_a U_{a,s} \Lambda_{a,s}^{-1/2}, \ \alpha \sim \mathcal{N}(0, \tau^2 I_{c_a-2}).
\]

We turn now to the 2-dimensional case and apply exactly the same method as in one dimension. We seek the SVD, \( U \Lambda U^T \), of the 2-dimensional penalty \( P(\lambda_a, \lambda_y) \) in (2). This can be expressed in terms of the eigenvectors and eigenvalues of the 1-dimensional penalties \( \lambda_a D_a^T D_a \) and \( \lambda_y D_y^T D_y \). We define \( U \) as

\[
[U_{y,n} \otimes U_{a,n} : U_{y,n} \otimes U_{a,s} : U_{y,s} \otimes U_{a,n} : U_{y,s} \otimes U_{a,s}]
\]

and \( \Lambda \) as the diagonal matrix

\[
\text{blockdiag}\{0_4, I_2 \otimes \lambda_a \Lambda_{a,s}, \lambda_y \Lambda_{y,s} \otimes I_2, I_{c_y-2} \otimes \lambda_a \Lambda_{a,s} + \lambda_y \Lambda_{y,s} \otimes I_{c_a-2}\}.
\]
It is straightforward to check that $U\Lambda U^T$ is indeed a singular value decomposition of the 2-dimensional penalty $P(\lambda_a, \lambda_y)$ in (2).

Lee et al. (2013) now come up with a clever, yet simple, idea. They make the following replacements in $\Lambda$

$$I_2 \otimes \lambda_a \Lambda_{a,s} \rightarrow \text{blockdiag}\{\lambda_1 \Lambda_{a,s}, \lambda_2 \Lambda_{a,s}\}$$

$$\lambda_y \Lambda_{y,s} \otimes I_2 \rightarrow \text{blockdiag}\{\lambda_3 \Lambda_{y,s}, \lambda_4 \Lambda_{y,s}\}$$

and there are five smoothing parameters instead of two. We now apply the SVD transformation as in the 1-dimensional case with $U$ defined in (15) but with $\Lambda$ defined in (17). We obtain a standard mixed model $y = X\beta + Z\alpha + \epsilon$

with

$$X = \begin{bmatrix} 1_y \otimes 1_a : 1_y \otimes x_a : x_y \otimes 1_a : x_y \otimes x_a \end{bmatrix}$$

$$Z = \begin{bmatrix} Z_1 : Z_2 : Z_3 : Z_4 : Z_5 \end{bmatrix}$$

$$= \begin{bmatrix} 1_y \otimes Z_a : Z_y \otimes 1_a : x_y \otimes Z_a : Z_y \otimes x_a : Z_y \otimes Z_a \end{bmatrix}.$$ (19)

Let $\alpha_i$ be the random effect corresponding to $Z_i$, $i = 1, \ldots, 5$. We take $\alpha_i \sim N(0, \tau_i^2 I)$ and the model may be fitted with standard software such as the lme function in the R package nlme.

The original 2-dimensional model defined in (1) and (2) fits a general surface. The present model has the advantage that it decomposes the fit into six components and each component has an interpretation and associated degrees of freedom; this is the PS-ANOVA of Lee et al. (2013). The results for female mortality for ages 50 to 90 and years 1975 to 2015 for data from the UK’s Office for National Statistics are given in Table 1.

<table>
<thead>
<tr>
<th>Term</th>
<th>Interpretation</th>
<th>Degrees of Freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>Bilinear surface in age and year</td>
<td>4</td>
</tr>
<tr>
<td>$Z_1$</td>
<td>Smooth in age</td>
<td>5.4</td>
</tr>
<tr>
<td>$Z_2$</td>
<td>Smooth in year</td>
<td>7.8</td>
</tr>
<tr>
<td>$Z_3$</td>
<td>Interaction: smooth in age, linear in year</td>
<td>4.1</td>
</tr>
<tr>
<td>$Z_4$</td>
<td>Interaction: linear in age, smooth in year</td>
<td>5.2</td>
</tr>
<tr>
<td>$Z_5$</td>
<td>Interaction: smooth in age, smooth in year</td>
<td>27.4</td>
</tr>
</tbody>
</table>

| $[X : Z]$ | Total | 54.0 |

The six components of mortality just described lend themselves naturally to a graphical display. Here the contribution of each component to the final fitted mortality is given a graphical interpretation which corresponds to the decomposition of the degrees of freedom of PS-ANOVA. Some examples of this graphical decomposition can be seen at
From the GLAM perspective, the mixed model defined in (18) and (19) is an additive GLAM with six GLAM style components. Problems with storage or computational time for this model do not arise with current computers and \texttt{lme} completes in effect instantly. For a larger problem the GLAM algorithms might help. The normal model can be replaced with the original Poisson assumption in which case GLAM might well be useful; Lee et al. (2013) is an essential reference here. The PS-ANOVA model can be extended to three dimensions by following the template laid out above; in this case GLAM is likely to be needed.

5 GLAM: The IWSM Story

María went to the IWSM in New Orleans in 1998. On her return she had no difficulty in convincing me that I should go to this conference. The arguments were simple and persuasive: a beautiful and interesting location, a small and friendly group of delegates, and an atmosphere that encouraged scientific discussion. My first IWSM was in Graz the following year. Of course, Paul was already a seasoned IWSM man.

The 2000 conference in Bilbao was key. I wanted to meet the man who, along with Brian Marx, had invented \( P \)-splines. There was something about the method that seemed right to me: it was just regression with a twist. I met Paul at the dinner. I remember it well: I was interested in smoothing and forecasting 2-dimensional mortality data; Paul was interested in extending \( P \)-splines to two dimensions. Our first paper on smoothing such data was presented at the IWSM in Crete (Durbán et al., 2002); we added forecasting at the Leuven conference the next year (Currie et al., 2003). These two papers were combined in a \textit{Statistical Modelling} paper (Currie et al., 2004).

We had already started work on array methods. Paul made the initial breakthrough with algebraic expressions for the linear and inner product functions in two dimensions (Eilers et al. 2006). The problem was how to extend this work to three and higher dimensions. It was only when the work was cast in array form that the full beauty of the GLAM method was revealed. It was a time of frenzied effort and sleepless nights but by the time of the IWSM in Florence in 2004 we had broken the back of the problem. María produced the mixed model formulation and the complete work appeared in our RSS B paper of 2006. Since then the focus has been on applications and increasingly the mixed model representation is used where the detailed insight of analyses is provided by PS-ANOVA.

The IWSM continues to play a key role in the development of GLAM in particular and \( P \)-splines more generally. The annual meetings provide the ideal forum for presenting work in progress, discussing future projects, receiving feedback and even recruiting new followers to the cause. GLAM
and $P$-splines may not be a religion but they certainly have their fanatics and for this we have the IWSM to thank!

**Acknowledgments:** It is a great pleasure to acknowledge the work of Paul Eilers and María Durbán. Our work together on GLAM was the most exciting time of my career. Thank you both!

**References**


Applied Statistics for Covariance Functions

John Aston

1 University of Cambridge, UK

E-mail for correspondence: j.aston@statslab.cam.ac.uk

Covariance Functions, the equivalent of covariance matrices for functional data, might not the first data object that usually comes to mind. However, in a number of applications, from neuroimaging to linguistics, they are the underlying statistical object of interest. Their analysis is not completely straight-forward though, as they lie in non-Euclidean spaces, which makes even simple statistical analysis somewhat more complex. In this talk, we will demonstrate how we can use different approaches from statistics and applied mathematics to carry out statistical analysis of covariance functions. We will look at covariances which change over space, and also covariances which only result from observations of an inverse problem. The talk will be illustrated with examples from the study of dialects and languages and from the investigation of brain connectivity. [Joint work with John Coleman, Eardi Lila, Davide Pigoli and Shahin Tavakoli]

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
Statistical boosting for Markov-switching distributional regression models

Timo Adam\textsuperscript{1}, Andreas Mayr\textsuperscript{2}, Thomas Kneib\textsuperscript{3}, Roland Langrock\textsuperscript{1}

\textsuperscript{1} Bielefeld University, Germany
\textsuperscript{2} LMU Munich, Germany
\textsuperscript{3} University of Göttingen, Germany

E-mail for correspondence: timo.adam@uni-bielefeld.de

\textbf{Abstract:} Markov-switching generalized additive models for location, scale and shape constitute a novel class of latent-state time series regression models that allow to model multiple state-dependent parameters of the response distribution — not only the mean, but also variance, skewness and kurtosis parameters — as smooth functions of some set of covariates. We introduce an estimation approach that is based on the EM algorithm, where gradient boosting is used to prevent overfitting and perform variable selection. The suggested approach is illustrated by modeling the conditional distribution of daily energy prices in Spain over time.

\textbf{Keywords:} Hidden Markov models; Gradient boosting; EM algorithm.

1 Introduction

Markov-switching models (MSMs; Goldfeld and Quandt, 1973; Hamilton, 1989) are popular tools for time series analyses in which the data follow certain patterns within some time periods but reveal different stochastic properties during other periods. While conventional MSMs are restricted to modeling the expected value of the response distribution (treating the remaining parameters as fixed), the idea of Markov-switching generalized additive models for location, scale and shape (MS-GAMLSS) is to model each parameter by its own state-dependent predictor (Rigby and Stasinopoulos, 2005; Langrock \textit{et al.}, 2018).

A basic MS-GAMLSS comprises an observed state-dependent process \( \{Y_t\}_{t=1}^T \) which is driven by a hidden state process \( \{S_t\}_{t=1}^T \). The state process is commonly assumed to be an \( N \)-state Markov chain, with transition

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{ij})$, where $\gamma_{ij} = \Pr(S_{t+1} = j \mid S_t = i)$, $i, j = 1, \ldots, N$, and initial distribution $\mathbf{\delta} = (\delta_1, \ldots, \delta_N)$, where $\delta_i = \Pr(S_1 = i), i = 1, \ldots, N$. The distribution of $Y_t$ depends on the underlying state at time $t$, $s_t$, and is characterized by some probability (or density) function

$$f_Y(y_t; \theta_{1t}^{(s_t)}, \theta_{2t}^{(s_t)}, \theta_{3t}^{(s_t)}, \theta_{4t}^{(s_t)}) = f_Y(y_t; \theta_{1t}, \theta_{2t}, \theta_{3t}, \theta_{4t}).$$

(1)

Here $\theta_{1t}^{(s_t)}$ typically is the conditional mean of $Y_t$ (given $s_t$). The remaining parameters in (1) may relate (but are not limited) to the conditional variance, skewness or kurtosis. To account for possible parameter constraints, we consider a set of link functions $g_k, k = 1, \ldots, 4$, which map the constrained parameters onto the corresponding real-valued predictors $\eta_k^{(s_t)}$, which in turn depend on the covariate vectors $x_t = (x_{jt}), j = 1, \ldots, P$, where $P$ denotes the number of covariates included in the model. The dependence structure of an MS-GAMLSS is illustrated in Figure 1.

2 The MS-gamboostLSS algorithm

Expressing the state sequence in terms of the indicator variables $u_i(t) = 1_{\{S_t=i\}}$ and $v_{ij}(t) = 1_{\{S_{t-1}=i,S_t=j\}}$, the complete-data log-likelihood (i.e. the joint log-likelihood of the observations and the states) is

$$\text{CDLL} = \sum_{i=1}^{N} u_i(1) \log(\delta_i) + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=2}^{T} v_{ij}(t) \log(\gamma_{ij})$$

dependent on $\delta_i$

$$+ \sum_{i=1}^{N} \sum_{t=1}^{T} u_i(t) \log \left( f_Y(y_t; \theta_{1t}^{(s_t)} ) \right).$$

dependent on $\eta_1^{(s_t)}, \eta_2^{(s_t)}, \eta_3^{(s_t)}$ and $\eta_4^{(s_t)}$

(2)

The MS-gamboostLSS algorithm alternates between an expectation (E-) step, which involves calculating the conditional expectations of the $u_i(t)$'s
and \( v_{ij}(t) \)'s given the current parameter estimates, and a maximization (M-) step, which involves the maximization of (2) with respect to the different parameters, until convergence.

- **E step:** Using the so-called forward and backward probabilities, \( \alpha_t(i) = f_y(y_1, \ldots, y_t, S_t = i|x_1, \ldots, x_t) \) and \( \beta_t(i) = f(y_{t+1}, \ldots, y_T|S_{t+1} = i, x_{t+1}, \ldots, x_T) \), respectively, the conditional expectations of \( u_i(t) \) and \( v_{ij}(t) \) are

\[
\hat{u}_i(t) = \frac{\alpha_t(i) \beta_t(i)}{\sum_{k=1}^{N} \alpha_T(k)},
\]

\( i = 1, \ldots, N \), and

\[
\hat{v}_{ij}(t) = \frac{\alpha_{t-1}(i) \hat{\gamma}_{ij} f_Y(y_t; \theta^{(j)}_t) \beta_t(j)}{\sum_{j=1}^{N} \alpha_T(j)},
\]

\( i, j = 1, \ldots, N \), respectively. The forward-backward algorithm is applied to efficiently evaluate \( \alpha_t(i) \) and \( \beta_t(i) \) for \( t = 1, \ldots, T \) and \( i = 1, \ldots, N \) (Zucchini et al., 2016).

- **M step:** Using (3) and (4), the first two summands in (2) can be maximized with respect to \( \delta_i \) and \( \gamma_{ij} \) analytically, which leads to

\[
\hat{\delta}_i = \hat{u}_i(1),
\]

\( i = 1, \ldots, N \), and

\[
\hat{\gamma}_{ij} = \frac{\sum_{t=2}^{T} \hat{v}_{ij}(t)}{\sum_{k=1}^{N} \sum_{t=2}^{T} \hat{v}_{ik}(t)},
\]

\( i, j = 1, \ldots, N \), respectively. The third summand in (2) reduces to the weighted log-likelihood of a conventional GAMLSS, which can
be maximized with respect to $\eta_{k}^{(i)}$ using the gamboostLSS algorithm (Mayr et al., 2012; Thomas et al., 2017). The gamboostLSS algorithm iteratively fits a set of base-learners (e.g. penalized B-splines; Eilers and Marx, 1996) to the partial derivatives of (2) with respect to the different predictors. In each iteration, only the base-learner that relates to the best-performing covariate $j^*$ and parameter $k^*$ is selected to update the $k^*$-th predictor, where some step length $0 < sl < 1$ is used to prevent overfitting (in most cases, $sl = 0.1$ provides a convenient choice). As only the most important covariates are selected in the first iterations, early stopping leads to automated variable selection. The stopping iterations are typically chosen by $K$-fold cross validation over some grid of possible values, where the stopping iterations which correspond to the highest average out-of-sample log-likelihood are chosen.

The MS-gamboostLSS algorithm is illustrated in Figure 2.

3 Spanish energy prices

To illustrate the suggested approach, we model the conditional distribution of daily energy prices in Spain, $Y_t$, given the oil price, $x_{1t}$, over time. The data (Sanchez-Espigares and Lopez-Moreno, 2014) comprise 1761 daily observations between February 1, 2002, and October 31, 2008. Assuming a normal distribution with state-dependent probability density function

$$f_Y(y_t; \theta_{1t}^{(s_t)}, \theta_{2t}^{(s_t)}) = \frac{1}{\sqrt{2\pi\theta_{2t}^{(s_t)^2}}} \exp \left( -\frac{(y_t - \theta_{1t}^{(s_t)})^2}{2\theta_{2t}^{(s_t)^2}} \right),$$

we fitted a 2-state MS-GAMLSS using penalized B-splines as base-learners, where the stopping iterations were chosen by 20-fold cross validation over the grid $(25, 50, 100, \ldots, 3200) \times (25, 50, 100, \ldots, 3200)$, which led to the optimal values 1600 (state 1) and 200 (state 2). Using a 3.6 GHz Intel® Core™ i7 CPU, model fitting took about 10.9 minutes. The off-diagonal t.p.m. entries were estimated as $\hat{\gamma}_{12} = 0.020$ and $\hat{\gamma}_{21} = 0.018$, which indicates strong persistence within the states (according to the estimated transition probabilities, the average dwell-times within a state were about 50 days for state 1 and 56 days for state 2). Figure 3 displays the estimated state-dependent distributions and the observed time series overlaid with the locally decoded underlying states. In both states, the oil price exhibits a (mostly) positive effect on the mean, whereas the effect on the variance is much stronger in state 2. As both parameters are higher in state 2, it may be linked to a rather nervous market regime, whereas state 1 relates to calmer periods.
FIGURE 3. Estimated state-dependent distributions for states 1 (blue) and 2 (red) (left plot) and locally decoded time series (right plot). Dashed lines indicate the 0.05, 0.15, 0.25, 0.75, 0.85 and 0.95 quantiles of the fitted state-dependent distributions.

4 Discussion

The MS-gamboostLSS algorithm constitutes a novel estimation approach for MS-GAMLSS that combines the EM and the gamboostLSS algorithm. By specifying appropriate base-learners for the different parameters in (1), e.g. intercept-only terms or simple linear models, the suggested approach can also be used to fit a variety of models that are nested special cases, including simple hidden Markov models (HMMs; Zucchini et al., 2016) and Markov-switching (generalized) linear and additive models (Langrock et al., 2017). Therefore, the MS-gamboostLSS algorithm may provide a promising method for model fitting and variable selection not only in MS-GAMLSS but also in a variety of other HMM-type models.

References


New approaches to postprocessing of multi-model ensemble weather forecasts

Clair Barnes1, Richard E. Chandler1

1 University College London, UK

E-mail for correspondence: clair.barnes.16@ucl.ac.uk

Abstract: We propose a Bayesian framework to infer the posterior probability distribution of weather quantities of interest, given a collection of ensemble forecasts from different modelling centres. The resulting forecasts are found to be marginally and jointly well calibrated, with no residual regional bias.

Keywords: Weather forecasts; uncertainty quantification; ensemble forecasting; Bayesian inference.

1 Introduction

Most weather forecasts use numerical weather prediction models, propagating a ‘best guess’ at the initial weather state through a mathematical model of the atmosphere in order to predict the future state. The uncertainty about the forecast is assessed by initialising models with an ensemble of perturbed starting conditions; if the forecast ensemble dispersion is small (large), the uncertainty about the issued forecast is assumed to be small (large). If a forecast is to be useful for decision making, it must be well calibrated, giving an accurate statement of the forecast uncertainty. The direct output from these models tends to display systematic biases and dispersion errors, which can be corrected by statistical postprocessing, with the necessary adjustments estimated from past forecast errors. Many postprocessing schemes exist to postprocess single ensemble forecasts, although many - such as Model Output Statistics (Gneiting et al. 2005) - apply only to univariate forecasts. Moreover, a single-ensemble forecast can only account (to some extent) for uncertainty about the initial conditions: uncertainty about the ‘correct’ choice of model may also contribute a significant amount of uncertainty to the forecast. Multi-model
ensembles (MMEs), consisting of ensemble output from several models, allow us to sample some of this model uncertainty, and so to obtain a more complete understanding of the uncertainty about the final forecast. A key challenge in MME postprocessing is to acknowledge the fact that the generating models are often more similar to one another than to reality: if, for example, one ensemble predicts too low a temperature, it is likely that other ensembles will display a similar tendency.

We present a Bayesian framework that explicitly allows for relationships between the component ensembles, allowing us to make inferences about the distribution of the ‘true’ weather state. The method is developed from that presented in Chandler (2013), where the focus was on inferring statistical properties of the distribution of future climate; here, we wish to make predictions directly about observable quantities. The multivariate-normal specification applies directly to forecasts of surface temperatures and wind vectors, and to other continuous weather variables after normalisation.

2 A new postprocessing framework

For a given forecast instance, we have a vector of weather quantities of interest, \( \mathbf{Y} = \{Y_1, \ldots, Y_p\} \), with \( p \) the number of variables we wish to forecast. The observed value of \( \mathbf{Y} \) is denoted \( \mathbf{Y}_0 \). For each \( i = 1, \ldots, m \), the \( i \)th ensemble provides a set of \( n_i \) forecasts of \( \mathbf{Y} \), with the \( j \)th forecast from the \( i \)th ensemble being labelled \( Y_{ij} \). The members of the \( i \)th ensemble forecast are assumed to be iid conditional on the ensemble’s population mean \( \mu_i \):

\[
Y_{ij} | \mu_i \sim MVN (\mu_i, C_i) \tag{1}
\]

Reflecting the fact that individual ensembles may often be more similar to one another than they are to reality, the ensemble mean forecasts are assumed to be iid around a mutual consensus, \( \xi \), which can be decomposed into the ‘true’ value, \( \mathbf{Y}_0 \), plus a shared discrepancy \( \Delta \).

\[
\mu_i | \xi \sim MVN (\mathbf{Y}_0 + \Delta, \Sigma) \tag{2}
\]

\[
\Delta \sim MVN (\eta, \Lambda) \tag{3}
\]

This conditional independence structure is encoded in the arrows in the directed acyclic graph shown in Figure 1. The distribution of \( \Delta \) can be estimated for any given forecast instance from an appropriate training set of past forecast-observation pairs. The Bayesian specification is completed by setting a prior distribution on the realised weather quantity, \( \mathbf{Y}_0 \):

\[
\mathbf{Y}_0 \sim MVN (\alpha, \Gamma) \tag{4}
\]

The conditional independence structure allows us to derive an expression for the postprocessed (posterior) forecast, conditioned on the ensemble fore-
Bayesian postprocessing of multi-ensemble weather forecasts

\[ \Delta \xi = Y_0 + \Delta Y_0 \mu_1 \mu_2 \mu_3 Y_1 Y_2 Y_3 \]

**FIGURE 1.** Schematic diagram of relationships between the elements of the multi-ensemble system. Quantities known at the time of forecasting are shown as filled nodes, with unknown quantities represented by open nodes.

casts. Defining \( \overline{Y}_i = n_i^{-1} \sum_j Y_{ij} \) and \( D_i = \Sigma + n_i^{-1} C_i \), we have

\[
Y_0 | Y_{ij} \sim MVN (\tau, S) \tag{5}
\]

\[
S^{-1} = \Gamma^{-1} + \left[ \Lambda + \left( \sum_{i=1}^{m} D_i^{-1} \right)^{-1} \right]^{-1} \tag{6}
\]

\[
\tau = S \left[ \Gamma^{-1} (\alpha + \eta) + \left( I + \sum_{i=1}^{m} D_i^{-1} \Lambda \right)^{-1} \sum_{i=1}^{m} D_i^{-1} \overline{Y}_i \right] - \eta \tag{7}
\]

3 Case study

The presented study considers forecasts of winter temperatures in 13 regions across the UK, at leadtimes of up to two weeks. Ensemble forecasts from the ECMWF, NCEP and UKMO were obtained from the TIGGE archive (Bougeault et al. 2010), along with ERA-Interim reanalyses (Dee et al. 2011) used to verify the forecasts.

Marginal calibration is assessed via Probability Integral Transform (PIT) histograms (Gneiting et al. 2008), with joint calibration assessed through modified Band Depth Rank (BDR) histograms (Thorarinsdottir et al. 2016); well calibrated forecasts produce uniform PIT and BDR histograms. Performance is compared to that of MOS forecasts and reference ‘superensemble’ forecasts, created by pooling the uncorrected ensemble members and defining a forecast density with the pooled mean and covariance.

Figure 2 shows typical examples of these histograms for the postprocessed forecasts. Unlike the MOS and superensemble forecasts, those produced by the Bayesian framework are found to be both marginally and jointly well calibrated at all leadtimes, displaying no evidence of systematic regional biases and with PIT and BDR histograms having minimal deviation from uniformity.
FIGURE 2. Histograms showing calibration of post-processed 5-day-ahead forecasts of surface temperatures: (a) PIT of marginal forecast for Bristol; (b) PIT of marginal forecast for Glasgow; (c) BDR of joint forecast across all sites.

Forecast discrepancies for each instance were initially estimated from the forecast errors in the 25 days prior to the forecast issue date. The use of training data consisting of ‘analogues’, selected on the basis of their initial-condition similarity to the forecast instance of interest, produces a small additional improvement in both marginal and joint calibration.

References


Dynamic Chain Graph Models for Ordinal Time Series Data

Pariya Behrouzi\textsuperscript{1}, Fentaw Abegaz\textsuperscript{2}, Ernst C. Wit\textsuperscript{3}

\textsuperscript{1} Biometris, Wageningen University and Research Center, Netherlands
\textsuperscript{2} Medical Genomics, University of Liege, Belgium
\textsuperscript{3} Statistics and probability, University of Groningen, Netherlands

E-mail for correspondence: pariya.behrouzi@wur.nl

Abstract: This paper introduces sparse dynamic chain graph models for network inference in high dimensional non-Gaussian time series data. The proposed method is parametrized by a precision matrix that encodes intra time-slice conditional independences among variables at a fixed time point, and an autoregressive coefficient that contains dynamic conditional independence interactions among time series components across consecutive time steps. The proposed model is a Gaussian copula vector autoregressive model. Estimation is achieved via a penalized EM algorithm where we use coordinate descent algorithm to optimize the penalized log-likelihood with the smoothly clipped absolute deviation penalty. We demonstrate our approach on simulated and psychological datasets.

Keywords: Chain graph models; time-series data; Vector autoregressive model

0.1 Dynamic chain graph models

A chain graph is defined as $G = (V, E)$ where $V$ is a set of nodes and $E$ is a set of ordered and unordered pairs of nodes, called edges, which contains the directed and undirected interactions. A dynamic chain graph model is associated with a time series chain graph model, where the dependence structure of the time series components can be divided into two sets: \textit{intra time-slice dependencies}, represented by undirected edges that specify the association among variables in a fixed time step, and \textit{inter time-slice dependencies}, represented by associations among variables across consecutive time steps. Links across time steps are directed, pointing from a set of nodes at a previous time step, $V_{(t-1)}$, to nodes at the current time step, $V_t$.

Let $Y(t) = (Y_1(t), \ldots, Y_p(t))^\top$, $t = 1, \ldots, T$ be an $p$-dimensional time series vector representation of $p$ variables that have been studied longitudinally

---

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
across $T$ time points. Each time series component $Y(t)$ is assumed to be sampled $n$ times. Thus, $Y_{ij}(t)$ represents the value of the $j$-th variable at time $t$ for the $i$-th sample, $i = 1, \ldots, n, j = 1, \ldots, p$. Here, we focus on non-Gaussian multivariate time series data such as ordinal-valued time series, taking values in $\{0, 1, \ldots, (c_k - 1)\}$, where $c_k$ is the number of possible categories, or mixed categorical-and-continuous time series data, as routinely occurring in real world settings.

1 Model definition

We assume a stable dynamic chain graph model, meaning that the structure of interactions within each time point remains stable for previous and current time steps, and interactions between consecutive time steps are also stable. We use a vector autoregressive process of order 1, VAR(1),

$$Z_t = \Gamma Z_{(t-1)} + \epsilon_t$$

(1)

to describe the directed latent interactions, where $\epsilon_t \sim N(0, \Theta^{-1})$ describes the undirected instantaneous interactions. We use the rank likelihood $f_Y(\Theta, \Gamma)$ as

$$n \sum_{i=1}^{T} \sum_{t=2}^{T} \log f(z_i^{(t)} \in A(y_i^{(t)}) | z_i^{(t-1)} \in A(y_i^{(t-1)}); \Theta, \Gamma) + \log f(z_i^{(1)} \in A(y_i^{(1)}) | \Theta, \Gamma)$$

where we ignore the second term and start from $t = 2$. We compute the conditional log-likelihood using the conditional distribution $f(z^{(t)} | z^{(t-1)})$.

According to (1) the conditional distribution $Z^{(t)} \mid Z^{(t-1)}$ follows

$$Z^{(t)} \mid Z^{(t-1)} = z^{(t-1)} \sim N(\Gamma z^{(t-1)}, \Theta^{-1})$$

(2)

whose density for $t$-th observation is defined as

$$(2\pi)^{p/2} \det(\Theta)^{1/2} \exp \left[ \frac{1}{2} \left( z^{(t)} - \Gamma z^{(t-1)} \right)' \Theta \left( z^{(t)} - \Gamma z^{(t-1)} \right) \right].$$

1.1 Penalized EM inference

In Gaussian copula, we treat marginal distributions as nuisance parameters since our main goal is to learn the dependence structure among time series components both at a fixed time step $t \in \mathbb{N}$ and also across consecutive time steps. We use EM algorithm, where the E-step can be written as

$$E(S_{\Gamma} \mid y_i; \Theta^*, \Gamma^*) = \frac{1}{n(T-1)} \left[ S_{cc} - S_{cp} \Gamma' - \Gamma S_{cp}' + \Gamma S_{pp} \Gamma' \right]$$

(3)

such that the conditional expectations inside (3) are defined as follow

$$S_{cc} = \sum_{i=1}^{n} \sum_{t=2}^{T} E[Z_i^{(t)}Z_i^{(t)'} | y_i; \Theta^*, \Gamma^*], \quad S_{pp} = \sum_{i=1}^{n} \sum_{t=1}^{T-1} E[Z_i^{(t)}Z_i^{(t)'} | y_i; \Theta^*, \Gamma^*]$$
TABLE 1. Comparison the two methods over 100 independent run, where t=5.

<table>
<thead>
<tr>
<th>Fixed at t=5</th>
<th>Performance Θ</th>
<th>Performance Γ</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=10 &amp; n=50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tsnetwork</td>
<td>0.37 0.37 0.85</td>
<td>0.44 0.43 0.7</td>
</tr>
<tr>
<td>SparseTSCGM</td>
<td>0.33 0.45 0.80</td>
<td>0.42 0.65 0.34</td>
</tr>
<tr>
<td>p=50 &amp; n=50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tsnetwork</td>
<td>0.13 0.08 1.00</td>
<td>0.32 0.24 0.95</td>
</tr>
<tr>
<td>SparseTSCGM</td>
<td>0.03 0.03 0.97</td>
<td>0.33 0.55 0.82</td>
</tr>
</tbody>
</table>

\[ S_{pc} = \sum_{i=1}^{n} \sum_{t=2}^{T} E_Z[Z_i^{(t-1)}Z_i^{(t)}|y_i; \Theta^*, \Gamma^*]. \]

We use the first and the second moments of the truncated normal distribution to approximate these conditional expectations. The M-step of the EM algorithm contains a two-stage optimization process where under the penalized log-likelihood with the SCAD penalty we optimize in the first stage

\[ \Theta^{(k)}_{\lambda} = \arg \max_{\Theta} \left\{ \log \det(\Theta) - \text{tr}(S^{(E)}_{\Gamma, \Theta}) - \sum_{j \neq j'} w_{jj'} |\theta_{jj'}| \right\}, \]

for previous \( \Gamma^* \). This optimization can be solved efficiently using the graphical lasso algorithm proposed by Fridman et al., (2008). We proceed in the second-stage to update the estimate of \( \Gamma \) given the updated \( \Theta \), where we optimize

\[ \Gamma^{(k)}_{\rho} = \arg \max_{\Gamma} \left\{ \log \det(\Theta^{(k)}_{\lambda}) - \text{tr}(S_{cc}^{(k)} \Theta^{(k)}_{\lambda}) - S_{cp} \Gamma' \Theta^{(k)}_{\lambda} - \right\}
\]

\[ \Gamma S_{cp}^{'} \Theta^{(k)}_{\lambda} + \Gamma S_{pp}^{'} \Theta^{(k)}_{\lambda} - \sum_{j,l} \nu_{jl} |\gamma_{jl}| \}

We use a direct coordinate descent algorithm (Fan et al., 2001) to calculate \( \Gamma^{(k)}_{\rho} \). Given the two-stage optimization problem inside the M-step, we update the \( S_{\Gamma} \) matrix in the E-step. This iterative procedure continues until to meet a convergence criteria. To determine the sparsity of contemporaneous and delayed interactions, we employ the Bayesian information criteria to tune \( \lambda \) and \( \rho \).

2 Data analysis

2.1 Simulations

We set up a simulation to generate sparse \( \Theta \) and \( \Gamma \) matrices. We compare the performance of our proposed method with SparseTSCGM (Abegaz and
FIGURE 1. Contemporaneous interactions (Left) and delayed interactions (Right) in NESDA data with five categories: (i) sleep, blue, (ii) mood, green, (iii) appetite, yellow, (iv) somatic, gray, (v) mental, red.

Wit (2013). The results of comparisons are provided in Table 1. We note that high values of $F_1$-scores indicate good performance. These results suggest that although recovering sparse network structure in ordinal time series data is a challenging task, the proposed approach performs well on model-based simulations.

2.2 Netherlands Study of Depression and Anxiety

We applied our method to the Netherlands Study of Depression and Anxiety (NESDA) to investigate the course of depression and anxiety disorders over a period of three years. The data consist of 1799 participants and 28 items. For each item there are four corresponding answers 0=None, 1=Mild, 2=Moderate, 3=Severe. Figure 1(Left) shows undirected links that suggest contemporaneous interactions among 12 items, and Figure 1(Right) displays directed edges that indicate Granger-causality relationships. We observe that item Feeling sad is the hub in both figures, suggesting that it plays a fundamental role in treating depression and anxiety disorders. Also, Figure 1(Right) shows several directed links pointing from mood category to mental category; this suggests that mood disorders influence the long-term development of mental disorders.

References


Modeling Ratio Outcomes with Gamma Distributed Components

Moritz Berger\textsuperscript{1}, Michael Wagner\textsuperscript{2}, Matthias Schmid\textsuperscript{1}

\textsuperscript{1} University of Bonn, Germany
\textsuperscript{2} German Center for Neurodegenerative Diseases, Bonn, Germany

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

Abstract: A regression model termed “extended GB2 model” is proposed, which is designed to analyze ratios of correlated gamma distributed random variables. Unlike regression modeling with a log-transformed response, the extended GB2 model directly links the expectation of the untransformed ratios to a set of covariates. We present a gradient boosting algorithm for maximum likelihood estimation of the model parameters and provide details on hypothesis testing. An application from the field of dementia research demonstrates the usefulness of the proposed modeling approach.

Keywords: Gamma distribution; Generalized beta distribution of the second kind; Generalized linear models; Ratio of random variables.

1 Introduction

In observational studies one frequently encounters the ratio of two positively correlated variables. Important examples are the LDL/HDL cholesterol ratio in cardiovascular research, the CD4/CD8 ratio in HIV research and the amyloid-\(\beta\) 42/40 ratio in dementia research. The latter will be considered in the application.

A common approach to model ratio outcomes is to apply a logarithmic transformation and to fit a regression model with log-transformed outcome. A well-known limitation of log-ratio analysis is that inference is only possible for the expectation on the log-transformed scale. This affects the interpretability of the predictor-response relationships and makes it difficult to draw conclusions from the results of associated hypothesis tests. An alternative approach is to model right-skewed positive data by gamma distributed random variables. While gamma regression for positive random...
variables is widely used in practice, no regression modeling strategy regarding the ratio of two correlated gamma distributed variables exists yet. To address these issues, we propose a regression model, which relates the expectation of the untransformed ratio directly to a set of covariates and also accounts for the correlation between the two variables.

2 Ratio of Two Correlated Gamma Variables

Let $U$ and $V$ be two gamma distributed random variables with probability density functions

$$f_U(u) = \frac{\lambda_u^\alpha}{\Gamma(\alpha)} u^{\alpha-1} \exp(-\lambda_u u), \quad (1)$$

$$f_V(v) = \frac{\lambda_v^\alpha}{\Gamma(\alpha)} v^{\alpha-1} \exp(-\lambda_v v), \quad (2)$$

where $\alpha > 0$ denotes a common shape parameter and $\lambda_u, \lambda_v > 0$ are individual rate parameters. The assumption of a common shape parameter ensures that the two density functions share the same basic form. To derive a distribution for the ratio $U/V$, we assume that the pair $(U, V)$ follows a bivariate gamma distribution with probability density function (p.d.f.)

$$f_{U,V}(u, v) = \frac{\lambda_u \lambda_v}{(1 - \rho) \Gamma(\alpha)} \left(\frac{uv}{\rho \lambda_u \lambda_v}\right)^{\alpha - 1} \times \exp\left(-\frac{\lambda_u u + \lambda_v v}{1 - \rho}\right) I_{\alpha - 1}\left(\frac{2\sqrt{\rho \lambda_u \lambda_v uv}}{1 - \rho}\right), \quad (3)$$

where $0 < \rho < 1$ and $I_{\alpha - 1}(\cdot)$ is the modified Bessel function of the first kind of order $\alpha - 1$. The p.d.f. defined in (3) was introduced by Kibble (1941) and is known as “Kibble’s bivariate gamma distribution” or “Kibble-Wicksell distribution”. It can be shown that the additional parameter $\rho$ equals the Pearson correlation coefficient of $U$ and $V$.

Let the joint distribution of $(U, V)$ be defined by the p.d.f. in (3). Using the results of Nadarajah and Kotz (2007) and Weinhold et al. (2016), it can be derived that the p.d.f. of the random variable $R := U/V$ is given by

$$f_R(r; \alpha, \rho, \theta) = \frac{\Gamma(2\alpha)}{\Gamma^2(\alpha)} \theta^{-\alpha} (1 - \rho)^\alpha$$

$$\times \left(\frac{1-\theta}{\theta} \frac{r}{1+r} + 1\right) \left(\frac{r^{\alpha-1}}{(1+r)^{2\alpha}}\right) \left(\left(\frac{1-\theta}{\theta} \frac{r}{1+r} + 1\right)^2 - 4 \rho \theta \frac{r}{(1+r)^2}\right)^{\alpha+0.5}, \quad (4)$$

where $\theta := \lambda_v / \lambda_u = E[U] / E[V]$ denotes the ratio of the two rate parameters. Importantly, for $\rho = 0$ the p.d.f. in (4) reduces to the p.d.f. of
FIGURE 1. Examples of the p.d.f. of the ratio $R = U/V$ in (4). The curves visualize the distribution of $R$ for $\alpha \in \{2, 4\}$, $\rho \in \{0.2, 0.5, 0.8\}$ and $\theta = 10$.

Under the assumptions of (4), the expectation of the random variable $R$ is given by

$$E\left[\frac{U}{V}\right] = \frac{E[U]}{E[V]} \cdot \frac{\Gamma(\alpha + 1) \Gamma(\alpha - 1)}{\Gamma^2(\alpha)} \ 2F_1(-1, 1; \alpha; \rho)$$

and

$$= \theta \ \frac{\Gamma(\alpha + 1) \Gamma(\alpha - 1)}{\Gamma^2(\alpha)} \ 2F_1(-1, 1; \alpha; \rho)$$

$$= \theta C(\alpha, \rho), \ \alpha > 1,$$

where $C(\alpha, \rho) := \Gamma(\alpha + 1) \Gamma(\alpha - 1) / \Gamma^2(\alpha) \ 2F_1(-1, 1; \alpha; \rho)$ and $2F_1(\cdot)$ is the generalized hypergeometric function (Candan and Orguner, 2013).

### 3 Extended GB2 Model

By equation (5), the mean ratio $E[U/V]$ can be written as the product of the ratio of means $\theta = E[U]/E[V]$ and a factor $C(\alpha, \rho)$ that is a function of the parameters $\alpha$ and $\rho$. Therefore, we propose to relate $\theta$ to a set of $p$ covariates $X_1, \ldots, X_p$ by the model equation $\log(\theta|X) = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_p X_p$, where $\gamma := (\gamma_0, \gamma_1, \ldots, \gamma_p)^T$ is a set of real-valued coefficients. The logarithmic transformation projects the positive random variable $\theta|X$ to the set of real numbers, ensuring that there are no restrictions on the coefficients. Moreover, we propose to treat $\alpha$ and $\rho$ as nuisance parameters that do not depend on the covariates. Because of the proportionality between $\theta$ and $E[U/V]$, the proposed extended GB2 model has the form

$$\log \left( E[U/V|X] \right) = \tilde{\gamma}_0 + \gamma_1 X_1 + \ldots + \gamma_p X_p,$$

where $\tilde{\gamma}_0 := \gamma_0 + \log(C(\alpha, \rho))$. 
In particular, since (6) can be re-written as
\[ E \left[ \frac{U}{V} \right] = \exp(\tilde{\gamma_0}) \cdot \exp(\gamma_1X_1) \cdot \ldots \cdot \exp(\gamma_pX_p), \] (7)
the expressions \( \exp(\gamma_1), \ldots, \exp(\gamma_p) \) have a simple interpretation in terms of multiplicative increases/decreases of the expected ratio. For example, if \( \gamma_k > 0, k \in \{1, \ldots, p\} \), increasing \( X_k \) by one unit implies that \( E \left[ \frac{U}{V} \right] \) is increased by the factor \( \exp(\gamma_k) \).

4 Optimization of the Likelihood and Inference

Estimates of the parameters \( \gamma_0, \gamma_1, \ldots, \gamma_p \) are obtained by maximizing the log-likelihood of the extended GB2 model evaluated on a set of independent realizations \((r_i, x_{i1}, \ldots, x_{ip})^\top, \ i = 1, \ldots, n\). To maximize the log-likelihood function, which is directly derived from equations (4) and (6), we propose to apply a gradient boosting algorithm with component-wise linear base-learners (Bühlmann and Hothorn, 2007). Gradient boosting is a generic optimizer that can be used to minimize any positive real-valued objective function \( R((r_i, \psi(x_i))_{i=1,\ldots,n}) \) over an unknown real-valued “prediction function” \( \psi(\cdot) \), provided that the derivative \( \partial R/\partial \psi \) exists. Because we require the base-learners to be linear in the covariates \( X_1, \ldots, X_p \), the function space of \( \psi(\cdot) \) is implicitly restricted to the linear subspace defined by \( \psi(x_i) = x_i^\top \gamma \). Consequently, the estimation of \( \psi(\cdot) \) reduces to the estimation of the coefficient vector \( \gamma \). Maximum likelihood estimates of the parameters of the extended GB2 model can therefore be computed by setting \( R \) equal to the negative of the log-likelihood and by running the gradient boosting algorithm until convergence. Furthermore, the algorithm is modified such that it allows for the estimation of the additional nuisance parameters \( \alpha \) and \( \rho \).

Because of the asymptotic normality of the maximum likelihood estimators, statistical tests of the hypotheses \( "H_0: \gamma_k = 0 \) vs. \( H_1: \gamma_k \neq 0", k = 1, \ldots, p, \) are obtained by plugging the estimates in the observed information matrix
\[ J(\gamma, \alpha, \rho) := -\partial^2 l(\gamma, \alpha, \rho; r_1, \ldots, r_n, x_1, \ldots, x_n) / \partial \gamma \gamma^\top \]
and by calculating the test statistics
\[ Z_k = \frac{\hat{\gamma}_k}{\sqrt{J^{-1}_{kk}(\hat{\gamma}, \hat{\alpha}, \hat{\rho})}}, \ k \in \{1, \ldots, p\}, \] (8)
where \( J^{-1}_{kk} \) denotes the \( k \)-th diagonal element of \( J^{-1} \). For a given significance (type I error) level \( \alpha_I \), the null hypothesis \( "H_0: \gamma_k = 0" \) is rejected if \( |Z_k| > z_{1-\alpha_I/2} \), where \( z_{1-\alpha_I/2} \) is the \((1-\alpha_I/2)\)-quantile of the standard normal distribution. Accordingly, asymptotic \((1-\alpha_I)%\) confidence intervals are defined by \( \hat{\gamma}_k \pm z_{1-\alpha_I/2} \sqrt{J^{-1}_{kk}(\hat{\gamma}, \hat{\alpha}, \hat{\rho})} \).
TABLE 1. Analysis of the DCN baseline data. Coefficient estimates and p-values were obtained by fitting an extended GB2 model to the data of MCI patients. The maximum likelihood estimates of $\alpha$ and $\rho$ were 10.110 and 0.344, respectively.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\gamma}$</th>
<th>se($\hat{\gamma}$)</th>
<th>p-value</th>
<th>exp($\hat{\gamma}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>-0.0093</td>
<td>0.0025</td>
<td>0.0002</td>
<td>0.9907</td>
</tr>
<tr>
<td>No. of years of education</td>
<td>0.0001</td>
<td>0.0070</td>
<td>0.9885</td>
<td>1.0001</td>
</tr>
<tr>
<td>Gender (female)</td>
<td>-0.0599</td>
<td>0.0422</td>
<td>0.1558</td>
<td>0.9418</td>
</tr>
<tr>
<td>ApoE $\epsilon$4 carrier (yes)</td>
<td>-0.2034</td>
<td>0.0411</td>
<td>&lt;0.0001</td>
<td>0.8159</td>
</tr>
</tbody>
</table>

5 Amyloid Beta for Diagnosis of Alzheimer’s Disease

We analyzed data from a multi-center observational cohort study conducted by the German Dementia Competence Network (DCN, Kornhuber et al., 2009). The study included patients older than 50 years who sought evaluation at one of the participating university memory clinics. All diagnoses were made using clinical and neuropsychological assessments. Dementia-related diagnoses were either mild cognitive impairment (MCI), Alzheimer’s disease (AD), or other dementia.

A major challenge in the diagnosis and prognosis of AD is the decades-long period between disease onset and the first clinical symptoms of AD. This problem is further aggravated by the fact that not all patients passing through the MCI stage will eventually suffer from underlying AD pathology. Since the amyloid-\(\beta\) 42/40 ratio is considered to be a strong predictor of AD progression, it is of high interest to relate amyloid-\(\beta\) 42/40 measurements to dementia-related risk factors in MCI patients for early AD diagnosis and prediction. Here, we focused on the amyloid-\(\beta\) 42/40 ratios of the DCN Study participants at baseline, which were measured in 379 of the 1095 patients diagnosed with MCI. Exclusion of patients with missing values in any of the analyzed variables resulted in an analysis data set with \(n = 336\) observations. The unconditional Pearson correlation coefficient between the amyloid-\(\beta\) 42 and the amyloid-\(\beta\) 40 concentrations was 0.422.

The following covariates were considered for inclusion in the extended GB2 model: (i) gender, (ii) age in years, (iii) educational level (measured by the number of years of education), and (iv) a binary variable indicating whether a patient was a carrier of the apolipoprotein E $\epsilon4$ (ApoE $\epsilon4$) allele. Table 1 presents the results obtained from fitting the extended GB2 model to the DCN baseline data. According to the p-values shown in the third column of Table 1, the AD risk factors age and ApoE $\epsilon4$ had highly significant effects on the amyloid-\(\beta\) 42/40 ratios of the study participants ($p = 0.0002$ and $p < 0.0001$, respectively). Each year of age was estimated to reduce the expected amyloid-\(\beta\) 42/40 ratio by the factor $\exp(-0.0093) = 0.9907$, corresponding to a yearly decrease of approximately 1%. Expected amyloid-
The quantile residuals presented in Figure 2 indicate a very good fit of the extended GB2 model.

β 42/40 ratios of ApoE ε4 carriers were reduced by an estimated 18% compared to patients not carrying the allele (exp(−0.2034) = 0.8159). In contrast, educational level and gender did not have significant effects on the amyloid-β 42/40 ratio at the 5% type I error level.

References


Modelling Cryptocurrencies using Robust Time Varying Undirected Graphs

Mauro Bernardi¹², Paola Stolfi²

¹ Department of Statistical Sciences, University of Padova, Padova, Italy
² Istituto per le Applicazioni del Calcolo “Mauro Picone” - CNR, Roma, Italy

E-mail for correspondence: paola.stolfi@uniroma3.it

Abstract: During the last two years the interest toward cryptocurrencies arose exponentially as well as the number of people investing on them. This paper analyses the joint dynamic evolution of cryptocurrencies which is often characterised by the presence of extreme values that highly deteriorates signal extraction. Within a static framework, sparse and high-dimensional dependence structures are estimated by undirected graphs. Here, to address the non-stationary behaviour of time series data, we consider sparse time-varying undirected Gaussian graphs, thereby leading to estimates that are very sensitive to the presence of outlying observations. Therefore, we propose a robust estimator which minimises the $\gamma$-divergence and we investigate its rate of convergence. We also provide an algorithm for efficiently handing parameter estimation.

Keywords: Undirected graphs; time varying models; robust methods; sparsity.

1 Introduction

The first cryptocurrency, namely the “Bitcoin”, has been founded in 2009. At the end of 2016 the number of available cryptocurrencies was around six hundred while at the end of 2017 it was over one thousand and three hundred. These numbers witnesses how the interest toward cryptocurrencies arose during the last years. At this time, they are mainly used as investments. Indeed many online exchange website offer the opportunity to sell and buy all the available cryptocurrency and to create investment portfolios to manage the related financial risk. However, due to their nature, these currencies are far from being similar to the traditional ones, an aspect that suddenly emerges by inspecting the data. The evolution over time of cryptocurrencies is characterised by highly non-stationarity and
by the presence of many outliers. Moreover, excluding the most important cryptocurrencies, such as Bitcoin, most of them should be conditionally independently distributed. To powerfully describe their joint dependence behaviour we consider sparse undirected graph. Sparse methods refer to statistical approaches specifically tailored to deal with estimation of large dimensional models with potentially many more features than observations. Despite their practical relevance, most of the recent contributions have been confined under the restrictive assumption of independently and identically distributed Gaussian observations, with only few exceptions: Lafferty at al. (2012), Finegold and Drton (2014), Vogel and Tyler (2014), and Bernardi et al. (2017). However, they do not deal with the potentially time–varying nature of variance–covariance matrices that may originate from dependent data. This case has been recently investigated by Zhou et al. (2010), where they introduce sparse time–varying undirected graphs, namely, graphs whose structure evolves smoothly over time. These models works well in high dimensional settings, that is when lots of parameters need to be estimated and few observations are available. Unfortunately, the Gaussian assumption makes the resulting estimates quite sensitive to the presence of outliers. The contribution of this paper is to solve this issue by proposing a robustification of the sparse time–varying graphs proposed by Zhou et al. (2010). Specifically, following Hirose et al. (2017), we propose a robust estimator which minimises the $\gamma$–divergence between the postulated and the true unknown distribution. Furthermore, we provide an algorithm to handle parameter estimation and we investigate the rate of convergence of the estimator of the variance–covariance matrix and of its inverse.

2 Methodology

We consider the dynamic model

$$W_t = W_{t-1} + Z_t, \quad Z_t \sim N\left(0, \Theta_t^{-1}\right), \quad \text{for } t > 0$$

$$W_0 \sim N\left(0, \Theta_0^{-1}\right).$$

Following Zhou et al. (2010), the Lasso–penalised maximum likelihood of the precision matrix $\Theta$ in the non i.i.d. case is given at time $t$ by

$$\hat{\Theta}_n (t) = \arg \min_{\Theta > 0} \left\{ \text{tr} \left( \Theta \hat{S}_n (t) \right) - \ln |\Theta| + \lambda \|\Theta\|_1 \right\}, \quad (1)$$

where $\hat{S}_n (t) = \sum_{s} r_{st} Z_s Z'_s \sum_{s} r_{st}$, is a weighted empirical covariance matrix with weights given by a symmetric nonnegative kernel $r_{st} = K (|s - t|/h_n)$ over time. In order to get a robust alternative to the estimator in equation (1) we follow the approach of Hirose et al. (2017) and we consider the $\gamma$–divergence between the postulated and the true unknown distribution. More precisely,
let
\[ \ell_\gamma (Z, \Theta) = -\frac{1}{\gamma} \ln \left\{ \frac{1}{n} \sum_s f(Z_s, \Theta)^\gamma \right\} + \frac{1}{1+\gamma} \ln \int f(Z, \Theta)^{1+\gamma} dZ, \] (2)

the negative $\gamma$–scoring function as implied by the $\gamma$–divergence then, the $\gamma$–Lasso maximum likelihood estimator of $\Theta$ is obtained as the solution of the following minimisation problem:

\[ \hat{\Theta}_n (t) = \arg \min_{\Theta > 0} \ell_\gamma (Z, \Theta) + \frac{\lambda}{2} \| \Theta - \text{diag} (\Theta) \|_1, \text{ for } t = 1, 2, \ldots, n. \] (3)

To handle the dynamic nature of the problem, we introduce the time–varying $\gamma$–Lasso maximum likelihood estimator as the solution of the following minimisation problem:

\[ \hat{\Theta}_n (t) = \arg \min_{\Theta > 0} \ell_{\gamma,t} (Z, \Theta) + \frac{\lambda}{2} \| \Theta - \text{diag} (\Theta) \|_1, \text{ for } t = 1, 2, \ldots, n, \] (4)

where
\[ \ell_{\gamma,t} (Z, \Theta) = -\frac{1}{\gamma} \ln \left\{ \sum_s \left( r_{st}^{1/\gamma} f(Z_s, \Theta) \right)^\gamma \right\} + \frac{1}{1+\gamma} \ln \int f(Z, \Theta)^{1+\gamma} dZ, \]

and $r_{st} = K (|s - t|/h_n)$ as defined before. In order to solve the minimisation problem we propose an MM algorithm, slightly modifying the one proposes in Hirose et al. (2017).

3 Application

We considered the daily time series of 39 cryptocurrencies from September 2016 to November 2017. We applied the time varying undirected graph proposed by Zhou et al. (2010) and our robustified version on daily log returns. We choose the uniform kernel with a bandwidth that consider a time window of ten days. We observe that the estimator proposed in Zhou et al. (2010) does not catch any conditionally independence structure among the cryptocurrency considered, namely, it always provides a graph edges–free. Our robustified version instead detects time–varying conditional independence structures. As an example, in figure 2 we plot the graph structure obtained using the time–varying approach of Zhou et al. (2010) (left panel) and our robustified version (right panel) estimated over the 10 days around the May 10, 2017, being the time period in which almost all the cryptocurrencies experienced an increase of their value, see Figure 1.
FIGURE 1. Left panel: 39 cryptocurrencies’ time series considered in the application in log scale. Right panel: log returns’ time series.

FIGURE 2. Left panel: estimated graph through the time varying approach of Zhou et al. (2010). Right panel: estimated graph through our robustified method. The estimation is in a neighbourhood of the May 10 2017, corresponding to a general increase of the cryptocurrencies’ value.

References


Attribution of large-scale drivers of peak river flows in Ireland

Aoibheann Brady¹, Julian Faraway¹, Ilaria Prosdocimi¹

¹ University of Bath, United Kingdom

E-mail for correspondence: a.brady@bath.ac.uk

Abstract: Several large flooding events in recent years have led to increased concerns that climate change may be affecting the risk of flooding. At-site tests assessing whether change can be detected in observed data are not very powerful and cannot fully differentiate between possible confounders. It is also difficult to detect fully climate-driven trends, and separate these from other anthropogenic impacts such as urbanisation. We propose a change in focus from detection only towards both detecting and attributing trends in peak river flows to large-scale climate drivers such as the North Atlantic Oscillation index. We focus on a set of near-natural “benchmark” catchments in Ireland in order to detect those non-human driven trends. In order to enhance our ability to detect a signal, we model all stations together in a Bayesian framework which is implemented through Stan.

Keywords: Bayesian hierarchical models; Detection & attribution; Flood risk.

1 Problem

Ireland has been hit by a number of severe floods in recent years, leading to concerns of an increased frequency and severity of floods. Climate change projections allude to increases in extreme precipitation (Bates (2009)), with the belief that this may contribute to an increase in peak river flows. These possible changes in flood risk are derived from climate change projections, however at-site tests using the (relatively short) observed river flow data records do not display compelling evidence of increasing trends. Such tests are not very powerful in a statistical sense. Current flood risk estimation approaches largely assume stationarity of the river flow process, which in practice means assuming the probability of an extreme event is constant. This approach can fail to accurately estimate the frequency of extreme flooding events, which can be extremely costly to the government and the
public alike. Consistently underestimating any trends in peak river flows may mean that flood infrastructures are unfit for purpose for future extreme events. Another issue is the focus on the detection of time trends, rather than attributing any such trends to a variable of interest. It is also often challenging to separate out anthropogenic changes from natural climate variability, making it difficult to accurately attribute any such trends.

We propose a new method for investigating potential drivers of peak river flows in Ireland, taking into account each of these issues. We focus on a combined approach of the detection and attribution of trends. We investigate the effect of large-scale climate indices such as the North Atlantic Oscillation and East Atlantic index on peak river flows in Ireland. We model all stations together within a Bayesian multilevel framework, both to improve the power of such models and to make use of the natural hierarchical structure of spatial data. Many of the river gauging stations are geographically or hydrologically close to each other, and we might expect trends of these nearby stations to be similar. Including all stations in a model together with a spatial random effect to account for correlation between sites may help us to obtain evidence of any trends in river flows that would have been too weak to detect otherwise. Finally, we focus on a set of near-natural “benchmark” catchments in order to detect those trends driven by natural climate variability.

2 Data

We focus on annual maximum river flow data from both the Republic of Ireland and Northern Ireland across a series of reference benchmark catchments introduced by Murphy et al. (2013) and Harrigan et al. (2017). These reference catchments have long records of good hydrometric quality, are relatively near-natural and representative of the land’s hydrology. They were chosen to overcome the difficulty in accurately attributing climate-driven trends, which may be due to human impact and changes in hydrometric performance in gauging stations over time.

The annual maximum series for benchmark catchments for Northern Ireland was obtained from the UK National River Flow Archive, while data for the Republic of Ireland was obtained through the Office of Public Works hydrometric database. The merged data set contains the largest observed instantaneous peak flows at each station in each water year (which run from October to September), measured in m$^3$/s. In total, there are 1,660 observations from 35 gauging stations, ranging from 1950-2015.

We focus on the combined approach of the detection and attribution of trends, by investigating the relationship between peak river flows and large-scale climate indices such as the North Atlantic Oscillation (NAO) or the East Atlantic (EA) Index. These indices are impacted by climate change in a more direct manner than precipitation, so they are proxies for climate
which is changing (but are also quite variable to begin with). Both the
NAO and EA are models of natural climate variability, which impact the
weather and climate of the Atlantic and Europe. Monthly data for both
of these indices was obtained from the National Oceanic and Atmospheric
Administration website.

3 Methods

We utilise the fact that nearby stations can be expected to be impacted
in a similar way by external variables. The model is written as the combi-
nation of some overall trend, a random effect to account for measurement
correlation at each station, and a random effect which has some spatial
structure. For station $i$ at time $t$, we have:

$$\log(\text{Flow})_{it} = \alpha + X_t \beta + r_i + s_i + \epsilon_{it},$$  \hspace{1cm} (1)

where $X$ is the matrix of explanatory variables we are investigating, $\epsilon_i \sim N(0, \sigma_i^2)$ is the measurement error, $r_i$ is a random effect to allow for vari-
ation between stations and $s_i$ is a spatial random effect ($s \sim \text{MVN}(0, \Sigma)$) to allow for correlation between nearby stations. We utilise an exponential
correlation structure, where distance is based on the Euclidean distance
between sites $i$ and $l$, which we model through a Gaussian process (GP)
given by

$$K(x|\eta, \rho)_{il} = \eta^2 \exp\left(\frac{-d_{il}}{\rho}\right).$$

The hyper-parameter $\rho$ describes the characteristic length-scale over which
sites $i$ and $l$ influence each other, while $\eta$ is the marginal standard deviation
controlling the magnitude of the function’s range. Following the recommen-
dations of Gelman et al. (2017), we use weakly informative priors for these
parameters. These models are implemented through Stan.

4 Results & Analysis

We investigate models relating the log of the annual maximum flows to
time, NAO, EA and combinations of these in a preliminary analysis. Fig-
ure 1 suggests that both EA and time are associated with the log of the
peak flows, but not NAO. However, when NAO is modelled alone (not
shown here), there appears to be a clear association with peak flows. This
suggests that there may be some time-varying confounding which we need
to account for when attributing the effect of climate indices on peak flows.
Mapping the spatial random effect $v_i$ in Figure 2 shows that there appears
to be some spatial clustering of trends, particularly in the south east. This
will be explored in a future analysis.
This preliminary analysis suggests that peak river flows in Ireland are associated with both time and climate indices. However, it is likely that both
the EA and NAO are confounded with time, which must be accounted for in future approaches. In addition, a causal framework must be explored in order to accurately attribute the impact of these climate indices on peak river flows in Ireland.

References


Modelling uncertainty and inaccuracy in ordinal data

Roberto Colombi¹, Sabrina Giordano²

¹ Department of Management, Information and Production Engineering, University of Bergamo, Italy
² Department of Economics, Statistics and Finance “Giovanni Anania”, University of Calabria, Italy

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

Abstract: In rating surveys, when ordinal categorical data are collected, an observed category is not necessarily the exact expression of the opinion of the interviewee. We provide a model which accounts for three possible mechanisms: accuracy, inaccuracy and uncertainty which drive individual judgements. We assume that a respondent, who does not respond accurately, either selects a category according to a probability function that models his inaccuracy or according to a Uniform distribution which models his random choice (uncertainty).

Keywords: Rating survey; Mixture models; Discrete distributions.

1 The model

In rating surveys people are requested to express a preference on a topic by selecting a category on an ordered scale. It may be plausible to believe that some respondents select the category which reflects exactly their personal knowledge, other people instead make mistakes in simplifying their belief in one category, someone else can also choose at random the answer. So we identify three components: accuracy, inaccuracy and uncertainty that in our opinion distinguish the three mentioned response styles and propose a model accounting for them.

Suppose that an individual is asked to rate an object on the scale \( S = \{1, 2, \ldots, m\} \) and assume that \( i^* \) is his correct evaluation of the item. An individual, due to lack of competence, interest, involvement, motivation etc., may be inaccurate and so the observed rating \( i \), in general, is different from \( i^* \).

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
In our approach, inaccurate interviewees report a rating \( i \), according to a probability function \( p_I(i; i^*, m) \) with a mode at \( i^* \) and support \( S \), while uncertain interviewees select the response at random according to the uniform distribution \( U(i; m) \) on \( S \). A convenient parameter free specification of \( p_I(i; i^*, m) \) is the triangular probability function \( T(i; i^*, m) \) with support \( S \) and mode at \( i^* \)

\[
T(i; i^*, m) = \frac{2}{m+1} \left\{ \frac{i}{i^*} \delta_{i<i^*} + \frac{m-i+1}{m-i^*+1} \delta_{i \geq i^*} \right\}, \quad i = 1, 2, \ldots, m.
\]

If \( \pi_I \) is the probability of an inaccurate answer, \( \pi_U \) of an uncertain one and \( \pi_A = 1 - \pi_I - \pi_U \), a rating \( i, i \in S \), is obtained, from an interviewee with true rating \( i^* \), with probability

\[
p(i; i^*, \pi_I, \pi_U, m) = \pi_A \delta_{i=i^*} + \pi_I T(i; i^*, m) + \pi_U U(i; m), \quad i = 1, 2, \ldots, m
\]

that, for every true rating \( i^* \), is called response probability function.

As the response probability function is unimodal with mode at \( i^* \), it is intuitive that the more it is peaked around the mode the less relevant are the effects of uncertainty and inaccuracy in answering. Let us explain the meaning of peakedness. Let \( X \) and \( \tilde{X} \) be two discrete unimodal random variables with support \( S \), same mode \( i^* \) and probability functions \( p(i) \) and \( \tilde{p}(i) \), \( i = 1, 2, \ldots, m \), respectively. The random variable \( X \) is more concentrated around the mode \( i^* \) than \( \tilde{X} \) or, analogously, the probability function \( p(i) \), \( i = 1, 2, \ldots, m \) is more peaked than \( \tilde{p}(i) \), \( i = 1, 2, \ldots, m \) if

\[
\sum_{j=i}^{i^*} p(j) \geq \sum_{j=i}^{i^*} \tilde{p}(j), \quad i = 1, 2, \ldots, m^*, \quad \sum_{j=i^*}^{i^*+1} p(j) \geq \sum_{j=i^*}^{i^*+1} \tilde{p}(j), \quad i = i^*, i^*+1, \ldots, m.
\]

The peakedness order for the response probability functions is equivalent to conditions \( \pi_U < \hat{\pi}_U, \pi_A > \hat{\pi}_A \), where \((\pi_U, \pi_I, \pi_A)\) and \((\hat{\pi}_U, \hat{\pi}_I, \hat{\pi}_A)\) are the probabilities of categories \( U: \text{Uncertainty}, I: \text{Inaccuracy} \) and \( A: \text{Accuracy} \) of two ordinal variables.

If the probabilities of the true ratings are denoted by \( \tau(i^*; \theta, m) \), \( i^* = 1, 2, \ldots, m \), where \( \theta \) is a vector of unknown parameters, the distribution of the observable ratings is

\[
q(i; \theta, \pi_I, \pi_U, m) = \sum_{i^*=1}^m \tau(i^*; \theta)p(i; i^*, \pi_I, \pi_U, m), \quad i = 1, 2, \ldots, m.
\]

Figure 1 illustrates an example of true, observation and response probabilities for a choice of \( \pi_A, \pi_U, \pi_I \). The observation probabilities \( q(i; \theta, \pi_I, \pi_U, m) \) depend on \( \pi_I, \pi_U \) and \( \tau(i^*; \theta, m) \), thus some identifiability restrictions on \( \tau(i^*; \theta, m) \) must be satisfied. In doing that, a flexible and parsimonious way of modelling different
shapes of the true rating distribution is specifying \( \tau(i^*; \theta, m) \) by the polynomial model of degree \( d \) \((d \leq m - 4)\)

\[
\log \frac{\tau(i^*+1; \theta, m)}{\tau(i^*; \theta, m)} = \theta_0 + \sum_{g=1}^{d} \theta_g l_g(i^*), \quad \theta = (\theta_0, \ldots, \theta_d), \quad i^* = 1, 2, \ldots, m - 1.
\]

For \( g = 1, 2, \ldots, d \), \((l_g(1), l_g(2), \ldots, l_g(m-1))\) is an orthogonal polynomial contrast of order \( d \) defined on \( \{1, 2, \ldots, m-1\} \).

When the observations come from \( H \) different strata, defined by the levels of a set of categorical covariates, a suffix \( h, s = 1, 2, \ldots, H \), is added. Effects of covariates can be investigated by the parallel logit model

\[
\theta_{0h} = x_{0h}'\gamma, \quad \theta_{gh} = \theta_{g}, \quad g = 1, 2, \ldots, d, \quad h = 1, 2, \ldots, H
\]

and through the linear models for local logits \( \log \frac{\pi_{1h}}{\pi_{0h}} \) and \( \log \frac{\pi_{4h}}{\pi_{3h}} \) to point out which group of respondents has higher inclination toward inaccuracy or accuracy or uncertainty in answering.

Under multinomial sampling in every stratum, inference based on maximum likelihood estimators is conducted considering that the model is correctly specified or misspecified. Confidence intervals, test statistics and their limiting distributions are accordingly determined (Vuong, 1989).

2 Example

As an example, we consider a study about the perception of Italian consumers on extra virgin olive oil quality. The survey involved 1000 subjects selected by the Nielsen Panel and the data were analyzed in Corduas (2015). We investigate the degree of importance that interviewees attach to flavour for their purchasing decisions, using a 5 points Likert scale, and we verify whether the judgements depend on the level of product knowledge \( K \) (low or high) that the consumers possess and the gender \( G \). The model we are
proposing is particularly useful for a sensory analysis and marketing perspectives because it contemplates that the respondents can be accurate or inaccurate in selecting a category, or uncertain in formulating at random their opinion after tasting. We use the additive models

\[
\theta_{0hl} = \gamma_0 + \gamma^G_h + \gamma^K_l, \quad \text{with } \gamma^G_1 = \gamma^K_1 = 0,
\]

\[
\log \frac{\pi_{Ih}}{\pi_{Uh}} = \beta_I + \beta^G_h + \beta^K_l, \quad \text{with } \beta^G_1 = \beta^K_1 = 0,
\]

\[
\log \frac{\pi_{Ah}}{\pi_{Ih}} = \beta_A + \beta^G_h + \beta^K_l, \quad \text{with } \beta^G_1 = \beta^K_1 = 0,
\]

where \( h = 1, 2, l = 1, 2 \), to explain how the true response and the propensity to being accurate/inaccurate/uncertain in answering differ according to gender (\( \gamma^G_2 = -0.7376, se = 0.4372, \beta^G_2 = 0.9879, se = 0.9497 \)) and product knowledge (\( \gamma^K_2 = 1.28, se = 0.6012, \beta^K_2 = 2.0398, se = 0.6252 \)). The estimated probabilities \( \pi_A, \pi_I, \pi_U \) for the four groups of consumers are represented in the ternary plot (Figure 2)

The most accurate are women with high knowledge of the product (\( \pi_A = 0.84, \pi_U = 0.004, \pi_I = 0.14 \)); men who don’t know very well the product are inaccurate in the choice and tend, more than other consumers, to answer at random (\( \pi_A = 0.14, \pi_U = 0.34, \pi_I = 0.51 \)); women with less information on the oil show a propensity to being inaccurate but not uncertain (\( \pi_A = 0.66, \pi_U = 0.027, \pi_I = 0.31 \)) and, finally, men with high personal knowledge rest as inaccurate as but less uncertain than men with low competences (\( \pi_A = 0.37, \pi_U = 0.12, \pi_I = 0.37 \)).

References


A Note on Latent Rating Regression for Aspect Analysis of User-Generated Content

Nelu Dan\(^1\), Ruggero Bellio \(^2\), Thomas Reutterer\(^1\)

\(^1\) WU Vienna University of Economics and Business, Austria
\(^2\) University of Udine, Italy

E-mail for correspondence: ruggero.bellio@uniud.it

Abstract: This work presents an approach for aspect analysis of user-generated text reviews about a service or a product. We build upon some proposals made in the machine learning literature to develop an approach from a statistical perspective. The core of the method is a random effects model for latent rating regression, estimated including some further regularization for sparsity.

Keywords: Online Reviews; Random Effects; Ridge Regression.

1 Introduction

Aspect analysis of user-generated contents is a class of methods developed in the machine learning literature for analysing online reviews of products or services. Of interest are reviews made of a fragment of text coupled with an overall rating expressed by the author of the review. The aims are (i) the identification of the main aspects of interest for the users, (ii) the attribution of the latent rating assigned in the review to each aspect, and finally (iii) the estimation of the latent weights placed on each different aspect by the reviewer. Among the contributions that investigated this topic in the machine learning literature, of particular interest are the two papers by Wang et al. (2010, 2011). The aim of our proposal is build on the methods of these two papers, re-interpreting and extending them from a statistical perspective to obtain more satisfactory inferences and predictions.

2 A two-step approach

The typical data of interest include a set of documents \(d = 1, \ldots, D\), each consisting of a set of words plus an overall rating \(r_d\). The vocabulary used

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
in the entire set of documents can be mapped to the set 1, . . . , V, where V is the number of different words employed. We propose the following two-step approach.

2.1 Step 1: Building a word-frequency matrix

The starting point of the entire analysis is the identification of the number of aspects K considered by the reviewers. We recommend to choose K based on domain-specific considerations. For instance, for hotel reviews, possible aspects may include Value for money, Room, Location, Cleanliness, Service, as in the example considered later.

The first step builds a feature matrix \( W_d \) for each document, that consists of the frequency of each word of document \( d \) for each of the \( K \) aspects, so it is a \( K \times V \) table of (processed) frequencies. Note that \( V \) is around many thousands in typical applications, so that \( W_d \) is extremely sparse. Its construction requires the assignment of each word to one of the \( K \) aspects. Differently from Wang et al. (2010, 2011), we endorse the direct application of Latent Dirichlet allocation (LDA) for topic models for this task (Blei et al., 2003). The crucial twist is to apply LDA in the sentence-based variant (Büschken and Allenby, 2016), which overcomes the bag of words assumption and results in a better generative model for online reviews.

2.2 Step 2: Late Rating Regression

The starting point of Latent Rating Regression (LRR) is a linear model for the observed rating \( r_d \) of the \( d \)-th document

\[
r_d = s_d^T \alpha_d + \varepsilon_d , \quad \varepsilon_d \sim \mathcal{N}(0, \delta^2) ,
\]

with independence assumed across reviews. Here \( s_d \) and \( \alpha_d \) are both vectors of length \( K \), containing the aspect ratings and the aspect weights of document \( d \), respectively. In particular, the \( k \)-th component of \( s_d \) is

\[
s_{dk} = \sum_{v=1}^{V} \beta_{kv} w_{dkv} ,
\]

where \( w_{dkv} \) is an entry of \( W_d \) and \( \beta_{kv} \) is the \((k,v)\) element of the matrix \( \beta \), of size \( K \times V \), collecting the sentiment polarities of each word of the vocabulary for each of the \( K \) aspects. Note that if a certain aspect is not detected in a review, the corresponding rating is undefined.

The assumptions for the aspect weights are a key feature of the model. The fact that the elements of \( \alpha_d \) are constrained to sum to 1 is handled by defining the transformation

\[
\eta_d = \log \left\{ \alpha_d / g(\alpha_d) \right\} ,
\]
where \( g(\cdot) \) is the geometric mean of its argument, so that
\[
\alpha_{dk} = \frac{\exp\{\eta_{dk}\}}{\sum_{k=1}^{K} \exp\{\eta_{dk}\}}, \quad k = 1, \ldots, K.
\]

We then assume a singular multivariate normal distribution for \( \eta_d \), with \( K \)-dimensional mean vector \( \mu \) and covariance matrix \( \Sigma \) of size \( K \times K \) (see Rao, 1973, §8.a4). Since \( \sum_{k=1}^{K} \eta_{dk} = 0 \) it follows that \( \sum_{k=1}^{K} \mu_k = 0 \) and the matrix \( \Sigma \) has rank \((K - 1)\), with \((K - 1)\) positive eigenvalues and one zero eigenvalue. The assumption implies a logistic normal distribution on the simplex for the aspect weight vector \( \alpha_d \) (see Aitchison, 1986, Ch. 5).

3 Estimation of the model parameters

The LRR step treats the word-frequency matrix of Step 1 as fixed by design. This facilitates the estimation of the corpus-level parameters \( \theta = (\mu, \Sigma, \beta, \delta^2) \). Of interest are also the document-level parameters \( \alpha_d \), which are random effects to be predicted for each document.

The estimation proceeds by jointly estimating \( \theta \) and \( \alpha_d, d = 1, \ldots, D \). Inferentially, this could be seen as an application of Henderson’s mixed equations, where fixed effects and random effects are jointly estimated based on observed normal data. The frequentist theory of Maximum a Posterior Estimation (MPE) (see Jiang, 2007, Ch. 3) provides a theoretical framework. Regularization is required for the estimation of \( \beta \), to account for sparsity. Some simple algebra based on (1) and (2) for given values of \( \alpha_d \) allows to re-express the linear model for the observed rating as
\[
r = X(\alpha)\beta_v + \varepsilon,
\]
where \( X(\alpha) \) is a design matrix of size \( D \times (KV) \), whose \( d \)-th row depends on \( \alpha_d \) and \( W_d \), and \( \beta_v = \text{vec}(\beta) \). Among the methods available to regularize the estimation of \( \beta_v \) (see Hastie et al., 2009), ridge regression seems the most suitable one. Regularization overcomes the fact that for the MPE method some bias on the estimated corpus-level parameters may arise when random effects are maximised rather than integrated out (Jiang, 2007, p. 136).

The final steps consists in post-processing the predicted aspect ratings, given by
\[
\hat{s}_{dk} = \sum_{v=1}^{V} \hat{\beta}_{kv} w_{dkv},
\]
and mapping them on the same scale used for the overall rating. A natural possibility is to perform the adjustment by using a link function, though possible refinements of this approach exist.
4 A simple example

To give a flavour of the results of the proposal, we applied it to a medium-size sample of about 100,000 hotel reviews, with a vocabulary size $V$ around 24,500. This is part of a larger data set, which is not disclosed for confidentiality reasons.

An important feature of this data set is the inclusion of the actual user preferences about the five aspects mentioned in §2.1. To visualise the accuracy of the proposed methodology, we can compare the averages of estimated aspect ratings with the averages of actual ratings. This is done in Figure 1 for about 200 hotels with at least 100 reviews, for the overall rating and the user rating for the five aspects. Here the aspect ratings computed by formula (2) have been transformed by a suitable logistic link.

![Figure 1](image_url)

**FIGURE 1.** Relation between averages of estimated ratings ($y$-axes) versus averages of actual ratings ($x$-axes) for 222 hotels with at least 100 reviews. The least squares line (red) and the $y = x$ line (dashed) are superimposed.

We note that the correlation for the overall rating is quite high (around 0.99), which is to be expected given the nature of LRR. Despite some underestimation for the *Cleanliness* aspect, all in all the results are quite encouraging also for the five latent aspects, with large correlations in all the cases but for the *Location* aspect (around 0.51). Indeed, the true rating of this aspect is the one least correlated with the overall rating, and this makes its prediction harder than the rest, since the the main idea of the LRR approach is to disentangle the overall rating.
5 Conclusion and ongoing research

The results presented here are still preliminary, and the research is still progressing in several directions. Here we discuss two points which seem of some importance, focusing in particular on the LRR step. First of all, there are some computational aspects to handle. The current R (R Core Team, 2018) implementation is built around two key ideas. The first one is to perform the update of the review random effects taking advantage of parallel computing. The review-specific computations are done in parallel and efficiently, but there is room for improvement. The other key point is to employ sparse linear algebra for the ridge regression step, which is currently carried out using the LiblineaR package (Helleputte 2017), the port to R of the LIBLINEAR C/C++ library for machine learning. A task that requires some attention is to improve the selection of the tuning parameter for ridge regression as the estimation algorithm proceeds.

On the inferential side, an important open point is the validation of the methodology by not relying entirely on observed data expressed by reviewers. Indeed, we noticed that the actual user ratings deviated from the results predicted by the model based on the review text in cases when users tend to evaluate different aspects very similarly. The latter is a well known empirical phenomenon in the context of online reviews, where we frequently observe bimodal (J-shaped) distributions with more positive than negative reviews and carry-over effects across items if multiple quality dimensions are evaluated by reviewers (Hu et al. 2009). Against this background, the proposed modelling framework is potentially capable of replacing the often biased customer ratings with a set of more accurate ratings for sub-components of product or service qualities which are directly inferred from customers verbal comments. Such an approach would provide marketers with much more valuable and reliable customer feedback information than conventional customer rating systems currently do. Whether the inferred aspect ratings better reflect the true underlying meaning of quality feedback remains an open hypothesis which surely requires some empirical support.

Acknowledgments: This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003.

References


Büschenken, J. and Allenby, G.M. (2016). Sentence-based text analysis for

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Sta-


have a J-shaped distribution? Overcoming biases in online word-of-

Applications*. New York: Springer.

R Core Team (2018). R: A language and environment for statistical com-


Wang, H., Lu, Y. and Zhai, C. (2010). Latent aspect rating analysis on re-
view text data: A rating regression approach. In: *Proceedings of the
16th ACM SIGKDD International Conference on Knowledge Discov-
ery and Data Mining*, Washington, D.C., 783–792.

out aspect keyword supervision. In: *Proceedings of the 17th ACM
SIGKDD International Conference on Knowledge Discovery and Data
Mining*, San Diego, California, 618–626.
Statistical modelling of time dependence in peak river flows

Emiko Dupont¹, Ilaria Prosdocimi¹, Daniel Simpson², Nicole Augustin¹

¹ University of Bath, UK  
² University of Toronto, Canada

E-mail for correspondence: e.dupont@bath.ac.uk

Abstract: The apparent prevalence and severity of recent flood events in the UK as well as perceived evidence of climate change lead to the belief that there is an upward trend in peak river flows over time. However, as data records are typically short (on average only around 45 years) it is difficult to obtain any compelling statistical evidence. We implement a new method for statistically testing for time dependence in peak river flows. Instead of considering gauging stations separately, we build a spatial model for all stations in Great Britain. This allows us to pool information from several stations in order to improve the statistical power of each test. Our results indicate that peak river flows have a significant positive time trend across the majority of Great Britain with the highest dependence in Northern England and Western Wales. Current flood management models may therefore be underestimating the risk of future extreme events by assuming stationarity of peak river flows.

Keywords: Flood risk; Time dependence; Spatial statistics; Statistical power; Partial pooling

1 Aims

The main aim of the project is to investigate whether a consistent time trend can be detected in peak river flow series across Great Britain. To this effect we investigate whether a new partial pooling approach to statistical hypothesis testing can be used to identify spatially coherent trends and geographical regions with stronger time dependence.

---

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
2 Data

Annual maxima of peak river flow was obtained from the National River Flow Archive (nrfa.ceh.ac.uk). We have modelled the flow data for water years up to 2013 from all gauging stations in Great Britain, excluding stations with fewer than 20 years of observations and stations whose data quality was deemed unsuitable. The total number of stations modelled was 748 with a median record length of 45 years and a total of 33,423 station years.

3 Methodology

3.1 No pooling

Our starting point is the ”no pooling” approach of testing for time dependence by modelling stations separately (see Prosdocimi et al (2014)): at a given gauging station, if \(x_1, \ldots, x_n\) is the observed annual maximum flow in water years \(t_1, \ldots, t_n\), then we assume that

\[ y_i = \log(x_i) = \beta_0 + \beta_1 t_i + \varepsilon_i \quad \text{where} \; \varepsilon_i \sim N(0, \sigma^2) \]

where \(\beta_0, \beta_1\) and \(\sigma^2\) are unknown parameters. We can test the significance of time as an explanatory variable by using a two-sided test on the null hypothesis \(H_0: \beta_1 = 0\). Under \(H_0\), the test statistic

\[ T = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}} \sim_{\text{approx}} N(0, 1). \]

Hence, at each station, time dependence is significant at the \(\alpha = 0.1\) level if the observed \(|T|\) is greater than 1.64. Fitting this model to our data, time dependence could be detected at around 20% of stations. One problem with this approach is that since the typical sample size \(n\) at a gauging station is small, then \(\hat{\sigma}_{\beta_1}\) is large and therefore the test statistic \(T\) will not be significant unless time dependence is particularly strong. In other words, the power of the statistical test is weak and therefore only large effect sizes can be detected using this method.

3.2 Partial pooling

In order to improve statistical power, we create a new test that uses spatial modelling or ”partial pooling” to ”borrow” information from nearby stations. Many gauging stations are geographically close together and take measurements on the same river system, and therefore for any given station, information from nearby stations should enhance the signal at the station in question. That is, we would want to pool data from several stations in
order to increase the effective sample size, thereby increasing power. This approach of using spatial statistical modelling to improve small area estimation was introduced by Chen et al (2014) in the context of data arising from complex surveys.

Let \( T = (T_1, \ldots, T_{748}) \) denote the vector of observed test statistics at the 748 gauging stations in our dataset. The stations can be divided into 97 hydrometric areas (HAs) which are geographical regions within which stations are considered to be relatively similar to each other. We define the following multilevel model

\[
T \mid \mu, \sigma_b = \mu + Zb + \epsilon \quad \text{where} \quad b \sim N(0, \sigma_b^2 I), \quad \epsilon \sim N(0, I)
\]

where \( \mu = (\mu, \ldots, \mu) \) is an overall intercept, \( b = (b_1, \ldots, b_{97}) \) is the vector of spatial random effects (one for each HA) with variance parameter \( \sigma_b^2 \) and \( Z \) is a matrix of ones and zeroes whose \( j \)'th row picks out the random effect \( b_j \) if station \( i \) lies in HA \( j \). The model was implemented using the R-inla package with standard priors, that is, a flat prior on the fixed effect \( \mu \) and a loggamma prior on the precision \( 1/\sigma_b^2 \) of the random effects.

We can now define a new hypothesis test for detecting time dependence at each station using the estimated test statistic \( \hat{T}_i \). In other words, if station \( i \) is in hydrometric area \( j \), then under the null hypothesis \( H_0 : \beta_1 = 0 \), the posterior estimate

\[
\hat{T}_i = (\mu + \hat{b}_j) \sim N(0, \sigma_{\hat{T}_i}^2)
\]

so that \( H_0 \) can be rejected at the significance level \( \alpha \) if the probability that a \( N(0, \sigma_{\hat{T}_i}^2) \)-distributed variable takes values greater than the observed \( |\hat{T}_i| \) is less than \( \frac{\alpha}{2} \). We estimated the null distribution of \( \hat{T}_i \) by simulating data under the null hypothesis, fitting the model for \( T \) and using the resulting empirical distribution of \( \hat{T}_i \).

4 Results

Under the partial pooling approach, the null-hypothesis is rejected at the \( \alpha = 0.1 \) level for a large majority of stations, i.e. time dependence is significant in most of Great Britain. The detected time trend is positive everywhere with the strongest time dependence in Northern England and Western Wales. Current flood management models could therefore be underestimating the risk of future extreme events by assuming a stationary river flow process if the estimated trends continue into the future.

5 Simulation study

In order to assess the effectiveness of the proposed methodology, we carried out a simulation study to compare the statistical power under the no pooling and partial pooling approaches. We also included a ”complete pooling”
model in which the test is the same at all stations and based on the overall average test statistic. We simulated data under five scenarios (S1 - S5):

\[
\beta_1 = \begin{cases} 
\beta^* & \text{at station 75002 (S1), in HA 75 (S2), HAs 73 – 76 (S3),} \\
& \text{HAs 68 – 76 (S4), at all stations (S5)} \\
0 & \text{otherwise}
\end{cases}
\]

The resulting power curves are shown in Figure 1. We conclude that when only one station has time dependence, no pooling is best at detecting it - partial pooling has low power for small signals but much more power than complete pooling. In contrast, when many stations have time dependence, partial/complete pooling is much better at detecting small signals.

FIGURE 1. Average power of hypothesis test at stations where \( \beta_1 = \beta^* \).

References


Unified Testing Framework for Feature-sets in Genomics

Mitra Ebrahimpoor\textsuperscript{1}, Roula Tsonaka\textsuperscript{1}, Pietro Spitali\textsuperscript{2}, Jelle Goeman\textsuperscript{1}

\textsuperscript{1} Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands
\textsuperscript{2} Human Genetics, Leiden University Medical Center, Leiden, Netherlands

E-mail for correspondence: m.ebrahimpoor@lumc.nl

Abstract: We introduce a general definition of the null hypothesis for testing feature-sets in genomics and illustrate how to test this new unified null based on the All-resolutions Inference. The strength of our proposal is that testing the competitive null in this framework is statistically sound. Furthermore, researchers do not need to decide beforehand the database and the features-sets they will test. This may even be postponed until after seeing the data.

Keywords: Genomics; Self-contained null hypothesis; Competitive null hypothesis; Gene-set testing.

1 Introduction

Analysing sets of features in genomics data rather than individual features has proved beneficial in interpreting results. Since its proposal, many statistical methods have been established for feature-set testing. These methods are broadly categorized, based on the definition of null hypothesis, into: self-contained or competitive (Goeman and Buhlmann, 2007). Self-contained null is designed to test if “None of the features in the set are active.”, and is false if any feature in the selected set is active. On the other hand, competitive null is designed to test if “Features in the set are at most as active as the background features.”; it is false if more features in the selected set are active compared to features outside the set. Each approach has its own advantages and disadvantages. In case of few active features, the two methods are similar but in general they yield different results. While self-contained methods have a statistically well-defined structure,
they are not advisable when many active features or polygenic processes are present. On the other hand, despite the favourable interpretation of the competitive null, current tests face methodological issues such as the feature independence assumption and feature-permutation. Furthermore, it has been suggested that some competitive tests do not in fact test the competitive null as defined earlier (Maciejewski, 2014; Debrabant, 2017). This article proposes a novel approach, unifying competitive and self-contained testing into a single framework. We introduce a new null hypothesis, which includes the self-contained and competitive types as special cases. We test the unified null hypothesis based on All-Resolutions Inference (ARI) of Goeman and Solari (2011). The proposed method includes calculation of adjusted \( p \)-values along with point estimates for the number of active features. Importantly, adopting ARI provides flexibility in choice of feature-sets to be tested.

2 Unified Null Hypothesis

2.1 Definition

Suppose that a genomics experiment is performed resulting in values for \( m \) features, denoted by \( S \). An unknown subset \( T \) of the studied features are truly active (A). There are \( 2^m - 1 \) possible feature-sets, or pathways, identified by \( S \). We denote the number of truly active features in \( S \) as \( A(S) = |S \cap T| \) and define \( \pi(S) = A(S)/|S| \). The competitive and self-contained null hypotheses can both be formulated in terms of \( \pi \). The self-contained null is \( H_{0\text{self}}(S) : \pi(S) = 0 \), and is rejected if any active features are present or if \( \pi(S) > 0 \). On the other hand, competitive null is \( H_{0\text{comp}}(S) : \pi(S) \leq \pi(S^c) \) where \( \pi(S^c) = A(S^c)/|S^c| \) is the unknown proportion of truly active features in \( S^c \). Hence, both hypotheses are a special case of \( H_U(S, c) : \pi(S) \leq c \), for \( c \in (0, 1) \). We name \( H_U(S, c) \) the unified null hypothesis. Choice of \( c \) would determine if the test is competitive (\( c = \pi(S) \)) or self-contained (\( c = 0 \)).

2.2 Inference

We test the unified null using ARI, which is a multiple testing procedure based on a set of \( p \)-values. In genomics context, \( p \)-values are obtained by individual feature tests. ARI constructs simultaneous confidence bounds for the proportion of truly active features in each set \( S \) (or \( \pi(S) \)). The \( (1 - \alpha) \) confidence bound is such that

\[
P(\hat{\pi}(S) \leq \pi(S); \text{for all } S) \geq 1 - \alpha
\]

(1)

The simultaneity here is over \( S \) and the method controls family-wise error rate (FWER) over all possible feature-sets. Testing the null hypotheses
for any \( c \) is straightforward based on these intervals. As \( m \) grows, calculation time increases exponentially. To resolve this, we adopt the Hommel’s shortcut introduced in Meijer et al. (2017). Adjusted \( p \)-values for the unified null are calculated for any combination of \( c \) and \( S \) by expanding the method of Goeman et al. (2017). Furthermore, a median point estimate for the proportion(number) of truly active features is calculated based on the confidence bounds by setting \( \alpha = 0.5 \).

### 3 Analysis of RNA-seq data

We will investigate pathways associated with muscular dystrophy using rnaseq data of muscle samples from 9 Dystrophic (mdx) and 10 healthy (wt) mice. Raw \( p \)-values per gene have been obtained using the limma pipeline established for RNAseq data. Next we applied our proposed method on the list of all possible pathways from Gene Ontology and Reactome databases. Also, We filtered the pathways according to log-Fold-change (logFC) values of the genes. So a sub-pathway of up-regulated genes (logFC > 0) was created. Figure-1 includes the estimated number of active genes in each set and among up-regulated genes. For illustration purposes, only 15 pathways with highest number of active genes and a significant competitive \( p \)-value (adjusted \( p < 0.05 \)), are shown.

![FIGURE 1. RNA-seq results for 15 pathways from GO(left) and Reactome(right) databases.](image)

### 4 Simulation study

A simulation experiment was designed to evaluate the power properties of the unified approach. Fisher’s exact test with FDR and FWER control were selected for comparisons. The feature-sets were defined according
to the *Gene Ontology* database. A *p*-value was generated for each of the 24,500 features in the database. The experiment is based on 5 parameters: size of pathway, effect size, proportion of active feature in the pathway and background. For the non-active features, *p*-values were calculated based on a random value from a standard normal distribution; while for active features, a normal distribution with different mean values is used. Simulated *p*-values are independent, which is required by Fisher’s test, but not the unified approach. 1000 simulations were performed. Power was defined as the proportion of *p*-values < 0.05. Figure-2 plots the power values for the three methods, for a small pathway (size=50). A similar pattern holds for the pathways of size 50 and 200. As evident by Figure-2, unified testing performs better when proportion of active features is small and has reasonable power in other cases considering the freedom it provides. In general, power for both methods is lower for small feature-sets than for large ones. ARI loses less power than Fisher’s exact test if the feature-set gets smaller, but conversely gains less if the feature-set gets larger.

![Figure 2: Power values for unified testing, Fisher's exact test with FDR and FWER correction](image)

**FIGURE 2.** The bold lines represents power for the unified testing; the dashed and dotted lines represent the fisher’s exact test with FDR and FWER correction, respectively.
5 Conclusion

Unified testing has several appealing features: first, both single features and features-sets are tested within the same framework with an overall error control. Second, testing for the competitive null does not depend on feature independence or gene-permutation. Finally, choices of test type and database are made after seeing the results, as the method controls FWER over all possible sets.

References


Goodbye moments, hello expectiles

Paul H. C. Eilers¹

¹ Erasmus University Medical Center, Rotterdam, the Netherlands

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

Abstract: The method of moments computes averages of powers of the variable under study and matches theoretical moments to them, by tuning the parameters of a known distribution. I propose to replace moments by expectiles and present interesting results of a first exploration of this idea.

Keywords: Asymmetric least squares; moment matching; density estimation.

1 Introduction

The method of moments was invented by Karl Pearson, at the end of the 19th century, as a procedure for estimating parameters of statistical distributions. The basic idea is simple: equate empirical and theoretical moments, where the moments are expected values of powers of $y$, the variable under study. An example is the estimation of expectation and variance of a (normal) distribution from averages of the observations themselves and their squares. A disadvantage of moments is that averages of third and higher powers of the observations are extremely sensitive to outlying values. That can make the method of moments unreliable.

Here I explore expectiles as alternatives for moments. Expectiles, based on asymmetric least squares, have pleasant properties and are easy to compute. One parameter, the asymmetry, can be varied continuously between 0 and 1, allowing a set of many expectiles to be computed. They are well defined for continuous and discrete distributions.

In many applications, conditional distributions are of prime interest; growth curves, of height or weight with age, provide a familiar example. Smooth expectile curves have been used to characterize growth data. The curves themselves already give a good impression of change. Parametric expectile matching (PEM) makes it easy to fit parametric distributions afterwards.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
FIGURE 1. Body mass index of Dutch boys. Left: data (n = 5265) and expectile curves. Right: estimated skew-normal density, based on the expectiles at age 5, (for \( p = 0.01, 0.02, 0.05, 0.1, 0.2 0.5, 0.8, 0.9, 0.95, 0.98 \) and 0.99). Crosses indicate the expectiles.

2 Theory and applications

It is well known that the weighted sum of squares \( S = \sum_i w_i(y_i - \phi)^2 \) is minimized by the weighted average \( \hat{\phi} = \sum_i w_i y_i / \sum_i w_i \). Equivalently, \( \sum_i w_i(y_i - \hat{\phi}) = 0 \), an estimating equation showing that \( \hat{\phi} \) is the root of the estimating function \( \psi(\phi, y) = \sum_i w_i(y_i - \phi) \).

We get interesting and useful results when we set \( w_i = p \) when \( y_i > \phi \) and \( w_i = 1 - p \) otherwise, where \( 0 < p < 1 \) is called the asymmetry. This is asymmetric least squares and \( \hat{\phi} \) is called the \( p \)th expectile (Newey and Powell, 1987). A convenient way to write the objective function and the estimating equations is

\[
2S = \sum_i r_i(r_i + q|r_i|) \quad \text{and} \quad \sum_i (r_i + q|r_i|) = 0, \tag{1}
\]

where \( r_i = y_i - \phi \), \( q = 2p - 1 \) and \( -1 < q < 1 \). I call \( q \) the balance.

Given \( q \), expectiles can be computed by an iterative weighted averaging algorithm, updating weights and \( \hat{\phi} \) in turn. Given \( \phi \), \( q \) is found explicitly from \( q = \sum r_i / \sum |r_i| \). The objective function is convex, so there is a unique minimum, and move down the gradient, so that minimum will be reached. In practice only a handful of iterations are sufficient.

Above, expectiles were introduced for observations, but they apply equally well to distributions. Let \( f(u; \theta) \) be a (continuous or discrete) density with parameter vector \( \theta \), then its theoretical expectiles follow from

\[
\int f(u; \theta)(u - \phi + q|u - \phi|)du = 0. \tag{2}
\]
Combining these ideas, I propose to estimate a parametric density by expectile matching, such that

$$\int f(u; \theta)(u - \hat{\phi}_j + q_j |u - \hat{\phi}_j|)du = g_j \approx 0. \quad (3)$$

for a vector of observed expectiles $\hat{\phi}$ with their corresponding balances $q$. An obvious choice is to minimize $\sum g_j^2$. To make this a practical scheme, I compute $f_k = f(u_k; \theta)$ on a fine grid (say 100 positions) and replace the integral by a sum. In general, numerical optimization will be needed to obtain $\hat{\theta}$.

Expectiles summarize the data and one can use them to investigate a variety of parametric distributions. With many expectiles, the summary can be quite detailed. Very conveniently, normalizing constants are not needed, simplifying the computations. The position of the root of an estimating function does not change if we multiply the function by an arbitrary constant.

The left panel of Figure 1 shows the body mass index of Dutch boys, as found in the R package AGD, with a set of 11 smooth expectile curves (Schnabel and Eilers, 2010). At the age of 5, very few observations are available. The right panel shows the estimated skew-normal density.

For the optimization I used the function `ucmfun` in the R package of the
same name. My initial choice was the standard function optim, but it turned out to be unreliable.

Figure 2 shows an application to discrete data. The Dutch province Groningen has a serious problem with earthquakes, caused by decades of massive extraction of natural gas. The Royal Dutch Meteorological Institute collects and provides earthquake data, from which I selected the last 15 years. The upper left panel shows counts per half year as a histogram. The curves represent smooth conditional expectiles. In the upper right panel a Poisson and a negative binomial distribution have been estimated by PEM, for the year 2006. The expectile values are indicated by the small crosses. From the estimated distributions follow theoretical expectiles, which can be compared to the observed ones. This is done in the lower left panel. It is clear that the negative binomial distribution works better than Poisson. A similar message is obtained from the lower right panel, where the mismatches, $g$ in (3), are plotted.

Instead of fitting parametric densities, we might go for a non-parametric one. Schnabel and Eilers (2013) showed how to model the logarithm of a smooth density that matches a given set of expectiles. The idea is to estimate $\eta = \log f$ by minimizing the penalized sum of squares

$$S = \sum_j \sum_k || \exp(\eta_k) t_{jk} ||^2 + \lambda \sum_i (\Delta^3 \eta_k)^2,$$

with $t_{jk} = u_k - \hat{\phi}_j + q_j |u_k - \hat{\phi}_j|$. The first term of $S$ states that expectations of the estimating functions should be (close to) zero. With $\eta$ having many more elements than available expectiles, these conditions do not give a unique solution. The second term in $S$ is a roughness penalty; it strives for a smooth result and thereby leads to a unique solution.

The problem is non-linear in $\eta$, but it is straightforward to solve, using a Newton-Raphson algorithm. An extra condition, $\sum_k f_k = 1$, is needed to make the solution truly unique. See the mentioned paper for details.

Figure 3 shows results for the earthquake data, using widely different values of $\lambda$. At the moment I do not have a recipe for using the data to suggest a good value. A still untested idea is to use cross-validation, leaving out each expectile in turn and computing the balance for it, from a density that is estimated form the remaining expectiles.

For $\lambda = 0.1$, the result is multimodal. This seems to make sense: in Figure 2 the spacing of the largest expectiles is relatively small, pointing to a higher probability density there.

### 3 Discussion

Parametric expectile matching (PEM) is simple to implement and it gives good results, in the examples I presented, and in many others I tried. PEM works for continuous and discrete distributions. Expectiles are unique
FIGURE 3. Nonparametric density estimates for the numbers of earthquakes in 2006, based on the expectile curves in the upper left panel of Figure 2. The values of $\lambda$ are indicated in the legend. The thin gray lines are drawn as a visual aid; the distributions are discrete.

for the later, in contrast to quantiles. The normalizing constant can be discarded, which is an advantage when experimenting with distributions, like the double Poisson, where it is hard to obtain.

Nonparametric expectile matching (NPEM) works well. It can be a help when searching for a suitable parametric distribution. One can also accept the estimated density as the final result. More work is needed to find a good recipe for choosing the value of the penalty parameter.

A series of (conditional) expectiles represents a detailed and relatively compact summary of a dataset. One use them to explore new distributions, even when the raw observations are not available.

This is only a beginning, a lot of work lies ahead. There is the question of design: how many expectiles should be used, and where should they be placed? How do we evaluate the quality of the fit? Can we find (equivalents for) confidence intervals? How robust is the method?

The minimum of $\sum_j g_j^2$ has a numerical value, but what is small and what is large? Can we use it to compare distributions? The plot of $g_j$ vs $\phi_j$ is informative; can we calibrate it? The plot of the theoretical expectiles against the observed ones is similar to Q-Q plots based on quantiles.

Small sample performance is interesting too. For the skew-normal distribution, Pewsey (2000) showed that the method of moments can give better results than maximum likelihood. How will expectile matching compare?

When promoting expectiles, I often encounter a reluctant (and sometimes
even aggressive) response from colleagues: “we already have quantiles and expectiles don’t have a strong intuitive interpretation, why should we use them?” The present application does not strain our intuition, and it gives useful results, as expectiles are just a vehicle to estimate densities. I hope it will convince many applied statisticians to take a better look at expectiles. A similar approach is possible with quantiles, but not for discrete distributions. Also it is my experience that quantile smoothing generally does not lead to very smooth curves, which tend to cross. Expectile smoothing is behaving better.

References


Functional principal component analysis for non-stationary dynamic time series

Amira Elayouty¹, Marian Scott², Claire Miller², Susan Waldron³

¹ Faculty of Economics and Political Science, Cairo University, Egypt
² School of Mathematics and Statistics, University of Glasgow, UK
³ School of Geographical and Earth Sciences, University of Glasgow, UK

E-mail for correspondence: a.ayouti@feps.edu.eg

Abstract: Motivated by a highly dynamic hydrological high-frequency time series, we propose time-varying Functional Principal Component Analysis (FPCA) as a novel approach for the analysis of non-stationary Functional Time Series (FTS) in the frequency domain. Traditional FPCA does not take into account (i) the temporal dependence between the functional observations and (ii) the changes in the covariance/variability structure over time, which could result in inadequate dimension reduction. The novel time-varying FPCA proposed adapts to the changes in the auto-covariance structure and varies smoothly over frequency and time to allow investigation of whether and how the variability structure in an FTS changes over time. Based on the (smooth) time-varying dynamic FPCs, a bootstrap inference procedure is proposed to detect significant changes in the covariance structure over time. Although this time-varying dynamic FPCA can be applied to any dynamic FTS, it has been applied here to study the daily processes of partial pressure of CO₂ in a small river catchment in Scotland.

Keywords: Functional Time Series; Frequency Domain; Smoothing; Principal Components; Non-stationarity; Functional spectral density

1 Introduction

Recent advances in sensor technology allow environmental monitoring programs to record measurements at high-temporal resolutions over long time periods, for processes which are in reality continuous in time. These High-Frequency Data (HFD) pose several challenges in terms of statistical modeling and analysis due to the complexity of such large volumes of data stemming from the persistent and dynamic dependence structure over the
different timescales (Elayouty et al. 2016). Functional Time Series (FTS) analysis and its recent developments (Hörmann et al. 2015) provide an appropriate framework for analyzing such HFD, taking into consideration these technical challenges.

This paper introduces a novel approach that identifies and accounts for volatility in FTS. This approach involves the development of frequency domain Functional Principal Components (FPCs) that vary smoothly over time, taking into account both temporal correlation and non-stationarity in the series. Using bootstrap procedures, these time-varying FPCs are employed to statistically assess changes over time in the covariance structure and variability modes of the underlying FTS process.

This work is motivated by highly dynamic HFD of excess partial pressure of carbon dioxide (EpCO$_2$) measured every 15 minutes over 3 years at a small catchment of the River Dee, Scotland. The long sequence of 15-minute measurements is segmented into daily intervals, which are then smoothed using B-splines to form a sequence of daily EpCO$_2$ functions. The data thus form a FTS and are viewed as realizations of a functional stochastic process \( \{X_k(t) : k \in \mathbb{Z}, t \in \mathcal{T}\} \) valued in the Hilbert space \( L^2(\mathcal{T}) \), with \( k \) denoting the day as a discrete time parameter and \( t \) being the intra-day time defined continuously on \( \mathcal{T} \).

## 2 Methodology

The proposed methodology relies on evaluating the Spectral Density (SD) of the FTS process \( \{X_k\} \) at each time point \( k \) and obtaining the dynamic FPCs (Hörmann et al. 2015) via the eigen-decomposition of the SD at each time point \( k \), assuming that the process varies smoothly over time. Because the SD contains information on the whole family of lag-\( h \) covariances, the novel FPCs accommodate the varying serial correlation. Due to the limited number of replicates at each time point \( k \), the local lag-\( h \) covariances and spectral densities are computed by smoothing the sample lag-\( h \) covariances over time using a weight kernel \( w_s(.) \) with smoothing parameter \( s \),

\[
\hat{V}_{k,h} = \frac{1}{\sum_{k' \in \mathbb{Z}} w_s(|k - k'|)} \sum_{k \in \mathbb{Z}} w_s(|k - k'|) X_{k'} \otimes X_{k'+h}. \quad (1)
\]

\( w_s(.) \) is a monotonically decreasing weight function of the distance \( |k - k'| \) regardless of the lag \( h \), ensuring that the highest weights are assigned to the pairs \( (X_{k'}, X_{k'+h}) \) near the target point \( k \). The neighborhood contributing to the covariance estimation is determined by the choice of the kernel and smoothing parameter. The choice of weight kernel is based on the nature of the variable of interest; a common choice is the Gaussian density. The smoothing parameter is chosen so that the process within each neighborhood is stationary without over-fitting the original process.
After computing $\hat{V}_{k,h}$, the local SD is estimated at each time point $k$ by:

$$\hat{F}_{k,\theta} = \frac{1}{2\pi} \sum_{h \in \mathbb{Z}} \hat{V}_{k,h} \exp(-ih\theta), \quad \theta \in [-\pi, \pi],$$

and the local eigenvalues $\hat{\lambda}_{k,m}(\theta)$ and eigenvectors $\hat{\phi}_{k,m}(\theta)$ of $\hat{F}_{k,\theta}$ are calculated. The local functional filters $\{\hat{\phi}_{kml}(t) : l \in \mathbb{Z}\}$ are estimated, via the inverse Fourier transform of $\hat{\phi}_{k,\theta}(\theta)$, and subsequently used to filter the original FTS across a number of lags and leads $l$ to obtain the $m^{th}$ local dynamic FPC scores at $k$ as:

$$\hat{Y}_{m,k}^{(k)} = \int_{t \in T} X_{k-l}(t) \hat{\phi}_{kml}(t) dt.$$

The original curves can thus be approximately reconstructed, based on these scores, using $q$-term (smooth) time-varying dynamic FPCs, $q < \infty$, as follows:

$$\hat{X}_k(t) \approx \sum_{m=1}^{q} \sum_{l=-L}^{L} \hat{Y}_{m,k+l}^{(k)} \hat{\phi}_{kml}(t), \quad \forall k.$$  

Based on the frequency domain local eigenvalues corresponding to the leading smooth dynamic FPCs, we propose a covariance stationarity test of FTS. The null hypothesis of the test, that the SD of the FTS does not vary throughout time, is investigated by evaluating whether the changes over time in the eigenvalues of the process SD are consistent with sampling variation. Following Miller and Bowman (2012), we propose a test statistic based on comparing the $m^{th}$ eigenvalue of the local SD $\hat{\lambda}_{k,m}(\theta)$ and the corresponding eigenvalue of the global SD obtained for the full FTS $\hat{\lambda}_m(\theta)$, both averaged over all frequencies $\theta$, at each time point. Using the parametric bootstrap, the null distribution of the test statistic is constructed which is then used to produce a point-wise reference band highlighting where in time there are significant deviations from the null hypothesis.

### 3 Simulation Study

To assess the performance of our proposed time-varying dynamic FPCs versus the stationary dynamic FPCs proposed by Hörmann et al. (2015) in approximating the original process, a simulation study for a variety of non-stationary data-generating processes was conducted. The simulation study was designed to mimic the EpCO$_2$ data presented in this paper.

Firstly, a FTS $\{X_k\}$ of 400 observations is generated from a functional auto-regressive of order 1, FAR(1). This simulation, in practice, is performed in a finite dimension $p$, using the basis expansion representation of the functions. The coefficients for the $p$ basis functions $z_k = (z_{k1}, \ldots, z_{kp})^\top$ associated with the functions $X_k(t) : k = 1, \ldots, 400$ are simulated according to the vector auto-regressive of order 1: $z_{k+1} = Rz_k + \epsilon_{k+1}^*$; where $R$ is the matrix of auto-regressive parameters whose norm defines the level of time dependence in the data and $\epsilon_{k+1}^*$ are i.i.d normally distributed noise with mean 0 and variance-covariance matrix $\Sigma$. This simulation is performed
based on the estimates of $R$ and $\Sigma$ obtained for the EpCO$_2$ data, assuming that the covariance structure does not vary with time. To construct a non-stationary FTS with a covariance structure that changes over time, an ordinary FPCA is performed on the FTS simulated above from the FAR(1) to obtain the eigenvalues $(\lambda_1, \ldots, \lambda_p)$ and corresponding eigenfunctions $(E_1(t), \ldots, E_p(t))$. These eigenvalues are then used to produce a $p$-dimensional vector of eigenvalues $\lambda_b$ such that both absolute and relative variance of the FPCs vary smoothly over a grid of $B$ time blocks. By naturally extending the work of Mardia et al. (1979) to a functional context, a functional process $\{X_k\}$ can be constructed by using:

$$X_k(t) = \bar{X}(t) + \sum_{m=1}^{p} S_{mk} E_m(t), \quad t \in \mathcal{T}, k = 1, \ldots, N$$

(4)

where $\bar{X}(t)$ is the functional mean, $S_{mk}$ is the score of the $m^{th}$ FPC for the $k^{th}$ observation and $E_m(t)$ is the $m^{th}$ eigenfunction. Based on this result, a sequence of locally stationary functions is generated by simulating at $b = 1, \ldots, B$, blocks of $N/B$ $p$-dimensional vectors of PC scores from a VAR(1) with a pre-specified level of dependence $\rho$ and normally distributed noise with mean 0 and variance-covariance matrix $\text{diag}(\lambda_b)$. This provides $N = 400$ functions, where $N/B$ functions share the same covariance structure. The above simulation procedure is repeated 200 times for different choices of $\rho$, reflecting weak to strong levels of dependence in the data. For each simulated non-stationary FTS $\{X_k\}$, we compute the stationary dynamic FPCs and the novel time-varying dynamic FPCs using the values $s = 10, 20, 40$ and $100$ for the smoothing parameter and estimate the corresponding scores, as per the methodology described in Section 2. These quantities are then used to recover the approximating series $\{\hat{X}_k(t)\}$ using $q = 1, 2$ and $3$ components. The performances of these approximations are then measured in terms of the normalized mean squared errors, NMSE, computed as:

$$\sum_k \|X_k(t) - \hat{X}_k(t)\|^2.$$  

(5)

Due to space limitations, we only present here the simulation results for $\rho = 0.1$ and 0.9. It is evident from Fig 1 that the (smooth) time-varying dynamic FPCs outperform the stationary dynamic FPCs in terms of NMSE. The performance of both methods improves as the dependence level in the data increases. This is a result of both methodologies accounting for the correlation structure in the data. However, as we may expect, the differences become more striking as the dependence level in the data decreases and the series becomes more dynamic. It is also noticed that the differences become negligible as the number of components $q$ used in the reconstruction increases and that using a smaller value for the smoothing parameter $s$ provides better approximations which deteriorate as $q$ increases. This justifies the trade-off between the smoothness of the FPCs over time and over-fitting.
FIGURE 1. Box-plots of the NMSE between the simulated curves and their reconstructed versions using $q = 1, 2$ and $3$ (from left to right) dynamic (red) and smooth time-varying dynamic FPCs with $s = 100$ (olive), 40 (green), 20 (blue) and 10 (purple), computed for 200 non-stationary simulation runs with $\rho = 0.1$ (top) and 0.9 (bottom).

4 Results and Discussion

The simulation results indicated that the novel non-stationary FPCs outperform their stationary counterparts in approximating the original process curves and simplifying the complexity of data in almost all settings and that improvements are more obvious as temporal dependence between curves weakens and the system becomes more dynamic. The time-varying dynamic FPCs are used to investigate and assess the dynamics and variability structure in the daily smooth profiles of EpCO$_2$ over time. The novel FPCs better approximate the original pattern as well as the within-day variability (Fig 2); the first time-varying dynamic FPC captured 94% of the variability. The proposed stationarity test identified significant changes in the covariance structure over frequency and time. The EpCO$_2$ system appears to involve high correlation throughout summer and winter, when the EpCO$_2$ daily pattern is ruled by biological activity. In transitional periods between summer and winter, the system is mostly determined by hydrological activity and therefore exhibits more variability responding to hydrological events like heavy storms or rainfall.
FIGURE 2. (a) 10 successive daily curves of de-trended EpCO₂ and the corresponding reconstructions based on (b) the first dynamic FPC and (c) the first time-varying dynamic FPC using s = 20 (chosen based on a sensitivity analysis)

5 Conclusion

Time-varying (smooth) FPCA proved to be an appropriate tool for reducing dimensionality, extracting the most important characteristics and simplifying the complexity in the variability structure of high-dimensional non-stationary time series. Traditional methods of FPCA may ignore the time dependence and non-stationarity in FTS and hence the variability modes identified may be biased.

Acknowledgments: A.E. is grateful to the University of Glasgow sensor studentship for funding. Authors are grateful to Susan Waldron for providing EpCO₂ data profiles generated through NERC Fellowship, NER/J/S/2001/00793.

References


Chain Regression Graph Models for Context-Specific independencies

Federica Nicolussi¹, Manuela Cazzaro¹

¹ University of Milan Bicocca, Italy.

E-mail for correspondence: federica.nicolussi@unimib.it

Abstract: Beyond the well known marginal and conditional independencies, the context-specific independence (CSI) is gaining attraction in the contingency table framework. This work focus on CSI statement inserted in the Hierarchical Multinomial Marginal (HMM) models context. In particular, we focus on the HMM models that can be represented through Chain Regression Graph models and we provide the conditions in order to apply, even in this context, the CSI statement. Application to real data will be provided.

Keywords: Multivariate regression model; context-specific independence; Hierarchical Multinomial Marginal models; Chain Regression Graph models; Stratified Chain Regression Graph models.

1 Introduction

The study of categorical variables under the multivariate point of view is deep dealt with in literature. The relationships of independence are often used to build a simplified model able to capture the connections among the variables. In particular, marginal and conditional independencies are widely used, while the context-specific independencies (CSIs) find a more recent application. With the term context-specific we refer to the independencies which hold only for a subset of modalities of the variables in the conditioning set. For instance, by considering three sets of variables, say $A$, $B$ and $C$, the CSI statement establishes that the variables in $A$ and in $B$ are independent given the variables in $C$ when they assume the modalities $i_C$ while the same statement does not hold when $C \neq i_C$, see for instance Højsgaard (2004). Formally

$$A \perp B|C = i_C \quad \text{and} \quad A \not\perp B|C \neq i_C. \quad (1)$$
The CSI focuses on different aspects. The former is to highlight the particular combination of modalities which discriminates between the existence or not of the independence relationship. The latter aspect simplifies the number of parameters involved in a multivariate regression models on categorical variables. In this last context, we can consider the variables having three roles: as purely responses, as purely explicatives or as mixed, that means that they are variables playing the role of covariate for a set of variables and they play as response for another set of variables. Following the approach of Marchetti and Lupparelli (2011), we take advantage of Chain Regression Graph (CRG) models to represent this regression environment, as we will describe in Section 2. In particular, we improve these models by adding labelled arcs such as Nyman et al. (2016) did with undirected graphs. These labelled arcs denote the CSIs in the graphical representation. In Section 3 we report, as it is shown in Marchetti and Lupparelli, (2011), a natural parametrization of CRG models based on the HMM models, see Bartolucci et al. (2007). Nicolussi and Cazzaro, (2017) show how to use the HMM models to represent also CSIs. The main results of these models are applied to a dataset on the innovation of Italian firms.

2 Stratified Chain Regression Graph models

Basically, CRG models take advantages from chain graphs to represent (in)dependence relationships among variables. A chain graph is a graph where the edges can be both directed or undirected and with the absence of (semi)directed cycles. In the CRG models, the vertices represent variables, the undirected arcs represent symmetrical relationships, such as the ones among a set of covariates, and directed arcs represent asymmetrical relationships, such as the one between dependent and explicative variables.

![FIGURE 1. The simple chain regression graph (a) and a chain regression graph with one labelled arc (b).]
In Figure 1 (a) is displayed an example of chain regression graph where the variable $X_1$ is a purely response variable, $X_2$ and $X_3$ play the role of explicative variables for $X_1$ and, at the same time, are response variables when we consider $X_4$ and $X_5$ as covariates. The rules to extract a list of independencies from a graph model are called Markov properties and they depend on the role of the variables involved. Following Marchetti and Lupparelli (2011) we use the so-called multivariate regression Markov properties. Similarly to Stratified Graphical models proposed by Nyman et al. (2016), we denote the CSIs through labelled arcs in the chain graph where the label denotes the modalities of the conditioning variables according to which the arc is missed. We call them Stratified Chain Regression Graph (SCRG) models. For instance, in Figure 1 (b), the arc between the vertices $X_2$ and $X_3$ reports the label $X_4, X_5 = (1, 1)$. This means that when the variables $X_4$ and $X_5$ assume respectively the modalities 1 and 1, the variables $X_2$ and $X_3$ are independent given $X_4$ and $X_5$. In symbol $X_2 \perp X_3 | X_4, X_5 = (1, 1)$. In Nicolussi and Cazzaro (2017) we provided new Markov properties to include also the CSIs.

3 Hierarchical Multinomial Marginal models

An appropriate parametrization of the CRG models is based on the HMM models, see for instance Marchetti and Lupparelli (2011). The vector of HMM parameters $\eta$ contains log-linear parameters that can be evaluated in marginal distributions by respecting the properties of hierarchy and completeness, see Bartolucci et al., (2007). Thus, each parameter $\eta^M_L(i_L)$ is described by an interaction set $L$, which contains the variables involved by the parameter evaluated in the modalities $i_L$, and a marginal set $M$, which denotes the marginal distribution where the parameter is defined. The full parametrization of the joint probability distribution, staked in the vector $\pi$ in lexicographic order, is explained from the matrix form

$$\eta = C \log(M\pi)$$

where $\eta$ is the vector of the parameters staked in the lexicographic order, $C$ is a contrast matrix which select certain logarithm of probabilities and $M$ is the matrix of marginalization which selects the marginal distribution where the parameters are evaluated. For more details see Bartolucci et al. (2007). CSIs can be represented through HMM models by imposing constraints on the sum of certain HMM parameters to be zero. These constraints depend on the type of coding we use for the variables, see for details Cazzaro and Colombi, (2014). For instance, coding all the variables by using the baseline approach, the constraints that describe the CSI in formula (1) are

$$\sum_{c \subseteq C} (-1)^{|C\setminus c|} \eta^M_L(i_v, i_c) = 0$$

(3)
where the marginal set is $\mathcal{M} = A \cup B \cup C$, $\mathcal{L} = v \cup c$, $v$ is a subset of $A \cup B$ such that $v \cap A \neq \emptyset$ and $v \cap B \neq \emptyset$, $i_v$ represents all the modalities of the variables in $v$ and $i_c$ are the modalities of the variables in $c \subseteq C$ according to the CSI statement. By considering now the multivariate regression environment, for any subset of response variables, says $r \subseteq R$ we consider the set of covariates, says $G$, and the CRG model can be represented through the HMM parameters $\eta^\mathcal{M}_\mathcal{L}(i_r|i_G)$ evaluated in the conditional distribution $r|G = i_G$

$$\eta^\mathcal{M}_\mathcal{L}(i_r|i_G) = \sum_{g \subseteq G} \beta_g(i_g)$$

where $\mathcal{M} = r \cup G$, $\mathcal{L} = r$ and $\beta_g(i_g)$ are the regression parameters. See, Nicolussi and Cazzaro (2017) for more details. Note that, including in the SCRG models also the CSIs leads to a high parameters simplification. In fact, Nicolussi and Cazzaro (2017) showed that imposing CSIs is equivalent to constrain to zero suitable $\beta$s parameters.

4 Application

We provide an application of the SCRG model, with the parametrization proposed in Section 3. At this aim we consider the survey on the innovation status of small and medium Italian enterprises during the 2010 – 2012, ISTAT (2015). The aim of this analysis is to study the effect of the innovation in some aspects of the enterprise’s life on the revenue growth without omitting the main features of the enterprise. Thus we take into account the following variables:

Var 1 **GROW**, revenue growth in 2012, (No=1, Yes=2);

Var 2 **IPR**, the innovation in products or services or production line or investment in R&D, (No=1, Yes=2);

Var 3 **IOR**, the innovation in organization system, (No=1, Yes=2);

Var 4 **IMAR**, innovation in marketing strategies, (No=1, Yes=2);

Var 5 **MRKT**, the main market (in revenue terms), (1= Regional, 2= National, 3= International);

Var 6 **DEG**, the percentage of graduate employers, (1= 0% ⊆ 10%, 2= 10% ⊆ 50%, 3=50% ⊆ 100%) ;

Var 7 **DIM**, enterprise size, (1= Small, 2= Medium).

The survey covers 18697 firms. In compliance with the aim of our analysis, we consider **GROW** as the only purely response variable, with possible covariates all the other variables. Then we consider also the **IPR, IOR**
and IMAR (the innovation variables) as response variables with covariates MRKT, DEG and DIM. With this considerations we collect the variables in chain components. We test different models (by imposing different constraints on the saturate HMM model), in order to find the best fitting one. In particular we start by testing only marginal and/or conditional independencies. Once we found a model characterized by a plausible set of independencies, we proceed to test also the CSIs. The best fitted model is represented by the chain graph in Figure 2.

The SCRG model in Figure 2 is described from two conditional independencies and one CSI. First, GROW is independent from IMAR given by IPR, IOR, MRKT, DEG and DIM. Secondly, GROW is independent from DEG given by IPR, IOR, IMAR, MRKT and DIM. Finally, according to the CSI, GROW is independent from IPR given by IOR=1, IMAR=*, MRKT=3, DEG=* and DIM=1, that is when there are no innovation in IOR when the innovation IMAR assumes any modality, when the firm works in an international market, when the percentage of graduated employers is whatever and when the firm is small. In correspondence of this model we have $df=121$, $Gsq=141.83$, $p-val=0.09$, $AIC=-192.17$, $BIC=1116.46$.

5 Conclusion

This work highlights important connection, in the study of categorical variables, between HMM models and CRG models. The graphical tool easily
represents complicate systems of relationships. The CSIs admits to consider modalities which discriminate among the dependence and the independence between variables. Furthermore, from the regression model point of view, CSIs, through labelled arcs, lead to reduce the number of parameters. However, the main limit of testing CSIs is the high number of the plausible models. In this work we did not test all plausible models but we reduced the models of interest through an exploratory analysis.

References


Handling Time Dependent Covariate Effects in Multiple Imputation: Application to Stroke in a Cox Modelling Framework.

Anna France¹, Deborah A. Costain¹, Andrew C. Titman¹

¹ Dept of Mathematics and Statistics, Lancaster University, UK

E-mail for correspondence: a.france1@lancaster.ac.uk

Abstract: Through examining risk factors for mortality post-stroke, we explore ways to appropriately handle time-varying covariate effects in the presence of missing data. By constructing a piecewise-proportional hazards model through interactions with time, non-proportionality can be accounted for. We consider multiple imputation of missing covariates, ensuring imputation models are congenial to the piecewise-proportional hazards model. In addition, we adapt standard methods of model validation to the piecewise-proportional hazards setting involving multiply imputed data sets.

Keywords: Time-varying covariate effects; Missing data; Multiple imputation; Stroke.

1 Introduction

Stroke is a serious, life threatening medical condition, with high mortality and morbidity rates (Rudd and Wolfe, 2002). The complexity and severity of stroke results in many potential risk factors for death, the effects of which may change over time. The aim of this work is to examine the risk factors for mortality post-stroke in the presence of missing data.

The data set was collected between 1996 and 2001 from an ongoing stroke audit programme. Participants were patients who were identified as having had an acute stroke and entered onto stroke registers at University Hospital Aintree or the Royal Liverpool University Hospitals, Broadgreen between January 1st and June 11th 1996. In total 538 patients were included. Audit data capture included patient demographics, characteristics of stroke, for example side of lesion, and resource use in hospital. Patients were followed up at discharge and at 3, 6, 12, 24, 36, 48, and 60 months post-stroke.
Over the 5 year period, 379 deaths were recorded. There were 38 baseline variables to be examined as risk factors for death post-stroke, some with up to 90% of intended observations missing. The prevalence and handling of missing data in research is a key focus of this work. Within previous studies which examined risk factors for death post-stroke, handling of missing data is rarely mentioned, and methods for assessing risk factors are limited and consider only univariate models, or adjust for age and gender which are often complete. We use multiple imputation using chained equations (MICE), as outlined by van Buuren (2012) and White et al. (2011), to handle the missing observations appropriately. Cox proportional hazards modelling will be used on the multiply imputed data sets to make inferences about the effect of covariates on mortality risk post-stroke, adjusted for other important risk factors.

Initial analysis highlighted issues with non-proportionality for several covariates including age, pre-stroke mobility and CT scan results. Other studies have found similar issues, but have used multiple models spanning different time periods post stroke to examine the risk factors for death, rather than adapting the Cox proportional hazards model to handle the non-proportional hazards jointly. We construct a piecewise-proportional hazards Cox model through interactions with time to account for time varying effects, and adapt the MICE framework to multiply impute missing covariates conditional upon this structure. Additionally, we consider how time dependencies affect standard methods for validating a Cox proportional hazards model, such as the Schoenfeld residuals, and explore how these methods can be appropriately adjusted to carry out model validation for multiply imputed data.

2 Methods

2.1 Piecewise-proportional hazards model

Denote the time-to-death for individual \( i \) as \( T_i \), and let \( C_i \) be the corresponding censoring time, with observation times \( Y_i = \min(T_i, C_i) \), \( i = 1, \ldots, n \). Let \( \delta_i \) denote the censoring indicator, where \( \delta_i = I(T_i \leq C_i) \), and let the covariate vector for individual \( i \) be denoted as \( X_i = (X_{i1}, X_{i2}, \ldots, X_{ip}) \). The Cox proportional hazards model is then defined as

\[
h_i(t) = h_0(t) \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}),
\]

where \( h_i(t) \) is the hazard for individual \( i \) at time \( t \), \( \beta_j \) is the coefficient for covariate \( j \) (\( j = 1, \ldots, p \)), and \( h_0(t) \) is the baseline hazard function.

Now suppose we have an additional covariate \( Z_i \), the effects of which vary across time, and let \( t_0 \) be a time point at which there is a change in effect of \( Z_i \) on survival. We can construct a piecewise-proportional hazards model
through interactions with time period within the Cox framework. This can be defined as

\[ h_i(t) = h_0(t) \exp(\beta Z_1 Z_i + \beta Z_2 I(t > t_0) Z_i + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}), \]

where

\[ I(t > t_0) = \begin{cases} 1, & \text{if } t > t_0 \\ 0, & \text{if } t \leq t_0 \end{cases} \]

is a time dependent covariate reflecting the time period. Note we now also require two censoring indicators

\[ \delta_{1i} = \begin{cases} \delta_i, & \text{if } Y_i \leq t_0 \\ 0, & \text{otherwise} \end{cases} \quad \delta_{2i} = \begin{cases} \delta_i, & \text{if } Y_i > t_0 \\ 0, & \text{otherwise} \end{cases} \]

### 2.2 Multiple Imputation using Chained Equations

Multiple imputation (MI) provides a computationally feasible approach to handling missing data within a large scope of problems under a wide range of missingness mechanisms (Carpenter and Kenward, 2013). The first stage of the MI procedure involves specifying an appropriate imputation model, which is then fitted to the data and from which posterior draws are used to fill in the missing values multiple times, say \( M \) times. Once the missing values have been imputed to create \( M \) imputed data sets, analysis can be carried out using standard, complete data techniques on each of the imputed data sets. The results of the \( M \) analyses can then be combined using Rubin’s rules (Rubin, 1976) to obtain an overall estimate and associated variance measure whilst accounting for imputation uncertainty.

MICE is a practical approach to carrying out the first stage of the MI procedure, and has the ability to handle a variety of covariate types. White and Royston (2009) outline the appropriate form for imputation models for a variety of covariate types within survival data. It is important for the imputation models to be congenial with the analysis model. In particular, it is essential that the response variable in the analysis model is included appropriately. For survival data, it is recommended that the imputation models include the Nelson-Aalen estimate of the cumulative hazard function evaluated at the follow-up time and the censoring indicator.

We further develop the theory and adapt the form of the imputation models to ensure congeniality between the imputation models and a piecewise-proportional hazards analysis model. By following similar arguments to White and Royston (2009), we show that the imputation model for a covariate \( Z \), with missing values, is a regression on the two censoring indicators \( \delta_1 \) and \( \delta_2 \), the cumulative baseline hazards associated with the time periods, \( H_0(\min(t_0, t)) \) and \( H_0(t) - H_0(\min(t_0, t)) \), and other predictive covariates \( X \). The Nelson-Aalen estimator is used to approximate the cumulative baseline hazard.
2.3 Model Validation

In order to validate the model fitted to multiply imputed data, we face two key issues: the model form of a piecewise-proportional hazards model does not allow for standard validation techniques to be used due to potential issue around scaling and residual constraints, and an overall diagnostic result over $M$ analyses is needed. Given that a piecewise-proportional hazards model is constructed to correct for non-proportionality, we focus on the Schoenfeld residuals. For a standard Cox proportional hazards model, if the proportionality assumption holds, the scaled Schoenfeld residuals plotted against some function of time should show no trend and be close to the zero line. Therneau and Grambsch (2000) developed a more formal test of proportionality which is also commonly used in practice. Both of these methods require the Schoenfeld residuals to be scaled, however given the piecewise form of our model, the scaling technique needs further consideration. We suggest a new technique for scaling the Schoenfeld residuals which takes into account their constraints in each period, and uses the Fishers information relating to each segment. We then adapt the work developed by Therneau and Grambsch (2000) to formally test the proportional hazards assumption. We test using a linear regression model of the scaled residuals against time, ensuring to account for the constraint that the residuals sum to zero, with a constant scaling on each time period and centered time, and use Rubin’s rules to give an overall test of proportional hazards over the $M$ data sets. As a visualisation tool, we plot the mean of the $M$ residuals and apply Rubin’s rules to the basis coefficients and variance components of the individual spline fits to yield a pooled smooth of the scaled Schoenfeld residuals against time.

3 Application

Initially MICE was carried out on the data to impute the missing covariate values. The imputation models are either logistic or linear regression dependent upon the type of covariate to be imputed. The imputation models were specified as outlined in Section 2.2 to ensure congeniality with the piecewise-proportional hazards model. Each of the multiply imputed data sets were analysed separately and identically. Backwards elimination via Wald tests was used to obtain the optimum combination of covariates. The coefficients in the models fitted to each of the imputed data sets were combined using Rubin’s rules to obtain a single hazard ratio for each covariate. Using $t_0 = 7$, a piecewise-proportional hazards Cox model was fitted across two time periods, 0 to 7 days post-stroke, and greater than 7 days. In order to carry out model validation, the residuals were extracted and scaled for each time period separately. For covariates in which the effects
are not time dependent, the residuals scaled by time period were combined
to test proportionality over the whole follow up time. For covariates fitted
with an interaction with time period, the scaled residuals were examined
separately for each time period to test for violation of the proportional
hazards assumption. Model validation techniques were carried out on each
imputed data set separately, and transformed appropriately to allow for
overall diagnostic results to be obtained using Rubin’s rules.

![Figure 1](image1.png)

**FIGURE 1.** Plots of scaled Schoenfeld residuals for pre-stroke mobility category
‘needs help’ against time for the Cox proportional hazards model from initial
analysis.

![Figure 2](image2.png)

**FIGURE 2.** Plots of mean scaled Schoenfeld residuals for pre-stroke mobility
category ‘needs help’ against time for piecewise-proportional hazards model with
pooled smooths (Left: Up to and including 7 days post-stroke; Right: After 7
days post-stroke).

Figure 1 and Figure 2 show the scaled Schoenfeld residuals for the ‘needs
help’ category of the pre-stroke mobility variable plotted against the Kaplan-
Meier scaling of time. Figure 1 shows the residuals for the Cox proportional
hazards model fitted to one of the multiply imputed data sets from our
initial analysis; this demonstrates the violation of the proportionality as-
sumption with evidence of a non-zero trend. Figure 2 gives a combined rep-
presentation of the residuals for the piecewise-proportional hazards model across all 10 imputed data sets; the left plot for the first time period up to 7 days post-stroke, and the right plot for after 7 days. Figure 2 shows a zero line and no trend against time for the scaled residuals for each time period, indicating that the piecewise-proportional hazards model has corrected the issues with non-proportionality.

4 Discussion

We propose methods for handling time varying covariate effects in the presence of missing data, developing an imputation framework with imputation models congenial to a piecewise-proportional hazards Cox model. Further, we develop testing procedures for validating the proportionality assumption within a piecewise-proportional hazards model fitted to multiply imputed data.

This work could be expanded to consider other diagnostic techniques and residuals. Further we could consider different types of non-proportional hazards, such as continuous time-dependent coefficients.

Acknowledgments: Special thanks to the Economic and Social Research Council (ESRC) for their financial support of this project.

References


Additive quantile mixed models with an application to longitudinal accelerometer measurements

Marco Geraci¹

¹ University of South Carolina, USA

E-mail for correspondence: geraci@mailbox.sc.edu

Abstract: Additive models are flexible regression tools that include linear as well as nonlinear (smooth) terms. Additive mixed models extend additive models to include random terms when the data are sampled according to cluster designs (e.g., longitudinal). In this paper, we propose a novel additive mixed model for quantile regression. Our methods are motivated by an application to physical activity using a dataset with more than half million accelerometer measurements. In a simulation study, we assess the proposed methods against existing alternatives.

Keywords: Linear quantile mixed models; Nonlinear effects; Random effects; Shrinkage; Smoothing.

1 Introduction

The benefits of regular physical activity (PA) on well-being and life expectancy as well as the detrimental health effects of sedentary behavior have been amply documented. In England, physical inactivity is estimated to cost more than eight billion pounds a year (National Institute for Health and Clinical Excellence). This includes both the direct costs of treating major, lifestyle-related diseases and the indirect costs of sickness absence. Establishing an active lifestyle at an early age is therefore an important form of prevention against morbidity and premature mortality.

The plot in Figure 1 shows accelerometer counts by time of the day in seven-year-old children of the Millennium Cohort Study (MCS), a UK-wide longitudinal multi-purpose survey. Our sample comprised 1154 English children for whom accelerometer counts were aggregated over 10-minute intervals between 7:00 and 20:00 (thus producing 79 time points), for seven days of the week. In total, this gave 638,162 accelerometer measurements (that is,
79 × 7 × 1154). During weekdays there are periods of higher PA levels that mirror traveling times to and from school, and lunch and break times (Sera et al, 2017). However, temporal trajectories at different quantile levels of the conditional distribution are not simply vertical shifts of one another. This suggests that the scale and possibly the shape of the counts distribution change with time of the day. In our analysis, we considered several predictors of PA pertaining to the socio-demographic, behavioral, and temporal domains. We also considered nonparametric terms for time of the day and body mass index.

![Graphs showing accelerometer counts over time for weekdays and weekend days.](image)

FIGURE 1. Accelerometer counts observed in 1154 English children of the UK Millennium Cohort Study, by days of the week (Monday through Friday, weekdays; Saturday and Sunday, weekend). Solid lines represent conditional quantile functions fitted with additive quantile mixed models, averaged over the distribution of the covariates, for 4 quantile levels (0.1, 0.5, 0.9, 0.95).

2 Methods

2.1 The model

We consider data from two-level nested designs in the form \((x_{ij}^T, z_{ij}^T, y_{ij})\), for \(j = 1, \ldots, n_i\) and \(i = 1, \ldots, M\), where \(x_{ij}^T\) is the \(j\)th row of a known \(n_i \times p\) matrix \(X_i\), \(z_{ij}^T\) is the \(j\)th row of a known \(n_i \times q\) matrix \(Z_i\) and \(y_{ij}\) is the \(j\)th observation of the response vector \(y_i = (y_{i1}, \ldots, y_{in_i})^T\) for the \(i\)th cluster. The quantile level is denoted by \(\tau\), \(\tau \in (0, 1)\).

We define the following \(\tau\)th additive quantile mixed model (AQMM)

\[
Q_{y_{ij} \mid u_i, x_i, z_i}(\tau) = \beta_{\tau, 0} + \sum_{k=1}^{p} g_{\tau}^{(k)}(x_{ijk}) + z_{ij}^T u_{\tau, i},
\]

\(j = 1, \ldots, n_i, i = 1, \ldots, M\), where \(g_{\tau}^{(k)}\) is a \(\tau\)-specific, centered, twice-differentiable, smooth function of the \(k\)th component of \(x\). The \(q \times 1\) vector \(u_{\tau, i}\) collects cluster-specific random effects associated with \(z_{ij}\) and its distribution is assumed to depend on a \(\tau\)-specific parameter. Without loss of
generality, let the components of \( x \) be ordered in such a way that the first \( s \) terms of the summation in (1) are nonlinear functions and the remaining \( p - s \) are linear.

To model nonlinear functions, we consider a spline model of the type
\[
g_{\tau}(x) \approx \sum_{h=1}^{H} v_{\tau,h} B_h(x), \quad \text{(e.g., cubic or B-Spline),}
\]
where the \( B_h \)'s and \( v_{\tau,h} \)'s, \( h = 1, \ldots, H \), denote, respectively, the basis functions and the corresponding coefficients, and \( H \) depends on the degrees of freedom or the number of knots. The quantile function in (1) is then approximated by
\[
Q^{\ast}_{y_{ij}\mid u_{i}, x_{i}, z_{i}^{\tau}}(\tau) = \beta_{\tau,0} + \sum_{k=1}^{s} \sum_{h=1}^{H} v_{\tau,hk} B_{h}^{(k)}(x_{ijk}) + \sum_{k=s+1}^{p} \beta_{\tau,k} x_{ijk} + z_{ij}^{T} u_{\tau,i}. \quad (2)
\]

We introduce a mixed-effects representation (Ruppert et al, 2003; Wood, 2006) by assuming that the vectors of the spline coefficients \( v_{\tau,k} \) are random with variance-covariance \( \phi_{\tau,k} I \). Further, we assume that the \( u_{\tau,i} \)'s are independent for different \( i \) (but may have a general covariance structure \( \Sigma_{\tau} \)) and are independent from \( v_{\tau} = (v_{\tau,1}^{T}, \ldots, v_{\tau,s}^{T})^{T} \). Our objective function is then given by
\[
\sum_{i=1}^{M} \rho_{\tau}(y_{i} - F_{i} \beta_{\tau} - Z_{i} u_{\tau,i} - B_{i} v_{\tau}) + \sum_{i=1}^{M} \| u_{\tau,i} \|_{\Sigma_{\tau}^{-1}}^{2} + \sum_{k=1}^{s} \phi_{\tau,k}^{-1} \| v_{\tau,k} \|^{2}, \quad (3)
\]
where \( \rho_{\tau}(r) = \sum_{j=1}^{n} \rho_{\tau}(r_{j}) \) and \( \rho_{\tau}(r) = r \{ \tau - I(r < 0) \} \) is the quantile regression check function. The \( L_{2} \)-penalty represents a suitable choice for modeling smooth changes in energy expenditure over time. Note that the \( \phi_{\tau,k} \)'s determine the amount of smoothing for the nonparametric terms.

The objective function (3) is that of a three-level linear quantile mixed model (Geraci and Bottai, 2014) and is maximized using a second order Laplacian approximation as detailed in Geraci (2018). Standard errors are computed via bootstrap. Due to the large sample size of the MCS accelerometer dataset, we used the method by Kleiner et al (2014).

### 2.2 Implementation and simulation study

These methods were implemented in the R package \texttt{aqmm} as an add-on to the package \texttt{lqmm} (Geraci, 2014). The add-on is currently available from the author’s website (https://marcogeraci.wordpress.com) and will appear in a future release of the main package. The core function made use of routines available from the \texttt{mgcv} and \texttt{nlme} packages. In a simulation study, we assessed the performance of AQMM as compared to existing alternatives (Wood, 2006; Fenske et al, 2013). The simulation study and the MCS accelerometer data analysis were performed on a 64-bit computer with a 3.60GHz quad core i7-4790 processor and 32 gigabytes of RAM.
3 Main results and conclusions

The simulation study showed that AQMM performed well as compared to alternative approaches. It provided advantages in terms of bias and mean squared error for both predictions and linear regression coefficients. In the analysis of the MCS accelerometer data, AQMM gave a complete picture of fixed and random effects showing that, for example, girls tend to have lower PA than boys especially at higher quantiles, and that individual heterogeneity in the trajectories is stronger at the median and higher quantiles. Figure 1 shows the resulting smoothed PA trajectories at different quantile levels by day of the week.

Acknowledgments: This research has been supported by the National Institutes of Health – National Institute of Child Health and Human Development (Grant Number: 1R03HD084807-01A1).

References


Spatio-temporal modelling of remote-sensing lake surface water temperature data

Mengyi Gong¹, Claire Miller¹, Marian Scott¹

¹ University of Glasgow, Scotland, U.K.

E-mail for correspondence: mengyi.gong@glasgow.ac.uk

Abstract: Remote-sensing technology is widely used in environmental monitoring. The coverage and resolution of satellite based data provide scientists with great opportunities to study and understand environmental change. However, the large volume and the missing observations in the remote-sensing data present challenges to statistical analysis. This paper investigates two approaches to the spatio-temporal modelling of remote-sensing lake surface water temperature data. Both methods use the state space framework, but with different parameterizations to reflect different aspects of the problem. The appropriateness of the methods for identifying spatial/temporal patterns in the data is discussed.

Keywords: state space; dimension reduction; FPC; STRE.

1 Introduction

The remote-sensing lake surface water temperature (LSWT) data are measured by the Advanced Along-Track Scanning Radiometer (AATSR), on board the European Space Agency’s Envisat platform. The retrieved LSWT data can be accessed from [http://www.ed.ac.uk/arclake/data.html](http://www.ed.ac.uk/arclake/data.html), the ARC-Lake v3.0 database. The observation period is from June 1995 to April 2012 and the spatial resolution is 0.05° × 0.05°. Ecologists are interested in the spatial/temporal patterns in the data to better understand the dynamics of the environmental system, for example, as part of the GloboLakes project ([www.globolakes.ac.uk](http://www.globolakes.ac.uk)).

The specific data set investigated here is the monthly LSWT of Lake Victoria. The data are stored in an array of dimension 65 × 66 × 203 (longitude by latitude by time). Although the number of observations cannot be regarded as ‘large’ in the sense of ‘big data’, this 3-dimensional dataset can still be challenging for statistical analysis. Therefore, spatio-temporal
models exploiting dimension reduction are investigated. Specifically, a functional data representation is used to reduce the data dimensionality and hierarchical dynamic spatio-temporal models (DSTM) constructed under the state space framework (Cressie & Wikle, 2011) are used to describe the spatio-temporal features of the data. Due to technical and atmospheric issues, remote-sensing data can have a substantial amount of missing data (see the data panels of Figure 1). Therefore, modifications are proposed here in the estimation algorithms to account for missing data.

2 The modelling framework

Consider a hierarchical DSTM in state space form with, (1) a data model relating the observation \( Z(s; t) \) to a ‘true’ spatio-temporal process \( Y(s; t) \) and (2) a process model describing the dynamics of the latent process through lagged dependence,

\[
Z(s; t) = Y(s; t) + \epsilon(s; t) = \Phi(s)\beta_t + \zeta(s; t) + \epsilon(s; t), \tag{1}
\]

\[
\beta_t = \sum_q M_q\beta_{t-\tau_q} + u_t. \tag{2}
\]

Dimension reduction comes from the representation of \( Y(s; t) \) as \( \Phi(s)\beta_t + \zeta(s; t) \), where \( \Phi(s) \) is a basis matrix and \( \zeta(s; t) \) is a non-dynamic component. If matrix \( \Phi(s) \) is of a lower rank \( K \) than the dimension of the data vector \( Z(s; t) \), being \( N \), then the dimension of the process model (2) would be reduced to \( K \). Substantial computational gains may be achieved if \( K \ll N \), which is often the case for remote-sensing data. The process dynamic follows a vector auto-regressive (VAR) model of order \( q \), reflecting the temporal dependence of the spatial process \( Y(s; t) \). To ensure identifiability, a parameter model putting constraints on model components may be included. Two parameterizations of this framework are proposed and investigated here. Both methods offer opportunities for dimension reduction, but with different emphases, where one focuses on the general spatio-temporal pattern and the other focuses on the spatio-temporal prediction.

2.1 The FPC parameterization

The functional principal component (FPC) parameterization is based on the empirical orthogonal function (EOF) parameterization in Xu & Wikle (2007). It maps the data to the leading EOFs extracted from the data,

\[
Z(s, t) = \sum_{p=1}^a \xi_p(s)\alpha_{pt} + \epsilon(s, t) = \Xi(s)\alpha_t + \epsilon(s, t), \tag{3}
\]

and parameterizes the covariance matrix of the vector of residuals \( \epsilon_t \) as \( \sigma^2I + \sum_{p=a+1}^P \lambda_p\xi_p\xi_p^\top \), where \( \xi_p \) is the vectorized EOF \( \xi_p(s) \) evaluated at
all locations $s$ and $\lambda_p$ is the corresponding eigenvalue. The process model is specified as a first order VAR, which is $\alpha_t = M\alpha_{t-1} + u_t$. Exploratory analysis suggests that first order dependence is appropriate for the LSWT data after removing the trend and seasonality. The component $\zeta(s; t)$ is not considered here. In the case of the remote-sensing LSWT data, functional representation is applied and functional PCs are extracted as the analogous of the EOFs for dimension reduction. Model estimation uses the EM algorithm with the Kalman filter/smoother. Model results provide information on spatial patterns through the functional PCs and temporal evolutions through the estimated process model.

2.2 The STRE model

The spatio-temporal random effect (STRE) model of Cressie et al., (2010) can be written in the same way as formulae (1) and (2). In particular, $\Phi(s)$ is usually taken to be a spatial basis and the component $\zeta(s; t)$ represents the non-dynamic random effect unique to each spatial image $Z(s; t)$, which cannot be captured by the dynamic of $Y(s; t)$. Again, dimension reduction can be achieved through a basis representation $\Phi(s)\beta_t$. Model estimation uses the EM algorithm, along with the fixed rank filter (FRF) and smoother (FRS) (Cressie et al., 2010; Katzfuss & Cressie, 2011). This method estimates the time-varying $\beta_t$ using the Kalman filter/smoother and the random effect $\zeta_t$ through a second filter based on the conditional distribution of $(\zeta_t, \beta_t) | Z_{1:t}$ (FRF) and $(\zeta_t, \beta_t) | Z_{1:T}$ (FRS), where $Z_{1:t}$ represents observed data $\{Z_1, \ldots, Z_t\}$. In particular, it is assumed that the non-dynamic component $\zeta_t$ only depends on the information of time point $t$. Temporal patterns can be extracted from the estimated process model. Spatial patterns may be modelled by assigning a correlation structure to the residual covariance matrix.

2.3 Implementation and results

The FPC parameterized model and the STRE models are then applied to a subset of the Lake Victoria LSWT data (dimension $36 \times 47 \times 202$). The subset is taken to minimize the influence of land pixels and lake border pixels, which tend to have larger uncertainties. As the data are stored on a regular grid and that shape is not critical to this analysis, a tensor spline basis is specified for the spatial basis $\Phi(s)$. A basis accounting for the shape of the lake may be used, but the implementation would require a lot more computational cost. The smoothness (degrees of freedom) of the basis is controlled directly by the number of knots of the tensor splines. This is out of the concern for the computational cost of tuning a smoothing parameter. Information criteria AIC and BIC are used to select the degrees of freedom. The variance proportion criterion is applied to select the number of FPCs in model (3). In this case, an 80% threshold gives 11 FPCs in
the dynamic component of the model; a 95% threshold selects a further 11 eigenfunctions to form the residual covariance matrix. A random walk model is used for modelling the process dynamic, i.e. $\alpha_t = \alpha_{t-1} + u_t$ for the FPC parameterized model and $\beta_t = \beta_{t-1} + u_t$ for the STRE model. This is appropriate considering the feature of the LSWT data after removing a seasonal mean.

The R code for implementing the two methods has been developed as part of this work. To accommodate missing data, the approach similar to the Kalman filter for sparse data (Shumway & Stoffer, 2006) is adopted. The implementation of FRF and FRS follows the procedure described in Katzfuss & Cressie (2011). In extreme circumstances where there are only a few observations available for a spatial image, a filtering threshold may also be applied to avoid over-interpolation.

The fitted LSWT images constructed using the smoothed $\alpha_t$ or $\beta_t$ and MLEs from the converged EM algorithms of the two models are shown in Figure 1. Both methods provide a good fit to the data. The residual sum of squares (RSS) from the FPC parameterized model is 0.1021; that of the STRE model is 0.0810. The fitted images from the FPC parameterized model are smoother, as the model is designed to capture the general patterns. The results from the STRE model show more detail, as the model is designed for interpolation and prediction. The contrast in the fitted images of the FPC parameterized model appears to be larger than that of the STRE model. In terms of very sparse images, imputation with smaller contrast maybe preferred to avoid over-interpolating the unobserved areas.

FIGURE 1. Examples of the Lake Victoria LSWT data and their fitted versions using the FPC parameterized model (upper) and the STRE model (lower).
An investigation can also be carried out on the smoothed process state to understand the patterns of the process dynamics. For the FPC parameterized model, the smoothed states reflect the temporal evolution of the corresponding FPCs. An example of the smoothed $\alpha_{1t}$ and $\alpha_{2t}$ time series from model (3) is given in Figure 2. In this case, $\alpha_{1t}$ seems to be showing certain seasonal fluctuations not covered by the seasonal mean; whereas $\alpha_{2t}$ displays mainly random fluctuation with a few spikes. The dynamics of $\beta_{t}$ in the STRE model (2) might be less straightforward to interpret, as they are spatial basis coefficients that do not always have a clear meaning in practice. Nonetheless, the time series of $\beta_{t}$ may still be useful to aid with the understanding of spatio-temporal patterns in the data.

![FIGURE 2. Examples of the dynamics of the FPCs from model (3).](image)

Finally, the model residuals are investigated. The images in Figure 3 present the pixel-wise RSS from the two models, reflecting the regional fit of the models as opposed to the overall fit. The pixels towards the northwest corner appear to have larger RSS values. However, there is not any big discrepancies between the RSS of different pixels, suggesting that the two methods are appropriate for providing missing data imputations despite the varying data availability in different areas of the lake. Katzfuss & Cressie (2011) also derived the formula of the mean squared prediction errors (MSPE) for the STRE model (1), which are defined as the diagonal elements of $E[(Y_{t} - Y_{t|T})(Y_{t} - Y_{t|T})^\top]$, where $Y_{t}$ is the vectorized $Y(s; t)$ and $Y_{t|T}$ is the FRS version of $Y_{t}$. In this case, the spatial pattern of the MSPEs corresponds to that of the missing percentages, but again, the values are at a similar scale.

3 Discussion

The FPC parametrized model and the STRE model provide two efficient approaches to the spatio-temporal modelling of the sparse high-dimensional remote-sensing data. Missing data imputation can be carried out while the spatial and temporal patterns are extracted. One criticism on the EOF (i.e. FPC in this case) based method is that the leading principal components may not be adequate in explaining the dominant system dynamics, despite
FIGURE 3. The pixel-wise RSS from the Lake Victoria LSWT data fitted using the FPC parameterized model (left) and the STRE model (right).

their power in describing the variation in the data. For the STRE model, the random component $\zeta_t$, while accounting for the individual effect unique to each spatial image, cannot provide a conclusive summary of the spatial variation. Unless the residual spatial structure is modelled, which could be computationally expensive, it is difficult to use the STRE model to identify the spatial variation patterns. Potential developments may be to parameterize the random effect $\zeta_t$ to reflect certain spatial patterns in a flexible manner. This will be investigated in the future to improve the modelling of the remote-sensing environmental data.

Acknowledgments: The authors are grateful to the ARC-Lake project for access to the data. The authors were partly funded for this work through the NERC GloboLakes project (NE/J022810/1).

References


Could bigger be better? Longer Atlantic salmon smolts seem more likely to return as adults

Stephen D. Gregory

1 Salmon & Trout Research Centre, Game and Wildlife Conservation Trust, UK

E-mail for correspondence: sgregory@gwct.org.uk

Abstract: Atlantic salmon stock declines are thought to be due to climate-driven changes in the marine environment that have rendered it hostile to migrating smolts. Recently, there has been a growing sense that factors affecting smolts during their development, i.e., in the freshwater environment, might play a larger role than previously judged. It has been hypothesised that smolt size is related to their marine survival, i.e., that larger smolts are better able to survive at sea than their smaller counterparts. With a large database of individual smolt capture histories and characteristics, I test the "bigger is better" hypothesis for Atlantic salmon smolts inhabiting the river Frome in Dorset, UK. By fitting and comparing multi-state mark-recapture state-space models with different covariates, I show support for the "bigger is better" hypothesis in these data, i.e., that longer smolts are more likely to return as adults compared to smaller smolts. This suggests that freshwater environments could be managed to maximise smolt quality (e.g., length) and thereby the numbers of returning adults.

Keywords: multi-state mark-recapture state-space model; smolt length; individual covariates; group covariates.

1 Introduction

Atlantic salmon *Salmo salar* stocks are declining (ICES, 2017). For a long time, the finger of blame has been pointed at the marine environment, where climate change and the processes it influences, such as spatial and temporal variation in algal blooms (Beaugrand and Reid, 2012), are thought to have rendered the environment hostile to seaward-migrating juveniles, known as smolts. More recently, however, there has been a growing sense that factors affecting juveniles during their development, i.e., in the fresh-
Myriad extrinsic and intrinsic factors, such as water temperature and body length (respectively), are thought to affect juvenile salmon development (for a review, see Russell et al., 2012). Among intrinsic factors, it has been long postulated that smolt mortality is inversely related to their body size, i.e., the inverse-weight hypothesis (Ricker, 1976). Many studies have provided some empirical evidence testing this "bigger is better" paradigm, which are reviewed in Gregory et al., (2018). For example, Armstrong et al. (2018) used individual-level data, together with covariates and their interactions, to show that the probability of adult salmon return was related to smolt size. However, in a review of studies that assessed the effect of salmonid smolt length on their subsequent survival, Gregory et al. (2018) found limited support for the "bigger is better" hypothesis.

In this paper, I use capture histories and characteristics of over 1500 smolts to parameterise multi-state mark-recapture state-space models, admitting individual- and group-level covariates, and compare their fits using model selection to measure the support for an effect of Atlantic salmon smolt length on their probability to return as an adult.

2 Methods

Since 2011, the Game and Wildlife Conservation Trust has been monitoring the seaward migration and adult return of a large sample of Atlantic salmon smolts on the river Frome in Dorset, UK, using Passive Integrated Transponders (PIT), an array of PIT detection antennae and a rotary screw trap (RST) (Figure 1). Each autumn, approximately 10,000 juvenile salmon are marked with a PIT tag and returned to their place of capture. Throughout the following spring, an RST is operated 24h per day to capture and measure a sample of emigrating smolts and monitor changes in their lengths. All smolts are processed within 30mins of capture and returned to the river ≈ 50m downstream of the RST within an hour of capture. All of the aforementioned procedures are carried out under a UK Home Office project license. Adult returns are collected passively on two PIT antenna approximately 3km apart, which are used as primary detection and secondary re-detection devices in a mark-recapture experiment to estimate the probability of adult return while accounting for imperfect detection. Approximately 97% of river Frome smolts emigrate one year after emerging from their gravel nests, and I limited samples to these individuals to reduce any potential effect of different life strategies of older (and larger) smolts.

Using these data, I parameterised multi-state mark-recapture state-space models (SSM) to explore the effect of smolt length on the probability to return as an adult. State-space models have the advantages of separating
Longer salmon smolts more likely to return

the ecological process from the observation process. In this way, I could investigate the effect of length on the ecological process, i.e., survival, while making no assumptions about the observation process, i.e., detection. Moreover, when specified in a Bayesian language, such as BUGS or stan, SSMs are highly flexible and admit many types of covariates as fixed or random effects.

In the simple case of estimating the effect of individual smolt length on its probability to return as an adult, then:

\[
y(i,t) | z(i,t) \sim \text{Bernoulli}(z(i,t)p)
\]

\[
z(i,t+1) | z(i,t) \sim \text{Bernoulli}(z(i,t)\varphi_i)
\]

where \( t > 0 \), \( z \) is a latent variable describing the state of smolt \( i \) at time \( t \), \( \varphi_i \) is the survival rate of smolt \( i \) from state \( z(i,t) \) to state \( z(i,t+1) \) and \( y \) is the observation of smolt \( i \) given the probability \( p \) of detecting it. From these equations, it can be noted that \( \varphi_i \) and \( p \) are time-invariant and \( p \) does not vary for individuals. To estimate the effect of smolt \( i \) length \( l_i \) on its survival, \( \varphi_i \) is specified as a deterministic function of logistic regression parameters:

\[
\text{logit}(\varphi_i) = \alpha + \beta l_i
\]

where \( \alpha \) is the estimated probability to return as an adult for any smolt
returning to the river and $\beta$ is the effect of smolt length on $\alpha$ while accounting for imperfect detection, i.e., $1 - p$.

I parameterised three SSMs: (1) a null model equivalent to Equation 1 with no effect of year or length on the probability to return as an adult; (2) a year model that included an effect of year, represented as annual deviations from the overall probability to return as an adult; and (3) a length model including both an effect of year and an overall effect of length on the probability to return as an adult.

Parameter values were estimated using Stan (http://mc-stan.org/) through the R interface in package rstan. Parameter values for each SSM were estimated using 3 parallel chains run for 2000 iterations including a 1000 iteration warmup that was discarded. All parameters were assigned weakly informative priors, i.e., Gaussian(0, 1) or Uniform(0, 100). For more details on estimation using Stan, see Stan Development Team (2017).

### 3 Results

Comparison of the SSM fits revealed the best support for the length model including a length effect and fixed year deviations (scaled approximate leave-one-out cross-validation estimate [looic]: 496.57), followed by the year model including fixed year deviations only (looic: 501.42) and then the null model omitting both terms (looic: 504.98). Moreover, the estimated parameter values and their lower and upper credible intervals were comparable between models (Table 1). Examination of fit diagnostics, include trace and autocorrelation plots, confirmed that all SSM fits were well-behaved.

<table>
<thead>
<tr>
<th>parameter</th>
<th>null</th>
<th>year</th>
<th>length</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$</td>
<td>0.034 (0.025 - 0.048)</td>
<td>0.029 (0.013 - 0.067)</td>
<td>0.026 (0.011 - 0.060)</td>
</tr>
<tr>
<td>$\psi$</td>
<td>0.730 (0.526 - 0.953)</td>
<td>0.717 (0.524 - 0.934)</td>
<td>0.725 (0.512 - 0.944)</td>
</tr>
<tr>
<td>$p$</td>
<td>0.714 (0.530 - 0.862)</td>
<td>0.720 (0.537 - 0.863)</td>
<td>0.721 (0.542 - 0.852)</td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td></td>
<td>0.449 (0.117 - 0.792)</td>
</tr>
</tbody>
</table>

My results suggest that an individual smolt’s probability to return as an adult is a function of its length; the length model received substantially higher empirical support compared to the alternative models examined that omitted a length effect, and the 97.5% credible interval of the length effect parameter $\beta$ that did not intercept 0 (Figure 2a).

It seems that the larger you are, the higher your probability to return as an adult. Moreover, the effect is not small: a smolt of approximately 16 cm
is between 3 and 4 times more likely to return as an adult compared to 12 cm smolt (Figure 2b).

**FIGURE 2.** (a) Posterior density plots computed from posterior draws with all chains merged. The shaded areas under the curves highlight the 50% uncertainty range and the distributions are trimmed to the 95% uncertainty range. (b) Fitted effect of length on probability that an individual smolt will return as an adult. Grey ribbon represents the 50% uncertainty range of the fit. Note that length is standardised by subtracting the mean and dividing by 1 standard deviation. Dashed red lines highlight the probability that a 12 and 16 cm smolt will return as an adult.

### 4 Discussion

A long-standing prevailing hypothesis is that "bigger is better" among Atlantic salmon smolts, i.e., that larger smolts will have a higher marine survival and therefore a higher probability to return to their natal river as an adult to spawn. This hypothesis has, however, been untested, in large part because of insufficient or inadequate data to test it robustly. Here, I show that smolt length is indeed related to the probability to return as an adult in these data. It seems that longer smolts are between two and three times more likely to return as an adult than shorter smolts. Aside from suggesting that bigger is indeed better among Atlantic salmon smolts, this finding has potentially important implications for their management; it is far easier to manage their freshwater habitats, i.e., rivers, than it is to manage their marine habitat, i.e., the nearshore coast or high sea. Accordingly, these findings suggest that we might manage smolt freshwater habitats to maximise their condition, here represented as length, to maximise the probability to return as an adult, i.e., to bolster declining adult salmon stocks.

An ability to improve the probability that an individual smolt will return as an adult could be particularly important in light of the observation
that Frome smolts (and possibly smolts elsewhere), like Atlantic salmon juveniles (Gregory et al., 2017) appear to be getting shorter. We must, however, be cognisant of alternative factors; it is unlikely that their length at emigration is the only factor influencing their probability of adult return: marine conditions are undoubtedly deteriorating and the timing of salmon migrations are changing.

Acknowledgments: Special thanks to all the people who helped collect the data used in this study, including present and past members of the GWCT Fisheries Department, Cefas and the many volunteers who have given up their time to help the imperiled salmon population of the river Frome.

References


Multilevel modelling with level 2 missing covariates: the relationship between student ratings and teacher beliefs and practices

Leonardo Grilli¹, Maria Francesca Marino¹, Omar Paccagnella², Carla Rampichini¹

¹ Department of Statistics Computer Science Applications ‘G. Parenti’, University of Florence, Italy
² Department of Statistical Sciences, University of Padua, Italy

E-mail for correspondence: carla.rampichini@unifi.it

Abstract: We analyse the relationship between student ratings of university courses and several characteristics of the student, the course and the teacher. In particular, we exploit data from a survey collecting information about teacher beliefs and practices at the University of Padua in academic year 2012/13. Student ratings are nested into classes, calling for multilevel modelling. However, due to survey non-response, the information about beliefs and practices is missing for about half of the teachers, posing a serious issue of missing data at level 2. To avoid listwise deletion, we make multiple imputation via fully conditional specification, exploiting information at all hierarchical levels. The proposed approach turns out to be effective. From a substantive point of view, some of the considered teacher beliefs are found to be significantly related to student ratings.

Keywords: fully conditional specification; multiple imputation; multivariate mixed model.

1 Case study

We analyse student satisfaction, as measured by student evaluation of teaching (SET). The peculiarity of the study lies in the availability of many variables about teacher characteristics and beliefs, and teaching practices. This work exploits data from the University of Padua for academic year 2012/13 about bachelor degree courses. The data set is obtained by merging three different sources: (i) the traditional SET survey with 18 items, measured on a ten-point scale (1: low, 10: high); (ii) administrative data

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
on students, teachers and didactic activities; (iii) a survey carried out by the PRODID project on teacher beliefs and practices. The data have a two-level hierarchical structure, with 56,775 student ratings at level 1 and 1,016 classes at level 2. The average class size is 79 (min 5, max 442).

We are interested in student satisfaction about two key aspects of teaching, i.e. teacher ability to involve students (item D06 of the SET questionnaire) and teacher clarity (item D07).

The analysis is based on the following bivariate two-level linear mixed model for item \( m \) (\( m: 1 \) for D06, 2 for D07) recorded on student \( i \) in class \( j \):

\[
Y_{mij} = \alpha_m + \beta_m' x_{ij} + \gamma_m' w_j + u_{mj} + e_{mij}
\]

(1)

where \( x_{ij} \) is the vector of student covariates (level 1), and \( w_j \) is the vector of teacher and class covariates (level 2). Level 1 errors \( e_{mij} \) are assumed independent across students. Level 2 errors (random effects) \( u_{mj} \) are assumed independent across classes, and independent from level 1 errors. We make standard assumptions for the distributions of the model errors, including homoscedasticity (within each outcome) and normality. Therefore, the response vector \( Y_{ij} = (Y_{1ij}, Y_{2ij})' \) has residual variance-covariance matrix \( V ar(Y_{ij}) = \Sigma_u + \Sigma_e \), where \( \Sigma_u \) and \( \Sigma_e \) are the covariance matrices of the errors at level 2 and level 1, respectively.

The survey on teacher beliefs and practices has about fifty percent of missing questionnaires, posing a serious issue of missing data at level 2. An analysis based on list-wise deletion would discard the entire set of student ratings for non responding teachers, causing two main problems: (i) a dramatic reduction of the sample size, and thus of the statistical power, and (ii) possibly biased estimates if the missing mechanism is not MCAR.

2 Handling missing data at level 2

In multilevel models, the treatment of missing data requires special techniques. In fact, the data have a hierarchical structure and, thus, missing values can be at any level of the hierarchy. Moreover missing values can alter the variance components and the correlations. Multiple imputation (MI) is the most flexible approach to missing data. MI has been extended to the multilevel setting to deal with these special issues, following two main approaches (Mistler and Enders, 2017; Grund et al., 2018): fully conditional specification, also known as multivariate imputation by chained equations, and joint modelling.

In our case study, the substantive model (1) is multilevel, however missing data are only at level 2. This feature makes the imputation simpler than in the general multilevel setting. Indeed, we can apply standard MI techniques to level 2 data, then merge level 1 and level 2 data to obtain complete datasets.
However, the MI step is challenging, since we have to impute many categorical variables with a high percentage of missing. In particular, about 50% of the teachers did not respond to the whole questionnaire, producing missing values on 10 binary items (teacher practices) and 20 ordinal items (teacher beliefs on a seven-point scale).

The imputation model at level 2 should include all the level 2 covariates and information on level 1 variables, in particular the response variable. Several ways are possible to summarize information from level 1 variables (Erler et al., 2016; Grund et al. 2017). We choose the cluster mean, which is effective in general and easy to implement in our case, where level 1 variables (including the response) are completely observed.

We perform multivariate imputation by chained equations (MICE). In our case, the imputation model is binary logit for the 10 binary items (teacher practices), and cumulative logit for the 20 ordinal items (teacher beliefs). The imputation model of a given item includes the following covariates: the items already imputed, the fully observed class and teacher characteristics, the cluster means of level 1 variables (covariates and outcomes), and the cluster size.

3 Results

The bivariate two-level model (1) is fitted by maximum likelihood on ten imputed data sets, and the results are combined with the standard MI rules. The analysis is conducted using the *gsem* and *mi* commands of Stata, version 15 (StataCorp, 2017).

We first fit the model without covariates, to explore the correlation structure of the two outcomes. We find out that the ICC is about 30% for both outcomes, i.e. teacher ability to involve students and teacher clarity. The two outcomes are highly correlated (0.83), especially at level 2 (0.933).

Then, we add the available covariates. In particular, teacher practices are included as binary indicators, while teacher beliefs are summarized into 6 indicators averaging the relevant seven-point ordinal items, i.e. *passion for teaching* (2 items), *passion for research* (2 items), *need for teaching support* (4 items), *care about student needs* (3 items), *role of active learning* (4 items), *interest in innovative teaching methods* (3 items).

The final model has a total of 6 student characteristics and 22 covariates at the second level (5 class variables, 3 teacher characteristics, 8 teacher practices and 6 teacher beliefs).

To quantify the missing data influence on the sampling variance of a parameter estimate, we can consider the fraction of missing information (FMI), i.e. the ratio of the sampling variance between imputations over the total sampling variance. For imputed covariates FMI ranges from 0.15 to 0.68, with a median value of 0.44. The fraction of missing data is about 0.5; therefore, for the majority of imputed covariates the trade-off between the
sampling error inflation due to MI and its reduction due to sample size increase is favourable.
The main substantive finding is that teacher practices and beliefs from the PRODID survey are significantly related with SET ratings. In particular, the practices *active learning* and *contribution from experts* are positively related to the outcomes. As for teacher beliefs, *passion for teaching* and *interest in innovative teaching methods* raise the ratings, while *need for teaching support* reduces the ratings.
To evaluate the robustness of the results, it is advisable to apply alternative imputation methods. We tried joint modelling (JM) imputation, using the R package *jomo* (Quartagno and Carpenter, 2017), which allows to impute binary and ordinal data through latent normal variables. However, in our case with 10 binary and 20 ordinal variables this approach is not feasible due to the computational burden, unless we treat the ordinal variables as continuous. We plan to study how the JM approach can be effectively implemented in our challenging setting, in order to compare the performances of MICE and JM.

References


StataCorp. (2017). *Stata: Release 15. Statistical Software*. College Station, TX: StataCorp LLC
Boosting Methods for Effects Selection in Cox Frailty Models

Andreas Groll\textsuperscript{1}, Trevor Hastie\textsuperscript{2}, Thomas Kneib\textsuperscript{3}, Gerhard Tutz\textsuperscript{4}

\textsuperscript{1} Technische Universität Dortmund, Faculty of Statistics, Vogelpothsweg 87, 44227 Dortmund, Germany
\textsuperscript{2} University of Stanford, Department of Statistics, 390 Serra Mall, Sequoia Hall, California 94305, United States
\textsuperscript{3} Georg-August-Universität Göttingen, Chairs of Statistics and Econometrics, Humboldtallee 3, 37073 Göttingen, Germany
\textsuperscript{4} Ludwig-Maximilians-Universität München, Department of Statistics, Akademies-straße 1, 80799 Munich, Germany

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

Abstract: As in many other sorts of regression problems, also in survival analysis it has become more and more relevant to face high-dimensional data with lots of potentially influential covariates. These generally can have time-constant or time-varying effect types, which a priori is often unknown to the modeler. A possible solution is to apply regularized estimation methods that allow to select relevant covariates and distinguish between these effect types. To address these model selection issues, a likelihood-based component-wise boosting approach for a specific Cox-type model, the so-called Cox frailty model, is proposed. Via Gaussian frailties heterogeneity in the population is modeled, resulting in flexible and sparse hazard rate models for survival data. The method will be applied to model time until pregnancy of German women, illustrating that the complexity of the influence structure can be strongly reduced by using the proposed approach.

Keywords: Variable selection; Effects Selection; Boosting; Cox frailty model; Time-varying coefficients.

1 Introduction

When modeling continuous event time data, proportional hazards (PH) models, and particularly the semi-parametric Cox model (Cox, 1972), play a major role. The classical Cox model is based on the semi-parametric hazard

\[ \lambda(t|x_i) = \lambda_0(t) \exp(x_i^T \beta), \]

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
representing the hazard for observation $i$ at time $t$, conditionally on covariate information $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. The vector $\mathbf{\beta}$ collects linear regression effects that are usually estimated via maximization of the corresponding partial likelihood, while the shared baseline hazard $\lambda_0(t)$ is estimated non-parametrically via the Nelson-Aaalen estimator. In general, $\lambda(t|\mathbf{x}_i) = \lim_{\Delta t \to 0} \frac{P(t \leq T < \Delta t|T \geq t, \mathbf{x}_i)}{\Delta t}$ defines the hazard rate for continuous event times and represents the instantaneous risk of a transition at time $t$. This approach performs well in classical problems with more observations than predictors. To combat the $p > n$ case, two of the most commonly used regularization techniques are penalization and boosting.

A first penalization approach was proposed by Tibshirani (1997). It uses the so-called least absolute shrinkage and selection operator (LASSO) penalty in the Cox model. Since then, several extensions have been proposed, compare Park and Hastie (2007) or Goeman (2010), just to mention two. An alternative regularization strategy and powerful learning idea developed in the machine learning community is boosting. Though originally designed for classification problems, it turned out to be also highly beneficial in regression. Here, we propose a likelihood-based component-wise boosting approach for Cox frailty models, addressing several model selection issues.

2 Cox frailty model with time-varying coefficients

Dependencies within clusters of observations or heterogeneity between clusters can be captured effectively by frailty models. However, parameter estimation in frailty models is more challenging than in the Cox model, since the corresponding profile likelihood has no closed form solution. In the Cox PH frailty model the hazard rate of subject $j$ from cluster $i$ is given by

$$
\lambda_{ij}(t|\mathbf{x}_{ij}, b_i) = b_i \lambda_0(t) \exp(\mathbf{x}_{ij}^T \mathbf{\beta}),
$$

conditionally on covariates $\mathbf{x}_{ij}$ and shared frailty $b_i$. The frailties $b_i, i = 1, \ldots, n$, are frequently assumed to follow a gamma distribution because of its mathematical convenience. While some multiplicative frailty distributions, such as the gamma and inverse Gaussian, have already been extensively studied and closed form representations of the log-likelihoods are available, the log-normal distribution is often more intuitive and allows for more flexible and complex predictor structures, though parameter estimation in the corresponding model is computationally more demanding. The hazard function with multiplicative, log-normal frailties has the general form

$$
\lambda_{ij}(t|\mathbf{x}_{ij}, \mathbf{u}_{ij}, b_i) = \lambda_0(t) \exp(\mathbf{x}_{ij}^T \mathbf{\beta} + \mathbf{u}_{ij}^T \mathbf{b}_i),
$$

where $\mathbf{u}_{ij}^T = (u_{ij1}, \ldots, u_{ijq})$ is the covariate vector associated with random effects $\mathbf{b}_i$, which follow a multivariate Gaussian distribution, i.e., $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{Q}(\theta))$ with mean $\mathbf{0}$ and covariance matrix $\mathbf{Q}(\theta)$, which is depending
on unknown parameters $\theta$. In this case, a penalized quasi-likelihood (PQL) approach based on Laplace approximation can be used for estimation (Breslow and Clayton, 1993). In this context, it is especially important to provide effective estimation algorithms, as standard procedures for determination of tuning parameters such as CV are usually very time-consuming.

While for Cox frailty models with simple predictor structure

$$\eta_{ij} = x_{ij}^T \beta + u_{ij}^T b_i$$

in the hazard function some solutions have already been given, often more complex structures of the linear predictor need to be taken into account. In particular, time-varying effects $\gamma_k(t)$ can be incorporated into the linear predictor. A standard way to estimate these effects $\gamma_k(t)$ is to expand them in equally spaced B-splines yielding

$$\gamma_k(t) = \sum_{m=1}^M \alpha_{k,m} B_m(t; d),$$

where $\alpha_{k,m}$, $m = 1, \ldots, M$, denote unknown spline coefficients, which need to be estimated, and $B_m(t; d)$ is the $m$-th B-spline basis function of degree $d$. For a detailed description of B-splines, see, e.g., Ruppert et al. (2003). Estimation can be stabilized, if, similar to the time-varying effects, instead of a non-parametric a semi-parametric baseline hazard is considered, which can also be flexibly estimated within the B-spline framework. Then, using the transformation $\gamma_0(t) := \log(\lambda_0(t))$ and setting $z_{ij0} = 1 \forall i, j$, the hazard

$$\lambda_{ij}(t|x_{ij}, z_{ij}, u_{ij}, b_i) = \exp(\eta_{ij}(t))$$

is obtained, with

$$\eta_{ij}(t) := x_{ij}^T \beta + \sum_{k=0}^r z_{ijk} \left( \sum_{m=1}^M \alpha_{k,m} B_m(t; d) \right) + u_{ij}^T b_i.$$  

The covariates $z_{ij}^T = (1, z_{ij1}, \ldots, z_{ijr})$ are associated with both baseline hazard and time-varying effects. In general, estimation of the parameters in predictor (1) can be based on Cox’s well-known full log-likelihood

$$l(\beta, \alpha, b) = \sum_{i=1}^n \sum_{j=1}^{N_i} d_{ij} \eta_{ij}(t_{ij}) - \int_0^{t_{ij}} \exp(\eta_{ij}(s)) ds,$$

where $n$ denotes the number of clusters, $N_i$ the cluster sizes, and the survival times $t_{ij}$ being complete if $d_{ij} = 1$ and right censored if $d_{ij} = 0$. Note that, in general, all or parts of the covariates $x_{ij}, z_{ij}$ and $u_{ij}$ could also vary over time., which is then handled by data splitting.

The roughness of the time-varying effects $\gamma_k(t)$ is controlled by P-splines (Eilers and Marx, 1996). For each time-varying effect $\gamma_k(t)$ a difference penalty is subtracted from the log-likelihood (2), which is directly applied on the corresponding spline coefficients $\alpha_k^T = (\alpha_{k,1}, \ldots, \alpha_{k,M})$. The corresponding smoothing parameters are denoted by $\xi_k, k = 0, 1, \ldots, r$.

3 Boosting

With the emergence of ensemble schemes Freund and Schapire (1997) developed the well-known AdaBoost algorithm for binary classification. It uses a weak learner as classifier, e.g., a classification and regression tree (CART,
Breiman et al., 1984), and is able to reduce both bias and variance of the weak learner (Breiman, 1998). As boosting algorithms eventually overfit, the optimal number of iterations represents a tuning parameter, which has to be determined in some data-driven way, e.g., by cross-validation (CV). A first generic gradient boosting algorithm for right-censored survival data is introduced in Hothorn et al. (2006). Within the framework of parametric accelerated failure time models a flexible boosting algorithm for censored time-to-event data is found in Schmid and Hothorn (2008). For structured survival data, Hofner et al. (2013) propose a likelihood-based boosting algorithm that permits the inclusion of both parametric and nonparametric time-varying effects as well as nonparametric effects of continuous covariates, using penalized splines as the main modeling technique.

In this work, we present a likelihood-based component-wise boosting approach for Cox frailty models, which addresses the following model selection issues: it can determine which covariates should be included in the model, and, which of those included have a constant or time-varying effect. We account for complex structures of the linear predictor and incorporate time-varying effects $\gamma_k(t)$, which we expand in B-splines as in (1). A similar approach using penalization techniques has been proposed in Groll et al. (2017). However, we are confident that component-wise boosting methods have the potential to improve the effects selection performance compared to penalization approaches as the one of Groll et al. (2017), where all model selection aspects are controlled by just two penalty parameters. In contrast, component-wise boosting is based on a set of base-learners, i.e., functions that lead to (typically) small improvements of the estimation in a single iteration and, hence, are slowly approaching a solution. In general, within the fitting procedure each individual base-learner can be updated exactly as often as necessary in order to obtain a good fit, with the overall number of boosting steps officiating as the main tuning parameter. For example, a base-learner could be updated several times, yielding a rather complex final effect, while other effects are never updated. This way, model selection is implicitly obtained. In the following, we sketch how suitable base-learners can be chosen in order to achieve the desired effects selection.

Kneib et al. (2009) suggested a modified parametrization of the P-splines, which is essential for the capability of effects selection of the likelihood-based component-wise boosting approach proposed in the following. Therefore, a potentially time-varying effect $\gamma(t)$ of a covariate $z$ is split into a parametric part consisting of an unpenalized polynomial of order $d - 1$ and the non-parametric deviation from this polynomial $\gamma_{\text{centered}}(t)$:

$$
\gamma(t) \cdot z = \underbrace{\alpha_0 \cdot z + \alpha_1 t \cdot z + \ldots + \alpha_{d-1} t^{d-1} \cdot z}_{\text{unpenalized part}} + \underbrace{\gamma_{\text{centered}}(t) \cdot z}_{\text{smooth penalized part}}.
$$

Technically, this decomposition is achieved by decomposing the corresponding vector of regression coefficients $\alpha$ into $\alpha^T = (\alpha_{\text{unpen}}^T, \alpha_{\text{pen}}^T)$, i.e., into
an unpenalized and a penalized part using spectral decomposition of the penalty matrix. In particular, if first order differences are used, we obtain

$$\gamma(t) \cdot z = \alpha_0 \cdot z + \gamma_{\text{centered}}(t) \cdot z,$$

which simply decomposes the time-varying effect into a linear (time-constant) effect and a smooth time-varying part. This is a key step with regard to effects selection of our iterative component-wise boosting procedure: we can now specify two base-learners for each (potentially) time-varying effect: a linear base-learner \( \alpha_0 \cdot z \), and a smooth deviation from linearity for this time-varying effect, \( \gamma_{\text{centered}}(t) \cdot z \). Both variable selection and model choice can now simultaneously be obtained. The former is achieved due to the component-wise nature of the method, as for each covariate separate base-learners are specified and in a single iteration only the best fitting base-learner (with respect to a certain criterion) is updated. To incorporate model choice in the boosting framework separate base-learners for each modeling possibility are added. A covariate effect can then be added in one of the two modeling possibilities from (3), which corresponds to model choice. After all, a small number of boosting iterations corresponds to a certain amount of regularization: if the boosting procedure is stopped early, several base-learners might have never been updated. Altogether, if none of the base-learners associated with a certain covariate \( z \) is selected, the covariate is completely excluded from the model. By setting \( z \equiv 1 \) in equation (3), in principle, the (log-)baseline hazard \( \gamma_0(t) = \log(\lambda_0(t)) \) can easily be incorporated into the boosting framework as well.

However, for this strategy it is important to guarantee a fair comparability of the different base-learners in terms of complexity. Following Hofner et al. (2011a), this can be done by defining equal degrees of freedom for each base-learner. Of course, increasing the smoothing parameters \( \xi_k \) from Section 2 generally leads to decreasing degrees of freedom (dfs), while in the decomposition (3) the linear base learner always has one df. Hence, the idea is to fix the smoothing parameter corresponding to the smooth time-varying part in (3) such that it also corresponds to one df, which is possible through the relationship of the smoothing parameter and the dfs established in Hofner et al. (2011a). This way, the smoothness of a function can be defined directly by the (effective) dfs instead of the smoothing parameter \( \xi_k \). Eilers and Marx (1996) have shown that a polynomial of degree \( d - 1 \) remains unpenalized by a \( d \)-th order difference penalty if the degree of the B-spline basis is larger or equal than \( d \). So when using B-splines of degree 3 combined with difference penalties of order 1, the dfs of the smooth time-varying part in (3) can indeed be fixed to exactly one. In particular, following e.g. Hofner et al. (2013), the degrees of freedom in flexible survival models with penalized splines can be derived by \( \text{df} = \text{trace}(F \cdot F_{\text{pen}}^{-1}) \).
References


Inflated Beta Regression Modelling for Speech Intelligibility Testing

Gillian Heller¹, Wenli Hu¹, Brett Swanson²

¹ Macquarie University, Sydney, Australia
² Cochlear Australia

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

Abstract: Speech intelligibility tests are conducted on hearing-impaired people, for the purpose of evaluating the performance of hearing devices. The speech reception threshold (SRT) is defined as the signal-to-noise ratio (SNR) at which a subject scores 50% correct on a speech intelligibility test. The SRT is estimated on a track of 20 sentences, and the estimated SRTs have traditionally been analyzed with a repeated measures ANOVA. We propose an alternative approach for analysis, a zero-and-one inflated beta regression model, in which an observation is a single sentence score rather than an SRT. The proposed approach was applied retrospectively to a study that assessed speech perception using three sound processing algorithms under different listening conditions. The results of this approach were consistent with the traditional approach, but were more informative. The fitted mean curve of each condition revealed differences in slope, i.e. differential performance at different parts of the SNR spectrum, a feature not detectable under the traditional approach.

Keywords: Inflated beta regression; psychometric curves; speech intelligibility.

1 Introduction

The psychometric function is an S-curve used in fields such as psychophysics and biomedical engineering, to model the relationship between a stimulus and response (Wichmann and Hill 2001). Typically two-parameter versions are used, such as the logistic or Weibull distribution functions \( F(x; \alpha, \beta) \). In the simplest case a subject responds either correctly or incorrectly to a stimulus, where high values of the stimulus are associated with higher probability of correct response. We consider speech intelligibility tests, which are conducted on hearing-impaired people, for the purpose of evaluating the performance of hearing devices under controlled listening conditions in...
a sound laboratory. The simplest version is that the subject listens to a pre-recorded word, mixed with noise, and repeats it back. Here the stimulus is the signal-to-noise ratio (SNR) and the subject’s response is either correct (1) or incorrect (0). Of interest is the speech reception threshold (SRT) $F_{0.5}^{-1}$, which is the stimulus value at which a subject has probability 0.5 of correct response. Use of the logistic regression model implies a logistic psychometric curve.

More common usage is for the subject to listen to a short sentence (of length $n$ words) and repeat it back. The number of correctly identified words $y$ is recorded and the proportion correct ($y/n$) is the response of interest. Typically data are collected in tracks of 20 sentences each, in which experimental conditions (e.g. noise gender, device algorithm) are the same but the SNR is varied over its range. A sample data set for a single track, with fitted SRT, is depicted in Figure 1.

![Proportion correct vs SNR in single track of 20 sentences, with fitted psychometric curve and SRT.](image)

**FIGURE 1.** Proportion correct vs SNR in single track of 20 sentences, with fitted psychometric curve and SRT.

## 2 Statistical model

Traditionally, these data have been analysed by summarising each track by a single number, the derived SRT, and analysing the SRTs as observations in an analysis of variance (e.g. Dawson et al 2011). The obvious shortcoming in this approach is that use of SRTs rather than the raw observations results in a substantial loss of information, particularly at the extremes of the SNR range.

Using the sentence as the unit of observation, an obvious model for the proportion correct is the binomial distribution; however, the assumption of
independent recognition of words in a sentence is unlikely to be satisfied.
The histogram of the proportion correct for all observations (sentences) in a study (Hu et al 2015) is shown in Figure 2, in which the striking features are the high frequencies at zero and one. This is explained by the use of context in word recognition, which results in a large proportion of sentences which are either completely recognised, or not recognised at all.

We propose the use of the zero-and-one inflated beta distribution (Ospina and Ferrari 2010). The distribution of proportion correct in $(0, 1)$ is approximated by the beta distribution, and explicit probability masses at zero and one are specified. Models may be specified for the mean and shape parameters of the beta distribution and the zero and one probabilities. In addition, we specify a random effect for subject, which is particularly important in this case as subjects vary substantially in their degree of hearing loss. The following is the mixed discrete-continuous probability model:

$$f(y; \mu, \sigma, \nu, \tau) = \begin{cases} \frac{\nu}{1+\nu+\tau} & y = 0 \\ \frac{1}{1+\nu+\tau} \cdot f_1(y; \mu, \sigma) & y \in (0, 1) \\ \frac{\tau}{1+\nu+\tau} & y = 1 \end{cases}$$

where $f_1(y; \mu, \sigma)$ is the pdf of the beta distribution, parametrised in terms of its mean $\mu$ and shape parameter $\sigma$; $\mu, \sigma \in (0, 1); \nu, \tau > 0$. The probability masses of measures 0 and 1 are associated with the parameters $\nu$ and $\tau$ as $\nu = p_0/(1 - p_0 - p_1)$ and $\tau = p_1/(1 - p_0 - p_1)$.

The parameters $\mu, \sigma, \nu$ and $\tau$ are modeled with covariates, as well as random

FIGURE 2. Histogram of proportion correct (all observations in study).
effects to account for within-subject correlation. We specify the models

\[
\begin{align*}
    \text{logit}(\mu) &= x^\top \beta + u_1 \\
    \text{logit}(\sigma) &= z^\top \gamma + u_2 \\
    \log(\nu) &= h^\top \lambda + u_3 \\
    \log(\tau) &= k^\top \rho + u_4
\end{align*}
\]

where $x$, $z$, $h$ and $k$ are vectors of known covariates, which may be overlapping or distinct; $\beta$, $\gamma$, $\lambda$ and $\rho$ are corresponding coefficient vectors; and

$u_j \sim \mathcal{N}(0, \delta_j^2), j = 1, \ldots, 4$ are random effects for subject. Logit links are used for the parameters constrained to $(0, 1)$, i.e. $\mu$ and $\sigma$, and log links for those constrained to $R^+$ ($\nu$ and $\tau$), as is common practice in generalized linear modeling. Parameter estimation is achieved in the R package `gamlss` (Stasinopoulos et al 2017).

3 The study

The study (Hu et al 2015) compared the speech recognition of seven cochlear implant recipients in the presence of an interfering talker. The primary factor of interest was the sound processing algorithm (A, B, C), and the question was whether the three algorithms yielded differences in the subjects’ performance. The other factors were the direction of the interfering talker ("noise direction"), which was either from the front or from both sides; and the gender of the talker ("noise gender"). As the target speech was presented from the front, it was hypothesized that performance would be better for side interferers, as subjects could potentially use the difference in spatial location to segregate the two voices. As the target voice was female, it was hypothesized that performance would be better for a male interferer, as subjects could potentially use the difference in voice pitch to segregate the two voices. For noise direction and noise gender, interaction with algorithm would indicate that the algorithms differed in their effectiveness in conveying spatial or pitch information. For all factors, interaction with SNR would indicate differential effects of the factors over the SNR spectrum, which could not be evaluated using the traditional approach.

4 Results

*Traditional analysis*: In a 3-way repeated measures ANOVA, noise gender-algorithm interaction was significant. Noise direction was not significant.

*Proposed analysis*: No covariates were significant for $\sigma$. For $\mu$, $\nu$ and $\tau$, SNR, noise gender and algorithm were all significant. Significant interactions were: noise gender-algorithm for $\mu$ and $\tau$; SNR-algorithm for $\nu$; and SNR-noise gender for $\tau$. (Parameter estimates are not shown for reasons of
space.) To interpret these results, fitted values are shown, as parameter interpretation over multiple model parameters is complex. The fitted overall means of percent correct (i.e. \((1 - \hat{p}_0 - \hat{p}_1)\hat{\mu} + \hat{p}_1\)) for each algorithm are shown in Figure 3, separately for noise genders, with parametric bootstrap confidence intervals. The curves for algorithms B and C largely overlap at all SNRs, indicating little difference between those algorithms. However, the curve for algorithm A has a steeper slope than those for algorithms B and C; for female interferers, the three curves overlap at low SNRs, but start to separate at higher SNRs, indicating that the performance of algorithm A is superior to that of B and C, at high SNR and with female interfering noise. This informative difference in slope between algorithms was not detectable using the traditional approach.

![Graph showing percent correct vs SNR for different noise genders and algorithms](image)

FIGURE 3. Fitted overall means of percent correct, showing effect of algorithm. Confidence intervals were calculated by 500 bootstrap samples.

References


Estimating dose-response for time-to-remission with instrumental variable adjustment

Jennifer Hellier¹, Richard Emsley¹, Andrew Pickles¹

¹ Biostatistics and Health Informatics Department, Institute of Psychiatry, Psychology and Neuroscience, King’s College London.

E-mail for correspondence: jennifer.hellier@kcl.ac.uk

Abstract: Threshold regression (TR), in which time to remission is modelled as a stochastic drift towards a boundary, is an alternative to the proportional hazards (PH) survival model, and has a clear conceptual mechanism for examining the effects of drug dose. However, for both TR and PH models, when dose titration occurs during treatment, the estimated causal effect of dose can be biased by confounding. An instrumental variable (IV) analysis can be used to remove such bias. This paper was motivated by the analysis of an antidepressant randomised clinical trial (RCT). We examine these issues for dose response.

Keywords: Threshold regression; Inverse Gaussian; Instrumental variables; Dose response.

1 Introduction

A common complication of time to depression remission studies is the necessity to titrate each patient’s drug dose over the first few weeks to a level considered therapeutic and acceptable. This practise can result in bias in the naive estimates of drug dose-response. A possible solution to this is to estimate the effect of dose using an IV approach (Hernan and Robins 2014). This requires having a variable (the instrument) that influences dose but does not influence time to remission except through its influence on dose, and such a variable can be obtained by exploiting the random treatment assignment of a RCT. We demonstrate a novel IV application utilising an inverse Gaussian (IG) distribution of TR and contrast this to a PH approach.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
1.1 Cox Regression Model

The Cox PH regression model uses the exponential formulation for the hazard function, we note that the effect of predictors of time to remission enter the model multiplicatively:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m)$$

where the constant $\lambda_0(t)$ is a baseline hazard function, $\lambda_i(t)$ is the hazard function at time $t$, here $i$ is a subscript for observation and the $x$s are the covariates with effects estimated by their corresponding coefficient. The Cox model, though it has many strengths, is non-collapsible (Burgess 2015), since the marginal regression parameter for a covariate is not equal to the conditional parameter.

1.2 Inverse Gaussian Model

We model an inverse Gaussian distribution of threshold regression (Aalen and Gjessing 2001). The distribution of time to remission ($t$) is given by:

$$f_{ig}(t) = \left(\frac{c_i}{\sigma_i}/\sqrt{2\pi t^3}\right)^{-1/2} \exp\left(-\frac{(c_i - \mu_i t)^2}{2\sigma_i^2 t}\right)$$

The IG distribution depends on the mean and variance parameters of the underlying Wiener process ($\mu$ and $\sigma$) and the initial patient status ($c$). Where $c_i/\sigma_i$ is the patients initial distance from the threshold, $\mu_i/\sigma_i$ is the drift velocity. Most explanatory variables (denoted $x_1, x_2, \ldots, x_m$) can be linked to both initial severity (distance from the remission threshold) and the rate of symptom improvement (velocity). Initial distance is linked to baseline covariates with an exponential function:

$$c_i/\sigma_i = \exp(\alpha' x_i) = \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_m x_m) = \exp(\theta_1)$$

Randomised treatment or post randomisation variables such as dose (denoted $z_1, z_2, \ldots, z_m$), can only be associated with the rate of symptom improvement. The velocity is linked to covariates as a linear function, given the matrix $\gamma = (x_1, x_2, \ldots, x_m, z_1, z_2, \ldots, z_m)$:

$$\mu_i/\sigma_i = (\beta' \gamma_i) = (\beta_1 \gamma_1 + \beta_2 \gamma_2 + \ldots + \beta_m \gamma_m) = \theta_2$$

Where $\alpha$ and $\beta$ are coefficients to be estimated. The simple linear function implies that the predicted direction of drift velocity can be both towards and away from the boundary, and that remission is not inevitable even in the long-run.

1.3 Instrumental Variable

IV analysis rests on three fundamental assumptions: the instrument must be correlated with the exposure of interest (relevance); the instrument must be independent of the confounder (exogeneity); and the instrument affects the outcome only through its relation to the exposure of interest (exclusion restriction) (Angrist, Imbens et al 1996).
1.4 Data
Weekly antidepressant dose was measured in 380 men and women with major depression treated with escitalopram or nortriptyline for twelve weeks in the Genome-Based therapeutic Drugs for Depression (GENDEP) study (Uher R et al 2009). Baseline variables on sociodemographic characteristics, health status and personal depression history are included. The averaged dose relative to maximum prescribing dose was calculated from the twelve study weeks and tested for association with time to depression remission in the PH and IG models.

2 Modelling
Randomised antidepressant treatment was used as an instrumental variable using the control function approach (Tchetgen Tchetgen et al 2015). The first stage of this method involved estimating average relative dose from our IV variable; randomised treatment. This regression model was further adjusted for the baseline covariates and age at randomisation. Inclusion of additional covariates in this step can help improve precision of the IV estimator. The residuals from the first stage regression can be considered as estimating the effects of other uncontrolled factors that influence a patient’s dose. In the second stage of this method, we estimate the survival model, but we include among the predictor variables the residuals from the first stage together with the patients relative dose. The second stage was implemented in both survival models. The IG model was constructed with two linear predictors (LP), residuals considered in velocity LP during IV analyses. Confidence intervals and p-values were based on 1000 non-parametric bootstrap samples to account for the two staged approach.

3 Results
Application of a control function IV method is intended to give an unbiased estimate of dose response on time to remission. In the naive model averaged daily dose was not associated with reduced time to remission. By contrast, the IV analyses showed a clear and significant relationship between increased dose and faster time to remission, Table 1. Both the direction and magnitude of the dose response were changed by use of the IV approach, and though evident in both the IG and PH models, the non-collapsibility of the PH model made this model hard to interpret. The validity of these results is dependent on the validity of IV assumptions.

4 Conclusions
We have used a novel application of IV techniques in an IG model to estimate an unbiased relative dose response on time to remission relationship
TABLE 1. Regression output for dose response on time to remission under the Cox (PH) and Inverse Gaussian (IG) model with and without the IV stage 1 residuals introduced as a control variable. MADRS: Montgomery-Asberg Depression Rating Scale; BMI: Body Mass Index; LP: Linear Predictor **p<0.001 * P<0.05

<table>
<thead>
<tr>
<th>Variables</th>
<th>PH Model</th>
<th>IG Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log HR 95%CI</td>
<td>Coeff 95% CI</td>
</tr>
<tr>
<td><strong>Distance LP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression (MADRS)</td>
<td>-0.055** (-0.083, -0.026)</td>
<td>0.027** (0.018, 0.037)</td>
</tr>
<tr>
<td>with stage 1 residuals</td>
<td>-0.062** (-0.100, -0.024)</td>
<td>0.029** (0.018, 0.041)</td>
</tr>
<tr>
<td>Duration (weeks)</td>
<td>-0.008 (-0.018, 0.003)</td>
<td>0.003 (0.000, 0.007)</td>
</tr>
<tr>
<td>with stage 1 residuals</td>
<td>-0.008 (-0.020, 0.003)</td>
<td>0.004 (-0.001, 0.008)</td>
</tr>
<tr>
<td>Sex (female)</td>
<td>-0.087 (-0.431, 0.257)</td>
<td>0.119 (-0.022, 0.261)</td>
</tr>
<tr>
<td>with stage 1 residuals</td>
<td>-0.268 (-0.684, 0.147)</td>
<td>0.163 (-0.009, 0.337)</td>
</tr>
<tr>
<td><strong>Velocity LP</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age of onset (years)</td>
<td>-0.010 (-0.027, 0.006)</td>
<td>-0.003 (-0.007, 0.001)</td>
</tr>
<tr>
<td>with stage 1 residuals</td>
<td>-0.009 (-0.028, 0.010)</td>
<td>-0.003 (-0.008, 0.002)</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>-0.028 (-0.062, 0.006)</td>
<td>-0.009* (-0.019, 0.000)</td>
</tr>
<tr>
<td>with stage 1 residuals</td>
<td>-0.034 (-0.071, 0.003)</td>
<td>-0.011* (-0.021, -0.001)</td>
</tr>
<tr>
<td>Relative dose</td>
<td>-0.178 (-0.939, 0.583)</td>
<td>-0.029 (-0.234, 0.175)</td>
</tr>
<tr>
<td>with stage 1 residuals</td>
<td>3.170* (0.353, 5.987)</td>
<td>0.881* (0.136, 1.626)</td>
</tr>
<tr>
<td>Stage 1 Residuals</td>
<td>-3.660* (-6.570, -0.751)</td>
<td>-1.010* (-1.795, -0.225)</td>
</tr>
</tbody>
</table>

in the GENDEP study. We demonstrate, the benefits of these analyses estimating causal parameters accounting for confounding. This is achieved through using a model with mechanistic underpinnings, together with a two-stage IV estimator to account for the unmeasured confounding that arose from the process of individual patient dose titration. The results for the trial dataset illustrate emphatically how the method can recover scientifically meaningful estimates of the dose-response relationship even where standard analysis gives estimates in the wrong direction.

**Acknowledgments:** JH was supported by a National Institute for Health Research (NIHR) doctoral fellowship. AP and RE were part funded by NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and Kings College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The authors would like to thank Dr Rudolf Uher for his helpful collaboration and for providing the GENDEP data.
References


Treatment effects beyond the mean using distributional regression

Maike Hohberg¹, Peter Pütz¹, Thomas Kneib¹

¹ Chair of Statistics, University of Goettingen, Germany

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

Abstract: We introduce distributional regression as a modeling framework for analyzing treatment effects beyond the mean in program evaluation. By relating each parameter of the response distribution to explanatory variables, it models the treatment effect on the whole conditional distribution. We elaborate on the combination of distributional regression with program evaluation methods used in economics to establish causality, demonstrate its benefits, and provide practical guidance to the usage of distributional regression by reanalyzing data from the Progresa program. Contrary to expectations, we find no significant decline in food consumption inequality for average households after introducing conditional cash transfers.

Keywords: Treatment effects; GAMLSS; inequality

1 Introduction

Program evaluation in economics typically identifies the effect of a policy or an intervention as the average difference between treatment and control group with respect to the response variable. Concentrating on mean differences between treatment group and control group is likely to miss important information about changes along the whole distribution of an outcome, for example in terms of an unintended increase in inequality, or when targeting ex ante vulnerability to a certain risk. These concepts rely on additional measures such as the variance and skewness of the response. For a systematic and coherent analysis of treatment effects on all parameters of the response distribution, we introduce distributional regression (GAMLSS, Rigby and Stasinopoulos 2005) to the program evaluation literature. Our aim is to demonstrate the implementation and benefits of distributional regression models in program evaluation to practitioners. For
this purpose, Hohberg et al. (2017) provide a step-by-step guide that is applied to the Progresa data set.

Distributional regression allows all parameters of the response distribution to vary with explanatory variables and can hence be used to assess how the conditional response distribution changes due to the treatment. It encompasses a wide range of potential outcome distributions and practically every distribution functional can be derived from these models. Distributional regression assumes that the observed \( y_i \) are conditionally independent and that their distribution can be described by a parametric density \( p(y_i|\vartheta_{i1}, \ldots, \vartheta_{iK}) \) where \( \vartheta_{i1}, \ldots, \vartheta_{iK} \) are \( K \) different parameters of the distribution. Each parameter \( \vartheta_{ik}, k = 1, \ldots, K, \) of the response distribution is conditioned on several explanatory variables and can be related to a predictor \( \eta_{ik} \) via a link function \( g_k \) such that \( \vartheta_{ik} = g_k^{-1}(\eta_{ik}) \).

A generic predictor for parameter \( \vartheta_{ik} \) takes on the following form:

\[
\eta_{ik} = \beta_{0k} + f_{1k}(x_{i1}) + \cdots + f_{jk}(x_{ij}).
\]

While \( \beta_{0k} \) denotes the overall level of the predictor, functions \( f_{jk}(x_{ij}), j = 1, \ldots, J_k, \) can be chosen to model a range of different effects of the vector of explanatory variables \( x_i \). Estimation can be done via a back-fitting approach within the Newton-Raphson type algorithm that maximizes the penalized likelihood and is implemented in the \texttt{gamlss} package in \texttt{R}.

### 2 Distributional regression in program evaluation

Compared to an analysis with distributional measures as the dependent variable, distributional regression yields one model from which several measures of interest can be consistently derived. Furthermore, aggregated distributional measures as dependent variables mask the underlying individual information whereas distributional regression allows to estimate effects on aggregate measures on the individual level. To present some examples of beyond-the-mean-measures, which are of interest when for example evaluating anti-poverty programs, we focus in the following on (1) vulnerability to poverty and (2) inequality.

(1) Vulnerability as expected poverty is the probability of having a consumption (or income) level below a certain threshold. To calculate this probability and assuming normally distributed log incomes \( \ln y_i \), the estimated coefficients for distribution parameters \( \mu \) and \( \sigma \) obtained from a distributional regression model are plugged into the standard normal cumulative distribution function

\[
\Pr(\ln y_i < \ln z|x'_i) = \Phi\left( \frac{\ln z - (\hat{\beta}_0^\mu + x'_i\hat{\beta}_1^\mu)}{\sqrt{\hat{\beta}_0^\sigma + x'_i\hat{\beta}_1^\sigma}} \right),
\]
where $\hat{\beta}_0$ is an intercept, $x_i$ is a vector of household characteristics, $\hat{\beta}_1$ is a vector of coefficients of the same length. Consequently, $\hat{\beta}_0^t + x_i'\hat{\beta}_1^t$ is the estimated mean, $\sqrt{\hat{\beta}_0^\sigma + x_i'\hat{\beta}_1^\sigma}$ the estimated standard deviation, and $z$ the poverty threshold.

Traditionally, vulnerability as proposed by Chaudhuri et al. (2002) is calculated using the feasible generalized least squares estimator (FGLS, Amemiya 1977) and involves running separate regressions for mean and variance of log consumption and includes several estimation and weighting steps. In contrast to such a stepwise procedure, distributional regression allows to estimate the effects on mean and variance simultaneously which simplifies quantifying uncertainty.

(2) Inequality is considered an unintended but relevant effect of an anti-poverty program since a change in inequality affects welfare. By focusing on the conditional distribution of consumption or income, we do not measure unconditional inequality of consumption or income, e.g. for the entire treatment and control group, but inequality given that other factors that explain differences in consumption are fixed at certain values. Inequality can be measured using the Gini coefficient. For a given continuous consumption distribution function $p(y)$, the Gini coefficient can be written as

$$G = \frac{1}{2\mu} \int_{0}^{\infty} \int_{0}^{\infty} p(y)p(z) |y - z| \, dy \, dz,$$

with $\mu$ denoting the mean of the distribution.

3 Application: Treatment effects beyond the mean using experimental data

Angelucci and De Giorgi (2009) investigate how conditional cash transfers to eligible (poor) households affect food consumption per adult equivalent of both eligible and ineligible (nonpoor) households in treatment and control villages. They used data from the Mexican Progresa program which was randomized at the village level. Aside from the expected positive effect of the cash transfer on the eligible households’ food consumption, the authors find a considerable increase of ineligible households’ food consumption in the treatment villages. Since the identified average treatment effects are larger for the poor than for the nonpoor, a lower food consumption inequality in the treatment villages is expected.

We rely on the same data and control variables as Angelucci and De Giorgi (2009) and build a distributional regression model using the three parameter Singh-Maddala distribution and the sample of all households in control and treatment villages. Marginal treatment effects on the mean and variance are computed as well as on other interesting features of the distribution, i.e. the Gini coefficient and vulnerability as expected poverty. Figure 1 graphically shows the conditional distribution for an average household, i.e.
evaluated at mean values for the other explanatory variables, in a treatment village compared to a control village. A shift to the right of the treatment curve indicates a positive marginal treatment effect at means on food consumption. This is also reflected in the point estimates.

**TABLE 1.** Treatment effects for all people in treatment villages \((n = 14,740)\): Point estimates and 95% bootstrap confidence interval bounds for marginal treatment effects at means (MTE)

<table>
<thead>
<tr>
<th></th>
<th>Point Estimate</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTE on mean</td>
<td>25.907</td>
<td>19.571</td>
<td>32.251</td>
</tr>
<tr>
<td>MTE on variance</td>
<td>4870.407</td>
<td>1703.281</td>
<td>8852.551</td>
</tr>
<tr>
<td>MTE on Gini</td>
<td>0.007</td>
<td>-0.004</td>
<td>0.021</td>
</tr>
<tr>
<td>MTE on vulnerability</td>
<td>-0.056</td>
<td>-0.073</td>
<td>-0.039</td>
</tr>
</tbody>
</table>

The results in Table 1 show point estimates for an average household. Living in a treatment village induces an expected increase in food consumption of about 25.91 pesos per adult equivalent while vulnerability as expected poverty decreases by -0.056. If every household benefits equally by the same amount or the poorest even more, the Gini decreases. However, the Gini coefficient increases in treatment villages and the confidence intervals do not reject the null hypothesis of equal food consumption inequality between treatment and control villages. There is no evidence that the treatment decreases inequality in food consumption for an average household, a quite
sobering result for a poverty alleviation program.

4 Cluster-robust Bootstrap inference

We account for the village grouping structure and unobserved village heterogeneity by applying a pairs cluster bootstrap procedure to obtain cluster-robust inference. Cameron and Miller (2015) give a comprehensive overview on cluster-robust inference, also within the bootstrap machinery. As a method applicable to nonlinear models, they propose a nonparametric pairs cluster bootstrap to obtain cluster-robust inference. Assume that the aim is a significance statement on the marginal treatment effect at means on the Gini coefficient and that the sample consists of $G$ clusters or groups. Then, repeat the following procedure $B$ times:

1. Resample $G$ clusters $(y_1, X_1), \ldots, (y_G, X_G)$ with replacement from the $G$ clusters in the original sample, where $y_g$ denotes the vector of responses in cluster $g$ and $X_g$ the matrix of covariates.

2. Run the distributional regression model for the bootstrap sample obtained in step 1 and predict the respective conditional distributions at mean values for other covariates for treated and nontreated individuals. For these distributions, the respective Gini coefficients are computed and their difference is calculated. This difference between the coefficients is the estimated marginal treatment effect at means on the Gini coefficient and is denoted by $\hat{\theta}_b$ for the current bootstrap sample.

From the resulting $B$ bootstrap estimates $\hat{\theta}_1, \ldots, \hat{\theta}_B$, bootstrap inference can be conducted in different ways. One option is to perform a t-test based on the cluster-robust variance estimator

$$
\hat{V}_{\text{clu,boot}}[\hat{\theta}] = \frac{c}{B - 1} \sum_{b=1}^{B} (\hat{\theta}_b - \bar{\hat{\theta}})^2,
$$

where $\bar{\hat{\theta}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b$ and $c = \frac{G}{G-1} \frac{N-1}{N-K}$ is a finite sample modification with the number of estimated model quantities denoted by $K$.

To test for significance of the marginal treatment effect on the Gini, the t statistic

$$
t = \frac{\hat{\theta}}{\sqrt{\hat{V}_{\text{clu, boot}}[\hat{\theta}]}}
$$

can be used, where $\hat{\theta}$ is the estimate for the marginal treatment effect from the original sample or the mean of all bootstrap estimates.
5 Treatment effects beyond the mean using non-experimental data

The practical example presented here considers only the case of a randomized controlled trial. However, program evaluation also relies on analyzing non-experimental data, for example when a policy is introduced and no prior experiments are not possible or unethical. Popular methods in this context include regression discontinuity designs, differences-in-differences, panel data techniques, and instrumental variable regression. In Hohberg et al. (2017), we develop frameworks for combining each of these methods with distributional regression.

References


Mixture of Experts Models for Joint Insurance Claims Modelling with Multivariate Gamma Distributions

Sen Hu\textsuperscript{1,2}, Adrian O’Hagan\textsuperscript{1,2}, T. Brendan Murphy\textsuperscript{1,2}

\textsuperscript{1} School of Mathematics and Statistics, University College Dublin, Ireland
\textsuperscript{2} Insight Centre for Data Analytics, Dublin, Ireland

E-mail for correspondence: adrian.ohagan@ucd.ie

Abstract: In general insurance, it is a common phenomenon that claims from multiple risk categories can be correlated: in personal motor insurance an accident can lead to claims for policyholder’s own vehicle, others’ property damage or bodily injury. Dependence can also be present across multiple product lines: claim history and personal characteristics from a motor policy can reflect aspects of one’s home policy claims. Many current insurance pricing approaches assume independence among multiple categories, focusing on independent modelling and then the sum of the risks will be taken as the total risk the insurer is taking on. In this work, multivariate versions of the frequency and severity models are investigated using mixture of experts models with multivariate Poisson and multivariate Gamma distributions, where the dependence structure will be modelled directly using a GLM framework. In particular, focus is given to the multivariate Gamma distribution which has not yet been widely investigated in the regression and mixture model literature. It can be viewed as a model-based clustering approach which clusters policyholders into groups with different dependence structures, and covariates of characteristics of policyholders, insured objects and insurance policy can be taken into account in the process. It shows that, by applying to a real-world Irish GI insurer dataset, claim predictions can be improved.

Keywords: Multivariate Gamma distribution; Mixture of experts; Model-based clustering; Generalized linear model.

1 Introduction

In general insurance, it is a common practice that risks from different categories are modelled independently, and their sum is regarded as the total risk the insurer takes on in exchange for a premium. The dependence from
multiple risks is generally neglected, although one could assume that an incident may result in multiple claims from more than one category, or some policyholders are more prone to make more claims from different policies than others when holding multiple policies with the same insurer. General linear models (GLMs) (Nelder & Wedderburn, 1972) have become the industry’s standard approach for claim modelling, typically using a univariate Poisson distribution for the frequency aspect of modelling (how often claims are made from policies) and a univariate Gamma distribution for severity (size of loss for an insurer given a claim has been made). For modelling multiple risks’ marginals and the dependence structure simultaneously, multivariate distributions within the GLM framework provide a useful approach extended from their univariate counterparts. Many attempts have been made regarding using multivariate Poisson regression (either standalone or as finite mixtures) in the literature. For more detailed discussion, see Karlis and Ntzoufras (2003), Karlis and Meligkotsidou (2007) and Bermdez and Karlis (2012). However, multivariate Gamma distributions have received less attention, especially regarding GLMs or the mixture of experts framework. This will be the main focus of this work.

2 Multivariate Gamma distribution

A bivariate case of the multivariate Gamma distribution is illustrated here, the extension to higher dimensions could be constructed similarly. There are various definitions of bivariate Gamma distributions - for a detailed review see Kotz et al (2000). This work uses Cherian and Ramabhadran’s bivariate Gamma definition (Cherian, 1941; Ramabhadran, 1951; Mathal and Moschopoulous, 1992): let $X_1, X_2, X_3$ be independent Gamma random variables, where $X_i \sim \text{Gamma}(\alpha_i, \beta)$ for $i = 1, 2, 3$ and common rate parameter $\beta$. Then

$$
Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 + X_3 \\ X_2 + X_3 \end{bmatrix} \sim MG_2(\alpha_1, \alpha_2, \alpha_3, \beta).
$$

It has density (using trivariate reduction)

$$
f(y_1, y_2) = \frac{\beta^{\alpha_1+\alpha_2+\alpha_3} e^{-\beta(y_1+y_2)}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \int_0^m e^{\beta x_3} x_3^{\alpha_3-1}(y_1 - x_3)^{\alpha_1-1}(y_2 - x_3)^{\alpha_2-1} dx_3,
$$

where $m = \min(y_1, y_2)$. This definition has the benefit that the marginals are also Gamma distributions such that $Y_1 \sim \text{Gamma}(\alpha_1+\alpha_0, \beta)$ and $Y_2 \sim \text{Gamma}(\alpha_2+\alpha_0, \beta)$, $\text{Cov}(Y_1, Y_2) = \frac{\alpha_0}{\beta^2}$. It is also worth noting that this distribution can only model positive covariance on its own, which motivates the use of finite mixtures of bivariate Gamma distributions for more flexible modelling of either positive or negative covariance.
Extending the bivariate version to higher dimensions can be constructed similarly: as an example, a full covariance-structure trivariate Gamma distribution is constructed as

\[
Y = \begin{bmatrix}
Y_1 \\
Y_2 \\
Y_3
\end{bmatrix} = \begin{bmatrix}
X_1 + X_{12} + X_{13} + X_{123} \\
X_2 + X_{12} + X_{23} + X_{123} \\
X_3 + X_{13} + X_{23} + X_{123}
\end{bmatrix}
\]

where each \( X_i \sim Gamma(\alpha_i, \beta) \) for \( i = 1, 2, 3, 12, 13, 23, 123 \). This definition assumes distinct covariance between each pair of dimensions and a common covariance element across all dimensions. Simplified covariance structures can be attained if only assuming pair-wise covariance or only common covariance across all dimensions.

A common method of estimation for this bivariate Gamma distribution is the method of moments. Tsionas (2004) proposed the use of Bayesian Monte Carlo method for estimating this distribution. This work uses maximum likelihood estimation (EM algorithm) for parameter estimation. The estimation method is similar to the one used for MoE models below.

### 3 Clustering with mixture of experts models

A mixture model clusters outcome variables \( y_i \) without considering extra associated information in the data. For insurance claims modelling, covariates need to be included as predictors in regression for claim prediction. The mixture of experts model (MoE) extends the mixture model, allowing the parameters of the model to depend on covariates \( w_i \). Note that this is equivalent to mixture of multivariate Gamma regressions. For the purpose of insurance claim modelling, it has the following benefits: (1) both the marginals of multiple risks and their covariance in multivariate data are modelled directly; (2) there are typically heterogeneous groups among policyholders where different groups have different claim behaviours; (3) it is able to model both positive and negative correlation.

From now on, only bivariate Gamma is focused on for simplified argument. Higher dimensional cases can be implemented similarly. Suppose the population consists of \( G \) groups, for each group \( g = 1, \ldots, G \), \( y_i \sim f(y_i|\theta_g) \), where \( \theta_g \) is the parameters of group \( g \) and each \( f \) is a bivariate Gamma distribution. Then the density (conditional on covariates \( w_i \)) is

\[
f(y_i|w_i) = \sum_{g=1}^{G} \tau_g(w_0i)f(y_{1i}, y_{2i}; \alpha_{1ig}(w_{1i}), \alpha_{2ig}(w_{2i}), \alpha_{3ig}(w_{3i}), \beta_i),
\]

where

\[
\log(\alpha_{kig}) = \gamma_{kg}^T w_{ki} \quad \text{for} \quad k = 1, 2, 3,
\]
and \( w_i \) is a vector of covariates. If no covariates \( w_i \) are included, this model becomes the standard mixture model with the density

\[
f(y_i) = \sum_{g=1}^{G} \tau_g f(y_{1i}, y_{2i} | \alpha_{1g}, \alpha_{2g}, \alpha_{3g}, \beta_g).
\]

In the machine learning literature of mixture of experts, it refers to the component densities \( f(y_i | \theta_g(x_i)) \) as “experts network”, and to the mixing proportions \( \tau_g(x_i) \) as “gating network”. There are two latent variables: missing cluster labels \( z_i = \{z_{i1}, ..., z_{iG}\} \) where \( z_{ig} = 1 \) if observation \( i \) belongs to cluster \( g \) and \( z_{ig} = 0 \) otherwise, and the latent variable \( X_3 \) for each bivariate Gamma distribution. Hence the complete data likelihood is

\[
L_c = \prod_{i=1}^{N} \prod_{g=1}^{G} [\tau_g(w_{0i}) f(y_{1i}, y_{2i}, x_3 | \theta_g(w_i))]^{z_{ig}},
\]

where \( \theta(w_i) = \{\gamma_{1g}; \gamma_{2g}; \gamma_{3g}; \beta_{ig}\} \). The complete data log-likelihood is

\[
\ell_c = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log[\tau_g(w_{0i})P(y_{1i}, y_{2i}, x_3 | \theta_g(w_i))] = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log \tau_g(w_{0i}) + \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log P(x_{1i} | \alpha_{1ig}(w_{1i}), \beta_i) + \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log P(x_{2i} | \alpha_{2ig}(w_{2i}), \beta_i) + \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log P(x_{3i} | \alpha_{3ig}(w_{3i}), \beta_i).
\]

The four parts can be modelled separately. The EM algorithm follows:

**E-step:** calculate

\[
\hat{z}_{ig} = \mathbb{E}(z_{ig} | y_{1i}, y_{2i}, \theta_g)
\]

\[
\hat{s}_{1ig} = \mathbb{E}(X_{2ig} | y_{1i}, y_{2i}; \theta_g)
\]

\[
\log \hat{s}_{1ig} = \mathbb{E}(\log(Y_{1i} - X_{3i}) | y_{1i}, y_{2i}; \theta_g)
\]

\[
\hat{s}_{3ig} = \mathbb{E}(X_{3ig} | y_{1i}, y_{2i}; \theta_g)
\]

\[
\log \hat{s}_{3ig} = \mathbb{E}(\log(X_{3ig}) | y_{1i}, y_{2i}; \theta_g)
\]

**M-step:** Update \( \gamma_{kg} \) which means maximizing (for \( k = 1, 2, 3 \))

\[
\sum_{i=1}^{N} \hat{z}_{ig} \exp(\gamma_{kg}^T w_{ki}) \log \beta_{ig} - \sum_{i=1}^{N} \hat{z}_{ig} \log \Gamma(\exp(\gamma_{kg}^T w_{ki}))+\sum_{i=1}^{N} \hat{z}_{ig} \exp(\gamma_{kg}^T w_{ki}) \log \hat{s}_{1ig}.
\]

Update \( \beta_{ig} \):

\[
\hat{\beta}_{ig} = \frac{\hat{z}_{ig}(\exp(\hat{\gamma}_{1g}^T w_{1i}) + \exp(\hat{\gamma}_{2g}^T w_{2i}) + \exp(\hat{\gamma}_{3g}^T w_{3i}))}{\hat{z}_{ig}(\hat{s}_{1ig} + \hat{s}_{2ig} + \hat{s}_{3ig})}.
\]

When the mixing proportion is regressed on its covariates, it is modelled using a multinomial logistic regression, with

\[
\hat{\tau}_g(w_{0i}) = \frac{\exp(\hat{\gamma}_g^T w_{0i})}{\sum_{g=1}^{G} \exp(\hat{\gamma}_g^T w_{0i})}.
\]
4 Data and Results

A large motor insurance claims dataset is obtained from an Irish general insurance company. Among the different categories covered per motor policy, accidental damage (AD) and third party (TP) property damage (PD) represent the highest correlation, hence their dependence structure is investigated as an example of the bivariate case. Nine covariates are used including policyholder’s information (number of penalty points, licence category), insured car information (fuel type, transmission), policy information (no-claim discount, excess). When using mixture of bivariate Gamma regressions, the best number of potential clusters is one, i.e. there are no underlying groups (selected by BIC). As a comparison, classifications from $G = 2$ and 3 are also shown in Figure 1. When splitting the data into training (70%) and testing (30%) sets, prediction comparisons using Gini index and root mean square error (rMSE) are shown in Table 1, which indicates the claim predictions are improved by MoE model.

![Figure 1](image)

**FIGURE 1.** Classification based on mixture of bivariate Gamma regressions for accidental damage (AD) and TP property damage (PD). $G=1$ is the selected optimal case with BIC = 69176.

<table>
<thead>
<tr>
<th></th>
<th>Gini index</th>
<th>rMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoE prediction of AD</td>
<td>0.669</td>
<td>1328</td>
</tr>
<tr>
<td>GLM prediction of AD</td>
<td>0.541</td>
<td>1976</td>
</tr>
<tr>
<td>MoE prediction of PD</td>
<td>0.683</td>
<td>1316</td>
</tr>
<tr>
<td>GLM prediction of PD</td>
<td>0.518</td>
<td>2427</td>
</tr>
</tbody>
</table>
5 Conclusion

This work investigated the application of multivariate Gamma distributions in the framework of mixture of experts models for insurance claims modelling, showed the claim prediction is improved. One issue with this distribution is the computational complexity on its density function, which has been holding back much of its use in the past. When extending to higher dimensional cases, the computation becomes more complex. It is also noted that covariates for each marginal and covariance part can be different, which can involve variable selection issues for the purpose of model parsimony. This leads to more computational complexity too. It will be part of the future work to develop a more optimal model with variable selection for the insurance data set.

Acknowledgments: This work was supported by the Science Foundation Ireland funded Insight Research Centre (SFI/12/RC/2289).

References


Statistical Analysis of Semiconductor Images

Sarah Karasek\textsuperscript{1}, Herwig Friedl\textsuperscript{1}, Peter Scheibelhofer\textsuperscript{2}

\textsuperscript{1} Institute of Statistics, Graz University of Technology, Austria
\textsuperscript{2} ams AG, Austria

E-mail for correspondence: s.karasek@tugraz.at

Abstract: Wafer including bond voids are a major problem for semiconductor manufacturer because of their insufficient quality. In order to detect the bond voids we take a look at the grayscale image of the wafer where the bond voids occur as light spots. We employ the frequencies of the grayscale values and implement a mixture of distributions to describe the data. Maximum likelihood estimates are calculated by the EM or ECM algorithm and results are provided for mixtures of normals or gammas. This model is often able to correctly classify both, the intensity of the image contrast and the quality of the wafer.

Keywords: Finite Mixtures; EM/ECM Algorithm; Wafer Image Analysis.

1 Introduction

The data to be analysed consists of 113 grayscale images of wafer containing bond voids, which emerge during the production process and denote undesirable air between the layers of a wafer. An automatic detection of the bond voids is only possible, if the image contrast is sufficiently strong enough. In this paper we present a method which allows to classify a wafer referring to its image contrast and quality, especially if the contrast is weak. Instead of working with each pixel of an image, which would be more than 22 million data points, we consider frequencies of grayscales that vary between 0 and 255. While a value of 0 constitutes a black pixel, 255 denotes a pure white one.

Figure 1 shows an original grayscale image with corresponding histogram, where values of 0 have already been omitted because they describe the totally black area outside the wafer disc and are therefore redundant for this analysis. We are interested in the relation between the location of the first peak in this histogram (somewhere between 20 and 40) which describes the bonded area and the last one (in this case at about 80) which denotes...
the bond voids. The larger this distance the stronger the image contrast and therefore the more reliable an automatic detection of the bond voids works.

FIGURE 1. Grayscale image with bond voids as light spots (left) with respective frequency histogram (right)

Since the histogram shows a multi-modal behaviour, a model based on a mixture of distributions might be appropriate. Besides mixtures of normals, which are some sort of a default for such data, we also study mixtures of gammas as proposed in Li et al. (2016) for high-resolution radar images.

2 Mixture of distributions and the EM algorithm

A $K$ component mixture model is marginally defined as the probability density function (pdf)

$$f(y_i|\theta, \pi) = \sum_{k=1}^{K} \pi_k f(y_i|\theta_k), \quad i = 1, \ldots, 255,$$

where $y = (y_1, \ldots, y_{255})$ are the 255 observed grayscale values appearing in an image and $(\theta, \pi) = (\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K)$ denote the component specific parameters and the mixture weights, respectively. Conditionally, $f(y_i|\theta_k)$ stands for the pdf of $y_i$ in component $k$.

Let $z_i = (z_{i1}, \ldots, z_{iK})$ be the missing part of the $i$th grayscale value with

$$z_{ik} = \begin{cases} 1 & \text{if } y_i \text{ belongs to component } k, \\ 0 & \text{otherwise}. \end{cases}$$
Hence the complete pdf of the sample can be written as

\[ f(y, z | \theta, \pi) = \prod_{i=1}^{255} \prod_{k=1}^{K} f(y_i | \theta_k)^{m_i z_{ik}} \pi_k^{z_{ik}}, \]

(2)

with observed frequency \(m_i\) of the \(i\)th grayscale. In order to find the maximum likelihood estimates we apply the EM algorithm (Dempster et al., 1977), where the objective function to be maximised is given by

\[ Q(\theta, \pi | \theta^{(t)}, \pi^{(t)}) = \sum_{i=1}^{255} m_i \sum_{k=1}^{K} w^{(t)}_{ik} \log(\pi_k f(y_i | \theta_k)). \]

(3)

This iterative process starts with appropriate initial values \((\theta^0, \pi^0)\). In the \(t\)th E-step, posterior probabilities

\[ w^{(t)}_{ik} = \frac{\pi^{(t)}_k f(y_i | \theta_k^{(t)})}{\sum_{l=1}^{K} \pi^{(t)}_l f(y_i | \theta_l^{(t)})} \]

(4)

are calculated. For the subsequent \(t\)th M-step these posterior probabilities are considered to be fixed and maximisation of (3) results in updates \((\theta^{(t+1)}, \pi^{(t+1)})\), where the mixture proportions can be updated independently of the specified model and are

\[ \pi^{(t+1)}_k = \frac{1}{N} \sum_{i=1}^{255} m_i w^{(t)}_{ik}, \quad k = 1, \ldots, K. \]

Here, \(N = \sum_{i=1}^{255} m_i\) denotes the total number of pixels in the image. In what follows, mixtures of normals or gammas are considered as suitable choices for \(f(y_i | \theta_k)\). When normals are used, i.e. \(\theta_k = (\mu_k, \sigma^2_k)\), then the parameter updates from the M-step are explicitly given as

\[ \mu^{(t+1)}_k = \frac{\sum_{i=1}^{n} m_i w^{(t)}_{ik} y_i}{\sum_{i=1}^{n} m_i w^{(t)}_{ik}}, \quad \sigma^2_{k(t+1)} = \frac{\sum_{i=1}^{n} m_i w^{(t)}_{ik} (y_i - \mu^{(t+1)}_k)^2}{\sum_{i=1}^{n} m_i w^{(t)}_{ik}}. \]

For gamma distributed components with mean-dispersion parametrisation \(\theta_k = (\mu_k, \phi_k)\), we follow the ideas of Meng and Rubin (1993) and apply the ECM algorithm. This results in the same updates of the proportions and the means as above. However, a conditional M-step is utilized to get updates \(\phi^{(t+1)}_k\) of the dispersion parameters for which

\[ \frac{\partial Q}{\partial \phi_k} = \frac{1}{\phi^{(t)}_k} \sum_{i=1}^{n} m_i w^{(t)}_{ik} \left[ -\log \frac{y_i}{\phi_k \mu^{(t+1)}_k} - 1 + \psi \left( \frac{1}{\phi_k} + \frac{y_i}{\mu^{(t+1)}_k} \right) \right] = 0 \]
with digamma function $\psi(x) = \Gamma'(x)/\Gamma(x)$, has to be solved for $k = 1, \ldots, K$.

The initial values $(\theta^0, \pi^0)$ are drawn randomly from carefully prespecified intervals. The intervals for drawing $(\mu^0_1, \ldots, \mu^0_{K-1})$ are given by the equidistant partition of $(a - 10, a + 20)$, where $a = \arg\max_{y_i} (m_i)$ is the largest mode of the histogram. The outer rightmost $\mu_K$ is initialized by the largest grayscale value for which its associated frequency is above a specified limit like 10,000. In case of the normal model the intervals for the standard deviations $\sigma^0$ are given as $(5, 10)$ for the first $K - 1$ components and $(10, 20)$ for the last one. For a gamma model the respective procedure results in initialisations of the dispersion parameters defined as $\phi_k = \sigma^2_k/\mu^2_k, k = 1, \ldots, K$.

All these models can be fitted with the `flexmix` package by Leisch (2004). However, if we want to additionally restrict the parameters on positive ordered means $\mu_{k-1} < \mu_k$ and positive dispersions, this package either has to be extended or some proper functions have to be developed. A typical parametrisation for the means that ensures such a restriction is $\mu_k = \sum_{j=1}^{k} e^{\alpha_j}$ (partial sums) and a log-linear model $\log \phi_k = \eta_k$ could be used to get positive dispersion parameters. Notice that under such a mean parametrisation the estimates (besides those of the mixture proportions) can no longer be calculated separately for each component.

### 3 Results

To illustrate the knowledge gain of this class of models, we present an example where three gammas are mixed. The rightmost component in the histogram of Figure 2 can be associated with the area on the wafer surface affected by bond voids. Its estimated weight is $\hat{\pi}_3 = 0.034$ which means that about 3.4% of this image relates to bond voids. If we allow at most 5% to be bond voids, we would classify this wafer as a good one. Besides the classification of the wafer quality, we consider the distance between the centre of the leftmost to the rightmost component, $\hat{d} = \hat{\mu}_3 - \hat{\mu}_1$. If this distance is small, as in our example with $\hat{d} = 58.53$, the wafer will be classified as one with a weak image contrast, which makes an automatic detection of bond voids with acoustic microscopy almost impossible. In comparison, images with a strong contrast tend to have a mean component distance of about 150.

Based on the mixture model, we can also improve the visibility of the bond voids by colouring the pixels in the colour of the component they most likely belong to. That means we colour each pixel $y_i$ corresponding to the maximum of its weights $\hat{w}_{ik}$ as shown in the right part of Figure 2. When using a mixture of three components the surface structure of the wafer, especially the horseshoe-shaped contact area as also the individual
semiconductor chips, becomes more visible. A mixture of two components will be sufficient if this information is not needed.

We now apply a three component mixture of gammas onto each of the 113 grayscale images. Their contrast qualities have been also manually classified by an expert in advance. Figure 3 shows the distances $\hat{d}$ plotted against the proportion $\hat{\pi}_3$. The position on the vertical axis reflects the model based classification of the contrasts that precisely separates between the two groups of good and medium/bad wafers.

FIGURE 2. Histogram with mixture of gammas (left) and coloured image (right)

FIGURE 3. Model based classification of wafer’s contrast (good= ■, medium= ●, and bad= ▲) and quality in terms of $\hat{\pi}_3$. 
FIGURE 4. Model based classification of wafer’s contrast (medium = ⬤, and bad = ▲) and quality in terms of $\hat{\pi}_3$.

Distinguishing between images of wafers with medium and bad contrast needs another criterion. A successful strategy is to use the mean of the component with the largest mixture weight ($\hat{\mu}_{max}$) and plot it against $\hat{\pi}_3$. This is illustrated in Figure 4.

In addition, $\hat{\pi}_3$ provides important information about the wafer’s quality. Images of wafers of good quality (few bond voids) are found in the left side of the plot whereas those failing the quality criterion are placed right of a prespecified threshold, e.g. $\hat{\pi}_3 > 0.05$.

References


Effect Selection in Distributional Regression

Nadja Klein¹, Manuel Carlan², Thomas Kneib², Stefan Lang³, Helga Wagner⁴

¹ University of Melbourne
² University of Goettingen
³ University of Innsbruck
⁴ University of Linz

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

Abstract: We propose a spike-and-slab prior specification that allows to carry the concept of Bayesian variable selection over to general effect selection within the class of distributional regression models and with possible hierarchical regression structures. The spike-and-slab prior is assigned to the prior standard deviation of the regression coefficients which allows us to work with a scalar quantity instead of dealing with possibly high-dimensional effect vectors. Furthermore, we specify the model in a redundant parameterisation with parameter expansion that yields improved shrinkage and sampling performance compared to the classical normal-inverse-gamma prior.

Keywords: Additive predictors; hierarchical data; penalised splines; prior elicitation; redundant parameterisation; shrinkage properties; stochastic search variable selection.

1 Introduction

Distributional regression is an important model class that enables estimation of the complete distribution of (multivariate) dependent variables for a broad class of parametric distributions while attaining the predictor flexibility of Generalized Additive Mixed Models. In this paper, we develop an approach that allows us to start from a model including all potentially relevant effects in each predictor of the distribution parameters of any arbitrary parametric response distribution and to estimate the complete underlying conditional distribution with simultaneous effect selection. For that purpose, we use spike-and-slab priors, where a mixture distribution is considered with one component (the spike, which is either a point mass at
Effect Selection in Distributional Regression

zero or a continuous distribution concentrated around zero) corresponding to irrelevant effects that can be excluded from the model while the second component (the slab, which is a weakly informative continuous distribution) corresponds to effects of a size that warrants their inclusion in the model.

2 Bayesian Effect Selection in Distributional Regression

Let \((y_i, \nu_i), i = 1, \ldots, n\) denote \(n\) independent observations on the response variable \(y\) and covariates \(\nu\). We then assume that the conditional distribution of \(y_i\) given \(\nu_i\) is specified in terms of a \(K\)-parametric distribution with density

\[
p(y_i|\vartheta_{i1}, \ldots, \vartheta_{iK})
\]

where \(\vartheta_i = (\vartheta_{i1}, \ldots, \vartheta_{iK})'\) is a collection of \(K\) scalar distributional parameters \(\vartheta_{ik}, k = 1, \ldots, K\) The semiparametric predictors themselves are specified as

\[
\eta_{ik} = \eta_{ik}^{in} + \sum_{j=1}^{J_k} f_{j,k}(\nu_i)
\]

where the effects \(f_{j,k}(\nu_i)\) represent various types of flexible functions depending on (different subsets of) the covariate vector \(\nu_i\) that are to be selected via spike-and-slab priors while \(\eta_{ik}^{in}\) represents a second additive predictor consisting of all effects that are not under selection. Dropping the parameter index \(k\) and the function index \(j\) for notational simplicity, we assume that each effect \(f(\nu_i)\) can be approximated by a linear combination of basis functions such that

\[
f(\nu_i) = \tau \sum_{d=1}^{D} \tilde{\beta}_d B_d(\nu_i)
\]

where \(B_d(\nu_i), d = 1, \ldots, D\) are the basis functions, \(\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_D)^T\) is the vector of (standardised) basis coefficients and \(\tau\) is an importance parameter.

2.1 Constraint Prior for Regression Coefficients

Since for many specific types of effects the vector of basis coefficients \(\beta\) is of relatively high dimension, it is often useful to enforce specific properties such as smoothness or shrinkage. In a Bayesian formulation, this can be facilitated by assuming (partially improper) multivariate Gaussian priors,

\[
p(\tilde{\beta}) \propto \exp\left(-\frac{1}{2} \tilde{\beta}^T K \tilde{\beta}\right) 1\left[A \tilde{\beta} = 0\right]
\]
where we assume that the constraint matrix $A$ is always chosen such that all rank-deficiencies in $K$ are effectively removed from the prior distribution. This can, for example, be achieved by setting

$$A = \text{span} \left( \ker(K) \right)$$

where $\ker(K)$ denotes the null space of $K$ and $\text{span} \left( \ker(K) \right)$ a representation of the corresponding basis.

To achieve function selection in our model, we place a spike-and-slab prior specification on the squared importance parameter $\tau^2$. This hierarchical prior relies on a mixture of one prior concentrated close to zero such that it can effectively be thought of as representing zero (the spike component) and a more dispersed, mostly noninformative prior (the slab) and is specified as

$$
\begin{align*}
\tau^2 | \delta, \psi^2 & \sim \text{Ga} \left( \frac{1}{2}, \frac{1}{2r(\delta)\psi^2} \right) \\
\delta & \sim \text{Bi}(1, \omega) \\
\psi^2 & \sim \text{IG}(a, b) \\
\omega & \sim \text{Beta}(a_0, b_0) \\
r(\delta) & = \begin{cases} r & \delta = 0 \\ 1 & \delta = 1 \end{cases}
\end{align*}
$$

The scale parameter $\psi^2$ determines the prior expectation of $\tau^2$, which is $\psi^2$ for $\delta = 1$ and $r\psi^2$ for $\delta = 0$, $r \ll 1$ is a fixed small value and hence the indicator $\delta$ determines whether a specific effect $\beta = \tau\tilde{\beta}$ is included in the model ($\delta = 1$) or excluded from the model ($\delta = 0$), $\omega$ is the prior probability for an effect being included in the model and the remaining parameters $a$, $b$, $a_0$ and $b_0$ are hyperparameters of the spike-and-slab prior.

The complex hierarchical spike-and-slab specification involves a large number of hyperparameters and thus, while there are feasible guidelines for $a$, $a_0$ and $b_0$ we still have to deal with the selection of $r$ (which should be small to effectively remove components from the model) and the hyperparameter $b$ of the inverse gamma hyperprior of $\psi^2$. We start by considering a specific type of effect $f(\nu)$ and formulate marginal prior probabilities for the inclusion/exclusion of effects of a certain size. More precisely, we consider probability statements on the largest absolute effect $|f(\nu)|$ over a certain set of covariate values $D$ given the status of the inclusion/exclusion parameter $\delta$. Given $\delta = 1$ (inclusion of the effect), the marginal distribution of $f(\nu)$ does no longer depend on $r$ such that the parameter $b$ can be determined from

$$
\mathbb{P}\left( \sup_{\nu \in D} |f(\nu)| \leq c \mid \delta = 1 \right) = \alpha
$$

which is the probability of including an effect that is in absolute terms smaller than a pre-specified level $c$ for all design points $\nu \in D$. Both the
level $c$ and the prior probability $\alpha$ have to be specified by the analyst according to her prior beliefs. To derive $r$, we proceed similarly but consider the probability
\[
\mathbb{P} \left( \sup_{\nu \in D} |f(\nu)| \leq c \mid \delta = 0 \right) = 1 - \alpha
\]
i.e. the probability of excluding an effect $f(\nu)$ which is in absolute terms smaller than the pre-specified level $c$. Since in this case we would rather be interested in making this probability large (irrelevant effects shall be excluded), the probability is reversed to $1 - \alpha$. Note that the absolute value of the effects can be taken without loss of generality due to the centring constraint of each function to ensure identifiability.

The basic idea of these two equations is that such prior statements can be much more easily elicited in applications, in particular in distributional regression where the application of response functions such as the exponential function or the logit transform induce default ranges of plausible effect sizes.

### 3 Simulation Study

In our simulation study we follow Scheipl et al. (2012) and define 4 functions:

- $f_1(x) = x$
- $f_2(x) = x + \frac{(2x-2)^2}{5.5}$
- $f_3(x) = -x + \pi \sin(\pi x)$
- $f_4(x) = 0.5x + 15\phi(2(x - 0.2)) - \phi(x + 0.4)$

and generate 16 covariates an from $i.i.d \sim U[-2, 2]$. In the case of zero-inflated Poisson responses, we distinguish two scenarios in terms of predictor complexity. These are

- **un sparse** in which out of 16 included covariates 12 have non zero influence. The true linear predictor for $\lambda$ is $\eta_\lambda = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + 1.5 (f_1(x_5) + f_2(x_6) + f_3(x_7) + f_4(x_8)) + 2(f_1(x_9) + f_2(x_{10}) + f_3(x_{11}) + f_4(x_{12})$ and we distinguish the two cases with additional and without additional spatial effect $f_{spat}(s)$, labeled as ‘spatial/non-spatial’.

- **sparse** in which out of eight included covariates four have non-zero influence. The true linear predictor for $\pi$ is $\eta_\pi = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4)$ and we again distinguish the two cases with additional and without additional spatial effect $f_{spat}(s)$.

Additionally, we optimize for different combinations of probability $\alpha$ and threshold $c$. Results for the estimated inclusion probabilities are shown in Figure 1. Estimations are obtained by the software tool BayesX (Belitz et al. (2009)).
FIGURE 1. Posterior inclusion probabilities of effects from the NBPSS prior for the zero-inflated Poisson model with \( n = 1,000 \) observations, uncorrelated covariates and a true spatial effect in the predictor. Blue boxplots correspond to effects that are included in the true model while the red boxes correspond to the noise variables that do not have an effect in the data generating mechanism.
References


Identification issues and efficient inference in semi-parametric cure survival model with shared covariates

Philippe Lambert\textsuperscript{1,2} and Vincent Bremhorst\textsuperscript{2}

\textsuperscript{1} Faculté des sciences sociales, Université de Liège, Belgium  
\textsuperscript{2} Institut de statistique, biostatistique et sciences actuarielles (ISBA), Université catholique de Louvain, Belgium

E-mail for correspondence: p.lambert@uliege.be - www.statsoc.ulg.ac.be

Abstract: Cure survival models are used when one wants to acknowledge that a partially unidentified proportion of subjects will never experience the event of interest. We focus here on the promotion time cure model in the general case where covariates are used jointly to describe changes in the probability to be ‘cured’ (i.e. immune to the event of interest) and in the semi-parametric distribution of event times for susceptible subjects. It is shown that identification issues arise for regression parameters of shared covariates when the maximum follow-up duration is insufficiently long and that large posterior correlation between these parameters cannot be avoided otherwise. A reparametrization of the model is proposed to handle these issues and shown to provide well mixing chains when sampling the joint posterior in a Bayesian approach or easily converging algorithms in the derivation of MLEs in likelihood based frequentist estimation strategies. We illustrate the methodology with the analysis of the time to death in patients suffering from oropharynx carcinoma.

Keywords: Cure survival models; Efficient inference; Identification issues.

1 The promotion time model

Let $Y_i$ and $C_i$ be the event and right censoring times for the $i$th of $n$ units ($i = 1, \ldots, n$). Suppose that, conditionally on covariates, $Y_i$ and $C_i$ are independent and that $(T_i, \delta_i)$ is observed where $T_i = \min\{Y_i, C_i\}$ and $\delta_i = I(Y_i < C_i)$.

Further assume that among the right censored subjects, an unidentified proportion of them are immune to the event of interest: the ‘cured’ units.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
The promotion time model (Yakovlev & Tsodikov 1996) acknowledges this by assuming that (marginally over the subject cure status), the population survival function has a non-zero limiting value for large values of time:

$$S_p(y_i|x_i, z_i) = \Pr(Y_i > y_i|x_i, z_i) = \exp\{-\theta(x_i) F(y_i|z_i)\}, \quad (1)$$

with $S_p(+\infty|x_i,z_i) = \exp\{-\theta(x_i)\}$ for some covariates $(x'_i, z'_i)$ in $\mathbb{R}^{p_1+p_2}$.

Given that the conditional cumulative hazard, $H_p(y|x_i, z_i) = -\log S_p(y|x_i, z_i)$, must be a nondecreasing function of time, it implies that $F(\cdot|z_i)$ is a conditional distribution function and that $H_p(y|x_i, z_i) \leq \theta(x_i)$ at any time $y$, explaining why some authors name the promotion time model more explicitly the Bounded Cumulative Hazard (BCH) model. See also Yakovlev & Tsodikov (1996) for a biological motivation of the model.

We partially follow Bremhorst & Lambert (2016) with a log-linear model for $\theta(x_i)$, $\log \theta(x) = \eta_\theta(x) = \beta_0 + x'\beta$, and a semi-parametric proportional hazards model to describe the effect of covariates on the dynamics in the normalized cumulative hazard, $F(y|z) = 1 - S_0(y)^{\exp(\eta_F(z))}$ with $\eta_F(z) = z'\gamma$. The baseline survival function $S_0(t)$ is specified through the log of the baseline hazard $h_0(t)$ written as a linear combination of $K$ cubic B-splines associated to equidistant knots on $(0, t_{\text{max}})$ where $t_{\text{max}}$ denotes the minimum follow-up duration required to ensure that an event-free subject by that time will not experience the event of interest. We take a large number of knots and penalize $r$th-order differences of successive B-spline coefficients $\phi = (\phi_1, \ldots, \phi_K)'$ during estimation (Eilers & Marx 1996). In a Bayesian framework, it can be translated into a smoothness prior $p(\phi|\lambda)$ for $\phi$ with a robust (gamma mixture) prior (Jullion & Lambert 2007) for the penalty parameter $\lambda$.

2 Inference

2.1 Identification issues

To simplify the presentation, assume that the two set of covariates share one component $x_2$ with respective linear predictors $\eta_\theta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ and $\eta_F(x) = \gamma_2 x_2 + \gamma_3 x_3$. Then, when the follow-up time $t$ is insufficiently long to ensure that the event on non-immune subjects is observed, the baseline normalized cumulative hazard function $F_0(t) = F(\cdot|0) = 1 - S_0(t)$ typically has a small underlying value, implying that

$$F(t|x) = 1 - (1 - F_0(y))^{\exp(\eta_F(x))} = \exp(\eta_F(x) F_0(t)) + O\left(F_0(t)^2\right).$$

Thus, when $F_0(t)$ is small,

$$\log H_p(t|x) = \log(\theta(x) F(t|x))$$

$$\approx (\beta_0 + \beta_1 x_1 + (\beta_2 + \gamma_2) x_2 + \gamma_3 x_3).$$
Therefore, when the observed maximum follow-up time $t_R$ does not guarantee that non-immune subjects experience the event of interest by that time, identification issues arise for the intercept $\beta_0$ and the regression coefficients of shared covariate(s) (here $x_2$), as they only appear either summed to $\log F_0(t)$ or through their sum $(\beta_2 + \gamma_2)$ in the preceding approximation. A large negative posterior correlation is expected between $\beta_2$ and $\gamma_2$ as an increase (resp. decrease) in $\beta_2$ can be approximately compensated by a similar decrease (resp. increase) in $\gamma_2$. For a sufficiently long maximum follow-up duration, parameters can be identified, but a non negligible posterior correlation between regression coefficients of shared covariates is still expected.

2.2 Model reparametrization

For these reasons, we suggest to reparametrize the model to improve mixing in MCMC chains and to increase the chance of a worriless convergence in the derivation of MLEs in a frequentist estimation of the model parameters. More specifically, we suggest to consider $(\beta_0, \beta, \gamma) \rightarrow (\beta_0, \psi, d\psi)$ where, in our example, $\psi_1 = \beta_1, \psi_2 = (\beta_2 + \gamma_2)/2, d\psi_2 = (\beta_2 - \gamma_2)/2, \psi_3 = \gamma_3$. Thus, for regression parameters of shared covariates, $\beta_k = \psi_k + d\psi_k$ and $\gamma_k = \psi_k - d\psi_k$. In a Bayesian framework, we also found it useful to take $\phi_k = P - \exp(\zeta_k)$ for some number $P > 0$ known contextually to be larger than any of the possible values of the spline parameters. Then, the roughness penalty is taken on vector $\zeta$.

2.3 Inference

The log-likelihood is obtained from the (population) cumulated hazard and hazard functions, $\ell(\beta_0, \beta, \gamma; D) = \sum_i \{\delta_i \log h_p(t_i | x_i, z_i) - H_p(t_i | x_i, z_i)\}$. After reparametrization, the joint posterior is given by

$$p(\beta_0, \psi, d\psi, \phi, \lambda, \delta | D) \propto L(\beta_0, \psi, d\psi, \phi; D)p(\beta_0)p(\psi)p(d\psi)p(\phi|\lambda)p(\lambda|\delta)p(\delta).$$

It can be explored using the Metropolis-within-Gibbs algorithm with Gibbs steps for the penalty parameters. Non reported simulation results confirm the preceding theoretical expectations with unbiased estimation of the regression parameters when the follow-up is sufficiently long. When it isn’t, only regression parameters of non-shared covariates and the sum of regression parameters of shared covariates are identified. The coverages of credible intervals (CI) are compatible with their nominal values in models with correctly selected covariates. When a covariate is assumed to simultaneously influence the cure probability and the event timing for susceptible subjects, coverages of CI for the corresponding regression parameters remain compatible with their nominal values provided that the covariate truly affects the cure probability. A small undercoverage was revealed otherwise.
3 Application

We illustrate the methodology and identifications issues on oropharynx carcinoma data (Kalbfleisch & Prentice, 1980). We focus on the effects of Tumor staging (0: Primary; 1: Invasive) and Treatment (0: Standard; 1: Test) on the probability to be cured and on event timing for non-cured subjects. It involves \( n = 130 \) patients with 38 of them right censored. The Kaplan-Meier curves on Fig. 1 suggest a strong Tumor staging effect while the plateaus over 3 years probably indicate that the subjects still in the risk sets over that time are actually cured.

The flexible promotion time model of Section 1 (with \( t_{\text{max}} = 5 \) and \( K = 10 \)) with \( x_1 = z_1 = \text{Tumor} \) and \( x_2 = z_2 = \text{Treatment} \) was fitted after reparametrization and under the Bayesian paradigm, see Section 2. The covariates were present simultaneously in the two regression parts of the model. Figure 2 display the pointwise estimate and the 90\% quantile credible intervals for the regression parameters for increasing follow-up durations starting at year 1. The left panels clearly highlight the large uncertainty and negative correlation in the estimation of \( (\beta_k, \gamma_k) \) when the follow-up is insufficiently long to clearly reveal the cured fraction. On the other hand, their half-sum \( \psi_k \) is estimated with precision much earlier (see the top right panel of Fig. 2). After 2.5 years, we already have significant indications that, for a given treatment, patients with an invasive tumor are less likely to be cured (as we only have positive plausible values for \( \beta_1 \)). It is only after 3 years and the last death due to cancer that (conditionally on tumor staging) significant indications of a Treatment effect on event timing for non-cured patients can be pointed (with only positive plausible values for \( \gamma_2 \)). However, it does not translate into a significant effect on long-term prognosis (as 0 is part of the plausible values for \( \beta_2 \)).

![FIGURE 1. Tongue cancer data: Kaplan-Meier curves.](image-url)
FIGURE 2. Tongue cancer data: Posterior median and 90% credible interval for the regression parameters for increasing maximum follow-up durations in the $\beta$–$\gamma$ (left) and $\psi$–$d\psi$ (right) parametrizations.

References


International Arms Transfers: A Dynamic Separable Network Model With Heterogeneity Components

Michael Lebacher\textsuperscript{1}, Paul W. Thurner\textsuperscript{2}, Göran Kauermann\textsuperscript{1}

\textsuperscript{1} Department of Statistics, LMU Munich, Germany
\textsuperscript{2} Geschwister-Scholl Institute of Political Science (GSI), LMU Munich, Germany

E-mail for correspondence: michael.lebacher@stat.uni-muenchen.de

Abstract: We investigate network data on international transfers of major conventional weapons from 1950 to 2016. The statistical model is based on a separable temporal exponential random graph model (StERGM). In order to provide a realistic framework for the data we extend the model as follows. (i) We allow for time-varying covariate effects. (ii) The actor-related heterogeneity is incorporated by actor specific time-varying random effects. In an additional step, the smooth random effect curves are subjected to a functional principal component analysis.

Keywords: Arms Trade; Functional Principal Component Analysis; Generalized Additive Mixed Models; Network Analysis; Varying Coefficient Model.

1 Data and Modelling Strategy

In this paper we analyse the international exchange of arms. The data is collected by the Stockholm International Peace Research Institute (SIPRI, 2017) and provides information on the trade of major conventional weapons (tanks, aircrafts, battleships, etc.). Dichotomizing the information whether there is trading or not, leads to a sequence of yearly binary networks as shown for the year 2016 in Figure 1. In order to formalize the approach, let \( Y^t \) be the binary network at time point \( t \), which consists of a set of \( n_t \) actors, labelled \( A^t \) and a set of directed edges, represented by the index set \( E^t = \{(i, j) : i, j \in A^t\} \). We set \( Y^t_{ij} = 1 \) if country \( i \) exports weapons to country \( j \) and zero otherwise. The basic model is a separable temporal exponential random graph model (StERGM) as proposed by Krivitsky and Handcock (2014). This model describes transition dynamics by considering separately the formation of new edges and the persistence of existing edges and assuming a first-order Markov structure. Let therefore \( Y^+ = Y^t \cup Y^{t-1} \) represent the formation network, that consists of edges that are present in \( t \).
or in \( t - 1 \). The persistence network is defined as \( Y^- = Y^t \cap Y^{t-1} \), consisting of edges that are present in \( t \) and in \( t - 1 \). Additionally, we assume that for each discrete time step, the processes of formation and persistence are separable (i.e. \( Y^+ \) and \( Y^- \) are independent given \( Y^{t-1} \)):

\[
P(Y^t = y^t | Y^{t-1} = y^{t-1}; \theta) = P(Y^+ = y^+ | Y^{t-1} = y^{t-1}; \theta^+) P(Y^- = y^- | Y^{t-1} = y^{t-1}; \theta^-),
\]

with \( \theta^+ \) and \( \theta^- \) denoting the parameter vectors. A central model class for binary networks are exponential random graph models (ERGMs, see for example Robins et al., 2007), but we follow Almquist and Butts (2014) and restrict the model such that for some vector of network statistics \( \tilde{g}(\cdot) \) the following logistic model holds for the formation model

\[
\log \left\{ \frac{P(Y^+_{ij} = 1 | Y^{t-1} = y^{t-1}; \theta^+)}{P(Y^+_{ij} = 0 | Y^{t-1} = y^{t-1}; \theta^+)} \right\} = \tilde{g}^T_{ij}(y^{t-1}) \theta^+.
\]

The analogous model is assumed for the persistence network \( Y^- \). In order to identify the temporality of effects we replace the parameter \( \theta^+ \) by \( \theta^+(t) \), representing a smooth function in time, leading to a varying coefficient model (see Hastie and Tibshirani, 1993). We follow the modelling strategy of Durbán et al. (2005) and allow for actor- and time-related heterogeneity by inclusion of smooth, time-varying random effects for the sender and
receiver states. With this model we can investigate the (hyper-) dyadic network effects shown in Figure 2 together with political and economic covariates as for example the logarithmic GDP (Gleditsch, 2013; World Bank, 2017) of the exporting and importing state. Dissimilarities between political regimes are measured by the absolute difference of the so-called polity IV score of the exporter and the importer (Center for Systemic Peace, 2017). Additionally we control for formal alliances, the logarithmic military personnel of the receiver and whether the exporter had colonial power over the importer in the past (Correlates of War, 2017).

For the estimation we use the generalized additive mixed model framework provided by Wood (2017) implemented in the R package mgcv. The fitted multivariate time series of random effects are subjected to a functional principal component analysis (Ramsay and Silverman, 2005) in order to explain which share of variance is induced by different overall levels of the random effect curves and which share can be attributed to dynamics.

2 Summary of the Results

We omit the effects of political and economic variables due to space restrictions but our results regarding network effects are shown in Figure 3. The coefficients of the formation model ($\theta^+$) are shown in the left column while the coefficients for the persistence model ($\theta^-$) are given in the right column. Hence we can interpret positive coefficients on the left (right) side as increasing the log-odds of forming new (maintaining existing) arms trade relationships. In the top row it can be seen that the outdegree of the sender is not a significant determinant of the exports in neither of the two models. The outdegree of the importer, however, has a negative impact on the formation. We find strong and positive effects for reciprocity, transitivity and shared-suppliers in the formation model. In particular, the dynamics of the transitivity parameter mirror how massively the end of the cold war has disrupted the international arms transfers. In the persistence model transitivity has a constant positive effect. The initially negative effect of shared-suppliers becomes insignificant with time while its the other way round for reciprocity.
Regarding actor heterogeneity, the analysis of the random effects reveals high variation among countries. The functional principal component analysis for the random effect curves is shown in Figure 4 for the formation model (top six panels) and in the persistence model (bottom six panels). For each model-effect combination, the scores of the first two principal components are plotted on the left hand side where the latter are visualised on the right hand side. The first component represents the different overall levels of the random effect curves and the dynamic of the random effects is captured by the second component, delivering a tendency for an upward movement if positive and downward if negative. This allows to single out "special" countries, as for example Israel (ISR) and the United Arab Emi-
FIGURE 4. Functional Principal Component Analysis of the smooth subject-specific Random Effects for the Formation (top) and Persistence Model (bottom).
rates (ARE) that take polar positions in the plots. Israel has high values for the first component as sender in the formation and persistence model (Figure 4 first and third plot on the left), i.e. Israel has a strong tendency to be the exporter in new trade relations as well as in persistent ones. The scores of the United Arab Emirates (ARE) on the other hand are located in the first quadrant of the plots for the receiver effect in the formation and persistent model (Figure 4 second and fourth plot on the left), which means that the United Arab Emirates are strong importers but also that the country has strongly increased its tendency to be the importer in persistent and occasional arms transfers. Besides that, other noticeable countries are for example Japan (JPN), North Korea (PRK) and Ukraine (UKR).

A careful evaluation shows that the model is able to reproduce the observed global network structures and provides good out-of-sample forecasts.

References


Gleditsch, K.S. (2013). Expanded trade and GDP data. privatewww.essex.ac.uk/ksg.


Predicting above-ground tree biomass: a
Seemingly Unrelated Penalized Regression
approach

Dae-Jin Lee¹ and María Durbán²

¹ BCAM - Basque Center for Applied Mathematics, Bilbao, Bizkaia, Spain
² Escuela Politécnica Superior Leganés, Universidad Carlos III de Madrid, Spain

E-mail for correspondence: dlee@bcamath.org

Abstract: The estimation of above-ground biomass in forests is critical for the development of crop management strategies, carbon cycle modelling and climate change mitigation programs. Estimating above-ground biomass and its components requires statistical methods to improve the accuracy of predictions using easily obtainable information such as diameter at breast height (dbh) and height of trees. In this work, we propose a flexible approach based on the use of penalized splines with shared random effects, we called this approach: Seemingly Unrelated Penalized Regression (SUPR) model.

Keywords: Aboveground tree biomass; penalized splines, seemingly unrelated regression.

1 Introduction

Above-ground biomass includes all biomass in living vegetation, both woody and herbaceous, above the soil including stems, stumps, branches, bark, seeds and foliage. The work by Parresol (2001), has become the standard approach to develop non-linear tree biomass equations that ensures compatibility among total and component predictions. This author proposed a procedure that guarantees the property of additivity among the components of tree biomass and total tree biomass, based on Non-linear Seemingly Unrelated Regression (NLSUR).
2 Biomass models and additivity property

For generality, let $w_1, w_2, \ldots, w_M$ denote the $M$ biomass components of a given tree and $w_t$ the total biomass, i.e. $w_t = \sum_{m=1}^{M} w_m$. Parresol (2001) noted that in the non-linear setting, additivity of component estimates can be guaranteed only by restricting the total biomass equation to be the sum of the component biomass equations. In contrast, Sánchez–González et al. (2017), proposed the use of penalized splines within an additive mixed model framework for the estimation and prediction of above-ground biomass data. This approach results in a much simpler alternative to NL-SUR models which equations are based on complex non-linear biomass equations that have to be chosen ad hoc for each dataset. The additivity property is satisfied by fixing the smoothing parameter ($\lambda$) for the total quantity (i.e. $\hat{w}_t$) and then fit a penalized spline model for the rest $m$ components with $\lambda$. In general, fixing the amount of smoothness for the total quantity is not a problem, since, in most cases, the other measurements present similar optimal smoothing parameters.

3 Seemingly Unrelated Penalized Regression approach

We propose a model in which a set of penalized splines models are related using a shared random component that connects the set of equations (Eilers et al. 2011). To illustrate the approach, let us consider a simple case, where $w_t$ is the total above-ground biomass of a sample of $n$ trees, a single covariate ($x$), and two tree components (i.e. $M = 2$), for instance $s$ (stem) and $b$ (branch). We aim to model the set of equations:

$$
\begin{align*}
    w_s &= f_s(x) + u + \epsilon_s \\
    w_b &= f_b(x) + u + \epsilon_b
\end{align*}
$$

where $u$ is a “shared random-effect” which relates both equations and $\epsilon_s$ and $\epsilon_b$ are error terms. Let be $f_s(x)$ and $f_b(x)$ smooth unknown functions of the covariate $x$ modeled by P-splines with a combination of a B-spline basis $\tilde{B} = B(x)$ of size $n \times c$. We define the model matrix $B$ in compact notation as $B = (I_2 \otimes \tilde{B} : I_2 \otimes I_n)$, where the second block relates both Eq.(1) and Eq.(2) and $I_2$ is a diagonal matrix of size 2 and $I_2$ a $2 \times 1$ column vector of one’s. The vector of coefficients $\theta = (a_s, a_b, u)^T$ are penalized by the penalty matrix of this SUPR model is $P = \text{blockdiag}(\lambda_s D^T D, \lambda_b D^T D, \kappa I_n)$, where $D$ is a $q$ order difference matrix, $\lambda_s$ and $\lambda_b$ are the smoothing parameters for $w_s$ and $w_b$ respectively and $\kappa I_n$ is a ridge penalty on the shared random effects $u$. The systems of equations is sparse and it is given by:

$$
\begin{pmatrix}
    B^T B + \lambda_b D^T D \\
    B^T B + \lambda_b D^T D \\
    B \\
\end{pmatrix}
\begin{pmatrix}
    \hat{a}_s \\
    \hat{a}_b \\
    \hat{u}
\end{pmatrix}
= 
\begin{pmatrix}
    w_s \\
    w_b \\
    w_s + w_b
\end{pmatrix}.
$$
Note that, the last row of the system of equations ensures the additivity property, i.e. \( \hat{w}_t = \hat{w}_s + \hat{w}_b \). Moreover, this model allows for the estimation of the smoothing parameters \( \lambda_s, \lambda_b \) separately (i.e. \( \lambda_s \neq \lambda_b \)) and the parameter \( \kappa \) using standard optimization criteria such as Akaike or Bayesian Information (AIC or BIC). The extension to the additive model case is straightforward. Given a second covariate \( z \), we include \( f_s(z) \) and \( f_b(z) \) in Eq.(1) and Eq.(2) and extend the regression basis \( B \) and the penalty \( P \) block-wise accordingly with two extra smoothing parameters. A small ridge penalty, \( 10^{-4} \) is included for identifiability.

4 Application to Populus genus above-ground biomass data

We illustrate the SUPR approach with above-ground biomass data from Southern Mediterranean Populus genus. The data was analysed in Sánchez et al. (2017) for several clones. The production was measured by recording total aboveground dry biomass (Mg DM ha\(^{-1}\)-BT) and as covariate we consider the diameter (in mm) at 1.30 m (dbh). Stem and branch biomass were weighted separately and used to assess the biomass distribution. Figure 1, shows on the left the predicted above-ground biomass with the proposed SUPR model for “2000 verde” clone. Table 1 shows for each component,
the root mean squared errors (RMSE) of the P-spline approach proposed in Sánchez–González et al. (2017) and SUPR. Notice that, while the RMSE’s for $w_t$ and $w_s$ are very close, for $w_b$ the RSME value for $w_b$ is much lower.

TABLE 1. Root Mean Square Error of total ($w_t$), stem ($w_s$) and branch ($w_b$) above-ground biomass.

<table>
<thead>
<tr>
<th></th>
<th>$w_t$</th>
<th>$w_s$</th>
<th>$w_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-spline</td>
<td>0.1988</td>
<td>0.2078</td>
<td>1.3143</td>
</tr>
<tr>
<td>SUPR</td>
<td>0.1989</td>
<td>0.2076</td>
<td>0.0935</td>
</tr>
</tbody>
</table>

FIGURE 2. Predicted versus observed of total ($w_t$), stem ($w_s$) and branch ($w_b$) above-ground biomass.

5 Conclusions

We propose an alternative penalized spline approach for above-ground biomass prediction. We use a shared random effect approach similar to a
seemingly unrelated regression model with a simplified variance-covariance structure. The additivity property of the total biomass is achieved by setting the model equations. Also, the SUPR models allow for more flexibility for each tree component.

**Acknowledgments:** This research was supported by the Basque Government through the BERC 2018-2021, the Department of Education, Language Policy and Culture of the Basque Government IT-620-13 programs and by Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and through projects MTM2017-82379- R funded by (AEI/FEDER, UE) and acronym “AFTERAM” and MTM2014-52184-P.

**References**


Quantifying the impact of air pollution on respiratory prescription rates in Scotland

Duncan Lee

1 University of Glasgow, Glasgow, Scotland

E-mail for correspondence: Duncan.Lee@glasgow.ac.uk

Abstract: Air pollution and health studies focus on severe health endpoints such as deaths, but this underestimates its health impact because it ignores ill health treated in primary care. Here we quantify the impact of air pollution on respiratory medication rates in primary care in Scotland.

Keywords: Air pollution; Respiratory prescriptions; Spatio-temporal modelling.

1 Introduction

Epidemiological research into air pollution has focused on severe health endpoints, such as hospital admissions or deaths, which underestimates its health impact because a large amount of ill health is treated in primary (non-hospitalised) care. Blangiardo et al. (2016) is one of the only papers to have quantified the impact of air pollution on ill health in primary care, focusing on rates of respiratory medication. They aggregated the data to non-overlapping areal units and used a conditional autoregressive (CAR) model for inference. However, primary medical care is provided by doctors located within General Practice (GP) surgeries, who prescribe the required medications. Here we quantify the impact of air pollution on respiratory prescription rates at the GP surgery point level. Each GP surgery has a single geographical coordinate, while its patient population is drawn from the surrounding area. Exploratory analysis shows that some pairs of spatially close GP surgeries have very similar data values, suggesting autocorrelation, while other pairs have very different values, suggesting no autocorrelation. Furthermore, the autocorrelation does not always decay with increasing distance apart, because it arises regularly that for GP surgeries \((i, j, k)\) all close together, surgery \(i\) is geographically closer to surgery \(j\) than surgery \(k\), but has a data value closer to that from surgery \(k\) than...
surgery \( j \). Therefore we propose a novel locally adaptive bivariate spatio-temporal process-convolution (PC, Higdon 1998) model, where only nearby weights are non-zero, and within this neighbouring set the weights do not necessarily decay with increasing distance apart.

2 Data

Data are available at a monthly resolution between October 2015 and July 2016 \((T = 10\) months) for \( K = 939 \) GP surgeries across Scotland. For each GP surgery and month we have the total number of prescriptions for Salbutamol (100mcg or 200mcg) and Ventolin (100mcg or 200mcg), which are the two most commonly used respiratory medications. The number of prescriptions at each GP surgery will depend on the size and age-sex structure of its patient population. This is accounted for by computing the expected numbers of respiratory disease sufferers registered with each GP surgery using indirect standardisation, which are then scaled so that the total observed and expected numbers are the same. Concentrations of particulate air pollution (PM\(_{10}\) and PM\(_{2.5}\)) were estimated at each GP surgery location, using the modelled annual average concentrations provided by the Department for Environment, Food and Rural Affairs. Additionally, covariate information quantifying poverty and ethnicity were also collected, as they are also known to be drivers of respiratory prescription rates.

3 Methodology

Let \( s_k = (s_{1k}, s_{2k}) \) denote the geographical coordinate (easting, northing) of the \( k \)th GP surgery, and let \((Y_t(s_k), E_t(s_k))\) respectively denote the observed number of prescriptions and the scaled expected number of respiratory disease patients for the \( k \)th GP surgery and \( t \)th month. Finally, let \( x_t(s_k) = (1, x_{t2}(s_k), \ldots, x_{tp}(s_k)) \) denote a \( p \times 1 \) vector of covariates including an intercept term. Then the proposed Bayesian hierarchical model is given by

\[
Y_t(s_k) \sim \text{Poisson}(E_t(s_k)R_t(s_k)),
\]

\[
\ln(R_t(s_k)) = x_t(s_k)^\top \beta + \sum_{j=1}^{K} w_{kj}\theta_t(s_j).
\]

Here \( R_t(s_k) \) denotes the rate of prescription, and we model the residual spatio-temporal autocorrelation via the process-convolution \( \sum_{j=1}^{K} w_{kj}\theta_t(s_j) \). Spatial autocorrelation is induced by the \( K \times K \) matrix of weights \( W = (w_{kj})_{K \times K} \), while temporal autocorrelation is induced by the first order autoregressive process prior:
Here $\gamma$ controls the temporal autocorrelation in the spatially unstructured latent process. To reduce the computational demand of fitting the model we propose applying a tapering function $I(s_k, s_j)$ to the weights contained in $W$, so that only the $m$ spatially closest elements $s_j$ to $s_k$ have a non-zero weight $w_{kj}$. That is, $I(s_k, s_j) = 1$ if surgery $j$ is one of the $m$ closest to surgery $k$, and $I(s_k, s_j) = 0$ otherwise. The resulting $W$ matrix is sparse with only $mK$ non-zero elements, which means that when updating $\theta_t(s_k)$ in a Markov chain Monte Carlo (MCMC) algorithm the data likelihood only needs to be evaluated at a small number of data points $Y_t(s_j)$ rather than at all $K$. In the analysis we fit the models with $m = 4, 8, 16$, and find the results are robust to the choice of this tapering parameter. We compare two specifications for the weights, the first of which is based on a Gaussian distance-decay Kernel function.

\[
w_{kj} = \frac{I(s_k, s_j) \frac{1}{\sqrt{2\pi/\alpha}} \exp \left( -\alpha \frac{||s_k - s_j||}{2} \right)}{\sum_{i=1}^{K} I(s_k, s_i) \frac{1}{\sqrt{2\pi/\alpha}} \exp \left( -\alpha \frac{||s_k - s_i||}{2} \right)}.\]

(3)

The second approach we consider for $W$ allows the weights not to obey a distance-decay form, because after covariate adjustment for a given GP surgery the spatially closest neighbouring GP surgery had the most similar rate in only 18% of cases. Thus we ensure spatial autocorrelation by only allowing the weights for the $m$ closest GP surgeries to be non-zero, but then within this group allow random weights via the Dirichlet prior:

\[
w_{kj} = \begin{cases} 
\psi_{kr} & \text{if } s_j \text{ is the } r\text{th closest to } s_k \\
0 & \text{Otherwise}
\end{cases},
\]

(4)

\[
\psi_k = (\psi_{k1}, \ldots, \psi_{km}) \sim \text{Dirichlet}(\alpha_1 = 1, \ldots, \alpha_m = 1).
\]

Inference for both weight models is implemented within a Bayesian context via MCMC simulation.

4 Results

Inference was based on 30,000 MCMC samples generated from 3 parallel Markov chains, that were each burnt in for 100,000 iterations at which point convergence was assessed to have been reached. A further 100,000 samples were then generated and thinned by 10 to reduce the autocorrelation. The results relate to the tapering parameter value $m = 8$, although the results for the other values of $m$ (4 and 16) are almost identical and are not shown.
TABLE 1. Estimated relative rates and 95% credible intervals for the covariates. The results relate to both models (3) and (4) with $m = 8$. The numbers in brackets in the first row are standard deviations (SD) for the covariate, and are the size of the increase that the relative rates relates to.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Relative rate</th>
<th>Model (3) $m = 8$</th>
<th>Model (4) $m = 8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM$_{2.5}$ (SD=1.23µgm$^{-3}$)</td>
<td>1.016 (1.003, 1.030)</td>
<td>1.022 (1.005, 1.040)</td>
<td></td>
</tr>
<tr>
<td>PM$_{10}$ (SD=1.84µgm$^{-3}$)</td>
<td>1.016 (1.002, 1.030)</td>
<td>1.016 (0.999, 1.035)</td>
<td></td>
</tr>
<tr>
<td>Property price (SD = £57,400)</td>
<td>0.839 (0.829, 0.847)</td>
<td>0.861 (0.850, 0.873)</td>
<td></td>
</tr>
<tr>
<td>Proportion white (SD = 0.067)</td>
<td>1.017 (1.005, 1.028)</td>
<td>1.014 (0.998, 1.031)</td>
<td></td>
</tr>
</tbody>
</table>

for brevity. The Watanabe-Akaike Information Criterion (WAIC) for the Kernel smoothing (3) and locally adaptive (4) weight models are 92,839 and 74,363 respectively, suggesting a substantially improved fit of the locally adaptive weight model.

The impacts of each covariate on respiratory prescription rates are summarised in Table 1, which are presented as relative rates for a 1 standard deviation increase in each covariate. The main result is that fine particulate air pollution PM$_{2.5}$ has a substantial impact on prescription rates, with a 1.23µgm$^{-3}$ increase in PM$_{2.5}$ being associated with a 2.2% increase in prescription rates. The effects for PM$_{10}$ are similar but slightly smaller.

5 Discussion

Our main finding is that fine particulate air pollution impacts respiratory prescribing rates in primary care, suggesting that this form of air pollution has a detrimental impact on less severe ill health outside a hospital setting. In future work we will extend this work to focus on different medication types to see if these results are present more generally.

References


Bayesian Effect Fusion for Categorical Predictors in Logistic Regression

Magdalena Leitner¹, Helga Wagner¹

¹ Department of Applied Statistics, Johannes Kepler University Linz, Austria

E-mail for correspondence: magdalena.leitner@jku.at

Abstract: We extend two methods developed for Bayesian effect fusion in linear regression to logistic regression models. We investigate and compare the performance of both approaches for simulated data and employ them in the analysis of a data set on a marketing campaign.

Keywords: Categorical Covariate; Spike and Slab Prior; Sparse Finite Mixture Prior; MCMC

1 Introduction

Variables measured on an ordinal or nominal scale are widely used as predictors in logistic regression models in many application fields, e.g. in medicine or economics. As the effect of a categorical covariate is captured by a group of level effects this can easily lead to a high-dimensional vector of regression coefficients that have to be estimated. To achieve a sparse representation of the effects of categorical covariates Pauger and Wagner (2018) and Malsiner-Walli et al. (2018) suggested Bayesian methods that allow to shrink non-relevant level effects to zero and fuse levels with essentially the same effect on the response. Here, we extend these approaches to logistic regression, compare their performance in a simulation study and illustrate their application on real data.

2 Bayesian logit model

To model the effect of \( h = 1, \ldots, p \) categorical predictors with \( c_h + 1 \) ordered or unordered levels \( k = 0, \ldots, c_h \) on a binary response variable \( y \) we consider
the logistic regression model

\[ P(y = 1) = \frac{\exp(\zeta)}{1 + \exp(\zeta)}, \]

with the linear predictor given as

\[ \zeta = \mu + \sum_{h=1}^{p} \sum_{k=1}^{c_h} X_{h,k} \beta_{h,k}. \]

Here, \( \mu \) is the intercept, \( X_{h,k} \) denotes the dummy variable for level \( k \) of covariate \( h \) and \( \beta_{h,k} \) its effect with respect to the reference category \( k = 0 \).

To complete Bayesian model specification we specify a prior of the structure

\[ p(\beta, \mu) = p(\mu) \prod_{h=1}^{H} p(\beta_{h} | \xi_h), \]

where \( \xi_h \) denotes additional covariate-specific hyperparameters. A flat Normal prior \( N(0, M_0) \) is assigned to the intercept \( \mu \). For the regression effects \( \beta_h \) associated to one categorical covariate we consider two different priors encouraging effect fusion: the prior proposed in Pauger and Wagner (2018) is a multivariate Normal distribution specified hierarchically as

\[ \beta_{h} \sim N(0, \gamma_h \tau_h^2 Q_{-1}(\delta_h)), \]

\[ \tau_h^2 \sim G^{-1}(g_{h0}, G_{h0}), \]

\[ p(\delta_h) \propto |Q_h(\delta_h)|^{1/2} \sum (1 - \delta_{h,kj})/2, \]

where \( \gamma_h \) is a scale factor and the elements of the prior precision matrix \( Q_h \) depend on the vector \( \delta_h \) of binary indicators \( \delta_{h,kj} \) which are defined for all pairs of levels \( 0 \leq j < k \leq c_h \). Depending on the value of \( \delta_{h,kj} \) the prior partial correlation of the effects of levels \( k \) and \( j \) can be low (for \( \delta_{h,kj} = 1 \)), encouraging two distinct level effects, or high (for \( \delta_{h,kj} = 0 \)), encouraging fusion of the two level effects. As shown in Pauger and Wagner (2018) this prior can be derived alternatively by assigning spike and slab priors to all effect differences \( \beta_{h,k} - \beta_{h,j} \) and correcting for the linear restrictions among them.

The second prior distribution, proposed by Malsiner-Walli et al. (2018), encourages effect fusion using a Bayesian model based clustering approach. In this approach a finite mixture prior of spiky components is specified on the level effects \( \beta_{h,k}, k = 1, \ldots, c_h \) as

\[ p(\beta_{h,k}) = \sum_{l=0}^{L_h} \eta_{hl} f_N(\beta_{h,k} | \mu_{hl}, \psi_{hl}), \]

\[ \eta_h \sim Dir_{L_h+1}(e_0). \]
Here, $L_h + 1$ is the number of mixture components for covariate $h$ with location parameters $\mu_{hl}$ and scale parameter $\psi_h$. $\mu_{h0}$ is set to 0 to encourage fusion with the baseline and a Normal hyperprior $N(m_{h0}, M_{h0})$ is assigned to $\mu_{hl}$ for $l = 1, \ldots, L_h$. All mixture components of one predictor have the same variance $\psi_h$, but the variances might be different across predictors. Empty mixture components which result when several level effects are assigned to the same mixture component are encouraged by choosing a small value for the parameter $e_0$ of the Dirichlet distribution, see Malsiner-Walli et al. (2018) for details.

For ordinal predictors the ordering information on the levels suggests to restrict fusion to adjacent pairs. Such a restriction can be easily incorporated in the spike and slab effect fusion prior but taking into account the ordering information is not feasible in the sparse finite mixture prior.

### 3 Posterior inference

For the logistic regression model posterior inference can be accomplished by MCMC sampling using data augmentation with latent Pólya-Gamma variables suggested in Polson et al. (2013). Thus, the logistic regression model can be represented as a linear regression model with errors following a scale mixture of Normals and the sampling schemes of Pauger and Wagner (2018) and Malsiner-Walli et al. (2018) can be used for posterior inference with one additional step to sample the Pólya-Gamma latent variables. We select the final model as the model minimizing the expected posterior Binder loss for each covariate separately for the spike and slab prior and for the finite mixture prior by identifying the optimal partition of the effects using PAM clustering and the silhouette coefficient.

### 4 Simulation study

To compare the performance of the two methods we conducted a simulation study. We simulated 100 data sets with $n = 2000$ observations with 4 ordinal and 4 nominal covariates. For each type of covariate two predictors have 8 and the other two have 4 levels and the covariates are generated independently with level probabilities $\pi_h = (0.1, 0.1, 0.2, 0.05, 0.2, 0.1, 0.2, 0.05)$ and $\pi_h = (0.1, 0.4, 0.2, 0.3)$ for regressors with 8 and 4 levels, respectively. We set $\mu = -1$ and the regression coefficients to $\beta_1 = (0, 1, 1, 2, 2, 4, 4)$ and $\beta_3 = (0, -2, -2)$ for the ordinal and $\beta_5 = (0, 1, 1, 1, -2, -2)$ and $\beta_7 = (0, 2, 2)$ for the nominal predictors and $\beta_h = 0$ for $h = 2, 4, 6, 8$.

For the spike and slab method we choose $r = 5 \cdot 10^8$, $g_{h0} = 5$, $G_{h0} = 25$ and $\tau^2 = 10^9$ and we set $e_0 = 0.01$ and $\nu = 10^4$ for the model based clustering method. For comparison we consider also the full model without
effect fusion and the true model, where the correct fusion is assumed to be known. These two models were fit using independent flat Normal priors on the level effects. MCMC was run for 8000 iterations after a burnin of 2000 for both effect fusion methods and for 3000 iterations after a burnin of 1000 iterations for the full and the true model. Figure 1 shows boxplots of the mean squared estimation error, which is defined for covariate $h$ in data set $i$ as

$$MSE_h^{(i)} = \frac{1}{c_h} \sum_{k=1}^{c_h} (\hat{\beta}_{h,k}^{(i)} - \beta_{h,k})^2,$$

for the covariates with 8 levels.

![Boxplots for different covariates showing MSE for each method.](image)

FIGURE 1. MSE for ordinal covariates 1 and 2 and nominal covariates 5 and 6.

The median MSE is smaller than in the full model for the effects of all 4 covariates under both priors. The finite mixture prior performs better than the spike and slab prior for the ordinal covariate 1 and the nominal covariate 5. For the ordinal covariate 2 and the nominal covariate 6, it is clearly outperformed by the spike and slab prior which performs almost as well as the true model. None of the levels of these two covariates has a non-zero effect and hence all levels could be fused to the baseline, however this model cannot be selected for the finite mixture prior as the silhouette coefficient does not allow for one cluster solutions.

5 Success of bank telemarketing

We illustrate the proposed methods for data on a marketing campaign of a Portuguese retail bank from May 2008 to November 2010 (downloaded from [http://archive.ics.uci.edu/ml/datasets/bank+marketing](http://archive.ics.uci.edu/ml/datasets/bank+marketing)), see Moro et al. (2014) for a detailed description. The binary response variable indicates whether a customer subscribes to a term deposit or not. Table 1 describes the covariates used in the analysis. We restrict the analysis to
observations without missing values and exclude observations where covariate education has the level illiterate leaving 30477 observations for the analysis.

TABLE 1. Description of the covariates. The first category is the baseline.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Description / Variable values</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>metric</td>
<td>age of the client in years</td>
</tr>
<tr>
<td>campaign</td>
<td>metric</td>
<td>number of contacts performed during this campaign for this client</td>
</tr>
<tr>
<td>previous</td>
<td>metric</td>
<td>number of contacts performed before this campaign for this client</td>
</tr>
<tr>
<td>education</td>
<td>ordinal</td>
<td>educational level of the client (basic 4 years, basic 6 years, basic 9 years, professional course, high school, university)</td>
</tr>
<tr>
<td>poutcome</td>
<td>nominal</td>
<td>outcome of the previous marketing campaign for this client (non-existent, failure, success)</td>
</tr>
<tr>
<td>job</td>
<td>nominal</td>
<td>type of the client’s job (admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed)</td>
</tr>
<tr>
<td>marital</td>
<td>nominal</td>
<td>marital status of the client (married, divorced, single)</td>
</tr>
<tr>
<td>housing</td>
<td>nominal</td>
<td>indicates whether the client has a house loan contract</td>
</tr>
<tr>
<td>loan</td>
<td>nominal</td>
<td>indicates whether the client has a personal loan contract</td>
</tr>
<tr>
<td>day of week</td>
<td>nominal</td>
<td>day of the week (Monday, Tuesday, ...) on which the client was last contacted</td>
</tr>
<tr>
<td>month</td>
<td>nominal</td>
<td>month in which the client was last contacted (March, April,...; note: there are no observations for January and February)</td>
</tr>
<tr>
<td>contact</td>
<td>nominal</td>
<td>contact communication type (cellular, telephone)</td>
</tr>
</tbody>
</table>

We use the same priors as in the simulation study and run MCMC for 15000 iterations after a burnin of 15000. Figure 2 shows that both approaches give very similar results. The number of estimated regression effects is reduced from 39 to 13 under the spike and slab prior and to 15 under the finite mixture. For both methods the BIC of the selected model is considerably smaller than for the full and the intercept only model, see Table 2.

TABLE 2. Comparison of in-sample performance.

<table>
<thead>
<tr>
<th></th>
<th>Intercept only model</th>
<th>Full model</th>
<th>Spike and slab</th>
<th>Finite mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIC</td>
<td>23155.81</td>
<td>19699.59</td>
<td>19470.58</td>
<td>19500.15</td>
</tr>
</tbody>
</table>

6 Conclusion

Both priors proposed for Bayesian effect fusion encourage sparsity and automatic grouping of categorical covariate levels in logistic regression models. MCMC sampling is feasible using data augmentation and according to the
simulation study and the real data example both approaches perform well and yield similar results.

Acknowledgments: This work was financially supported by the Austrian Science Fund (FWF) via the research project number P25850 'Sparse Bayesian modelling for categorical predictors'.

References


Constrained log-likelihood for partial proportional odds models for ordinal response variables

Altea Lorenzo-Arribas¹², Mark J. Brewer ¹, Antony M. Overstall²

¹ Biomathematics and Statistics Scotland, UK
² University of Southampton, UK

E-mail for correspondence: altea.lorenzo-arribas@bioss.ac.uk

Abstract: Partial Proportional Odds Models where the Proportional Odds assumption can be relaxed on one or more of the covariates can be problematic particularly when one of the covariates is continuous. In addition to a potential for over-parameterisation, the main issues are lack of convergence and the presence of negative predicted class probabilities. We provide a simulation assessment of the frequency in which these issues appear and propose solutions by means of a Lasso penalisation and a re-parameterisation of the log-likelihood for the model. A case study looking at environmental attitudes has also been studied.

Keywords: PPOM; ordinal; Lasso; environmental attitudes.

1 PPOMs limitations

Partial Proportional Odds Models (PPOMs) have been proposed as a generalisation of Proportional Odds Models (POMs) where the Proportional Odds (PO) assumption can be relaxed for one or more of the covariates, corresponding to relaxing the assumption of parallel regression lines (Agresti, 2010). In the framework of the cumulative logit models, given an ordinal response variable \( Y_i \) with \( C \) ordered categories, PPOMs can be defined by the following expression:

\[
\logit(P(Y_i \leq j)) = \alpha_j + \sum_{k=1}^{p} \beta_k X_{ik} + \sum_{k=1}^{q} \gamma_{jk} Z_{ik}
\]
where: $i = 1, \ldots, n; j = 1, \ldots, C - 1; -\infty < \alpha_1 < \alpha_2 < \ldots < \alpha_{C-1} < \infty$ are the thresholds or cut-points (Peterson & Harrell, 1990), and we define $\gamma_{jk} = \gamma_k + u_{jk}$. $Z_{ik}$ are the covariates for which some or all coefficients vary by category (here $\gamma_{jk}$ from $k = 1$ to $k = q$, where for $j = 1$, $\gamma_{1k} = 0$ for all $k$) while the coefficients for the covariates $X_{ik}$ remain constant for all response categories (here $\beta_k$ from $k = 1$ to $k = p$). $q$ represents the number of covariates for which the PO assumption does not hold and $p$ is the number of covariates for which the PO assumption holds ($q \leq p$). For $q = 0$, we would have the expression for POM.

PPOMs are highly parameterised and this can cause problems. One problem is with convergence, and the full PPOM may not be even be identifiable. Another problem is just as serious; the PPOM can produce negative predicted class probabilities, since the non-parallel lines can intersect within the range of the continuous covariate (Hedeker et al., 2006; see Figure 1).

Although Williams (2016) reports that this issue is rare, we have found that it can be in fact more common than expected, particularly in cases of separation and quasi-complete separation. Unlike in ordinary least-squares regression for modelling a normally distributed response, when a logistic model perfectly or nearly perfectly predicts the response (that is, separates the response levels), unique maximum likelihood estimates do not exist. Some model parameters are not identifiable. This is a common result of the data being sparse, meaning that not all response levels are observed in each of the predictor settings, which often happens with small data sets or when the event is rare or a response option is unlikely to be chosen (e.g., due to social desirability bias).
2 Solutions

We propose solutions to these two problems. Firstly, for convergence problems caused by over-parameterisation, we use a Lasso estimation method that will also determine which covariates satisfy the PO assumption:

\[
L^* = L + \sum_{k=1}^{q} \lambda_k \sum_{j=1}^{C} |u_{jk}|,
\]

where \( L \) is the original log-likelihood:

\[
L = \sum_{i=1}^{n} \sum_{j=1}^{C} I_{ij} \log(P(Y = j|X_i)) = \sum_{i=1}^{n} \sum_{j=1}^{C} I_{ij} \log(\pi_{ij})
\]

with \( i = 1, \ldots, n; j = 1, \ldots, C, \) and where \( I_{ij} \) is an indicator variable for observation \( i \) such that:

\[
I_{ij} = \begin{cases} 
1 & \text{if } Y_i = j \\
0 & \text{if } Y_i \neq j
\end{cases}
\]

and the probabilities \( \pi_{ij} \) are defined as follows:

\[
\pi_{ij} = P(Y_i = j) = \begin{cases} 
P(Y_i \leq 1) & \text{if } Y_i = 1 \\
(P(Y_i \leq j) - P(Y_i \leq j - 1) & \text{if } 1 < Y_i < C - 1 \\
1 - P(Y_i \leq C - 1) & \text{if } Y_i = C
\end{cases}
\]

where \( P(Y_i \leq j) \) is the cumulative probability that a given observation is less than the \( j \)-th level (Peterson & Harrell, 1990).

Secondly, for the problem of negative class probabilities, we propose a geometric reformulation of the model which guarantees that class probabilities will be non-negative. The proposed parameterisation for the log-likelihood is derived from two straight lines for which we impose a restriction so that they do not overlap within the stated limits. In order to find the parameter values, we re-express the original definition of PPOMs where \( p = 0 \) and \( q = 1 \) as:

\[
\text{logit}(P(Y_i \leq j)) = \alpha_j + \gamma_j Z_i,
\]

where \( i = 1, \ldots, n \) and from which it follows that:

\[
(\alpha_1 - \alpha_2) + \gamma_1 z_i \leq \gamma_2 z_i
\]

We could have the following parameterisation for a 2 categories response variable:

\[
\alpha_1 = \alpha_1^*, \alpha_2 = \alpha_1 + \alpha_2^*
\]
where:

$$\gamma_1 = \gamma_1^* \cdot \gamma_2 = \gamma_1 + (\alpha_1 - \alpha_2) + \gamma_2^*$$

with $\gamma_1^*, \gamma_2^* \geq 0$. 
Or we could define more conveniently: $a_2 = \log(\alpha_2^*)$ and $g_2 = \log(\gamma_2^*)$ and apply this new parameterisation to the corresponding log-likelihood. This parameterisation could be extended to models with more covariates (see Figure 2 for the case of two covariates).

![Diagram](image)

**FIGURE 2.** Re-parameterisation for a PPOM with two covariates $z_a$ and $z_b$.

In order to avoid the crossing of the planes, we would have the following conditions for the four corners:

$$\begin{align*}
\alpha_2 &> \alpha_1 \\
\alpha_2 + \gamma_2^a &> \alpha_1 + \gamma_1^a \\
\alpha_2 + \gamma_2^a + \gamma_2^b &> \alpha_1 + \gamma_1^a + \gamma_1^b \\
\alpha_2 + \gamma_2^b &> \alpha_1 + \gamma_1^b
\end{align*}$$

In summary,

$$\alpha_0 + z^t \gamma_0' \leq \cdots \leq \alpha_{C-1} + z^t \gamma_{C-1}$$

In our simulation study we consider 6 scenarios; 3 where we generate data from a POM and 3 from a PPOM, and we compare the results for the default log-likelihood, the new parameterisation and the Lasso penalisation.
3 Case study

In addition to the simulation study, we apply our new parameterisation to data from the Scottish Environmental Attitudes and Behaviours Survey (Ipsos MORI, 2008) which evaluates respondents’ environmental attitudes. We have assessed different ordinal models for which we find crossing of regression lines on the probability scale (see Figure 3). We start with a simple example where we model educational attainment (i.e., highest level of qualification obtained) versus age via a PPOM. Although we acknowledge that for an appropriate analysis of the data, we would need to control for other covariates (e.g., sex), for the purposes of this methodological study, we start with one covariate only and follow by exploring more complex models.

![Figure 3. PPOM of educational attainment as a function of age without (a) and with (b) Lasso penalisation (λ = 0.9).](image)

We can see that even for a restrictive Lasso (see Figure 3 (b)), there are still crossing issues (e.g., the yellow category is overlapping the orange category). For this particular data set and the specific models we have looked at, we have found our re-parameterisation to be a more systematic and efficient technique than the Lasso penalty, with regards to avoiding the crossing of ordinal regression lines in the probability scale.

Acknowledgments: This work was supported by the Scottish Government’s Rural and Environment Science and Analytical Services Division (RESAS).

References


Penalised maximum likelihood estimation in multistate models with an application to cardiovasculary disease

Robson J. M. Machado¹, Ardo van den Hout¹, Giampiero Marra¹

¹ University College London, U.K.

E-mail for correspondence: robson.machado.14@ucl.ac.uk

Abstract: We propose an efficient method to estimate nonparametric multistate models with splines for interval-censored data. A Markov process framework is used to formulate the models. Hazards are specified with splines base functions to allow for flexible modelling over time. Estimation is undertaken using a penalised likelihood approach. Given a piecewise-constant approximation to the hazards, a Fisher scoring algorithm can be used for estimation of model parameters. An automatic method is used to estimate the multiple smoothing parameters. The new estimation procedure is made possible by rewriting the optimisation problem in a generalised likelihood based penalised method. The method is illustrated with data for the progression of cardiac allograft vasculopathy in posttransplantation patients.

Keywords: Multistate models; Interval censoring; Splines.

1 Introduction

The method is applied to data for cardiac allograft vasculopathy (CAV). CAV is a narrowing of the arterial walls and the main cause of death in heart transplantation patients. The data are a series of approximately yearly angiographic examinations of heart transplant recipients. The state at each time is a grade of CAV which can be normal, moderate or severe. Dead is the absorbing state and time of death is known within one day. The process is biologically irreversible and of particular interest is the onset of CAV. Diagnosis of ischaemic heart disease (IHD) and donor age are known to be major risk factors of disease onset (Titman, 2011). In order to investigate this, three-state progressive models can be defined. The states
are classified as normal (1) if the patient has not developed the disease, ill (2) if the patient has developed moderate or severe CAV and dead (3) if the patient has died, see Figure 1.

2 Multistate models with splines

Let $Y(t)$ be a continuous-time Markov chain on finite state space $S$, time-homogeneous transition probabilities are given by

$$p_{rs}(t) = P(Y(t + u) = s | Y(u) = r),$$

for $r, s \in S$, $u \geq 0$ and $t \geq 0$. Transition matrix $P(t)$ contains these probabilities such that the rows sum up to 1. The hazards are defined by

$$q_{rs} = \lim_{\Delta t \to 0} \frac{P(Y(t + \Delta t) = s | Y(t) = r)}{\Delta t},$$

for $r \neq s$. The matrix with off-diagonal entries $q_{rs}$ and diagonal entries $q_{rr} = -\sum_{r \neq s} q_{rs}$ is the generator matrix $Q$. Given $Q$, the solution for $P(t)$ subject to $P(0) = I$ is $P(t) = \exp(tQ)$, see, e.g., Kalbfleish and Lawless (1985). Time-dependent models can be defined by using proportional hazards model for transition $r$ to $s$, $r \neq s$

$$q_{rs}(t) = q_{rs,0}(t) \exp \left( \beta_{rs}^\top x \right),$$

where $q_{rs,0}(t)$ is the baseline hazard function, $x$ is a covariate vector and $\beta_{rs}^\top$ is vector of unknown parameters. We focus on the nonparametric estimation of $q_{rs,0}(t)$ with splines. Each hazard can be approximated by the exponential of a linear combination of $K_{rs}$ spline base functions $B_k(t)$ and regression coefficients $\alpha_{rs,k} \in \mathbb{R}$ as follows

$$q_{rs,0}(t) = \exp \left( \sum_{k=1}^{K_{rs}} \alpha_{rs,k} B_k(t) \right).$$

3 Penalised maximum likelihood estimation

For each hazard, let the number of splines basis dimension be large enough to allow for flexible modelling. Define the full set of parameter by $\theta$ and the
penalty matrix by $S_\lambda$. This is a block diagonal matrix with blocks $\lambda_{rs}S_{rs}$ for penalising splines parameters of transition $r$ to $s$ and zeros elsewhere. The amount of smoothing is controlled by adding a smoothness penalty to the log-likelihood function. Let $\ell(\theta)$ be the usual logarithm of the likelihood function for multistate models with interval-censored data (Jackson, 2011). The penalised log-likelihood function is

$$\ell_p(\theta) = \ell(\theta) - \frac{1}{2} \theta^\top S_\lambda \theta. \quad (3)$$

For the estimation, let $g_p^{[a]} = g^{[a]} - S_\lambda \theta^{[a]}$ and $I_p^{[a]} = I^{[a]} + S_\lambda$ represent the penalised gradient and negative of the penalised hessian matrix at iteration $a$, respectively, where $g^{[a]} = \partial \ell(\theta)/\partial \theta|_{\theta=\theta^{[a]}}$ and $I^{[a]} = -\partial^2 \ell(\theta)/\partial \theta \partial \theta^\top|_{\theta=\theta^{[a]}}$. For fixed value of $\hat{\lambda}$, the $a^{th}$ estimate of $\theta$ can be updated by

$$\theta^{[a+1]} = (I^{[a]} + S_\lambda)\^{-1} \sqrt{I^{[a]}} z^{[a]}, \quad (4)$$

where $z^{[a]} = \sqrt{I^{[a]}} \theta^{[a]} + \epsilon^{[a]}$ and $\epsilon^{[a]} = \sqrt{I^{[a]}}^{-1} g^{[a]}$ (Marra et al., 2017).

From likelihood theory, $\epsilon \sim N(0, I)$ and $z \sim N(\mu_z, I)$, where $I$ is the identity matrix, $\mu_z = \sqrt{I} \theta$ and $\theta$ is the true parameter vector. The predicted value vector for $z$ is $\hat{\mu}_z = \sqrt{I} \hat{\theta} = A_\lambda z$, where $A_\lambda = \sqrt{I}(I + S_\lambda)^{-1}\sqrt{I}$. The smoothing parameter vector, $\lambda$, is estimated by minimising the Un-Biased Risk Estimator

$$\mathcal{V}(\lambda) = ||z - A_\lambda z||^2 - c + 2tr(A_\lambda). \quad (5)$$

Equation (5) can be minimised using the automatic smoothing parameter selection method developed by Wood (2006) or by using a general-purpose optimiser.

### 4 Application to CAV data

We fit a progressive three-state model for the CAV data defined as in Figure 1. The proportional hazard model with splines is specified with dependence on donor age ($dage$) and primary diagnosis of ischaemic heart disease (IHD):

$$q_{rs}(t) = \exp \left( \sum_{k=1}^{10} \alpha_{rs,k} B_k(t) + \beta_1 dage + \beta_2 IHD \right), \quad (6)$$

where $(r, s) \in \{(1, 2), (1, 3), (2, 3)\}$ and $B_k(t)$ are known spline basis function. We use penalised cubic regression splines. The estimated smooth hazards for subjects with IHD and donor age of 26 (solid lines) and 95% confidence intervals (dashed lines) are presented.
FIGURE 2. Estimated smooth hazards for subjects with IHD and with donor age of 26 (solid lines), with 95% confidence intervals (dashed lines)

in Figure 2. The risk of moving from state 1 (healthy) to state 2 (CAV) increases until approximately 8 years after transplant, but decreases afterwards. The risk of going from state 1 to state 3 (dead) is very low and almost constant until approximately 10 years since transplant, but increases pretty steep afterwards. The transition intensity from state 2 to state 3 is quite volatile and upwards until 10 years after transplant and decreasing afterwards. The confidence intervals are fairly wide after approximately 10 years. That is because data become scarce after 10 years. For the parametric part of the model, $\widehat{\beta}_1 = 0.018$ and $\widehat{\beta}_2 = 0.274$ indicating that donor age and IHD increase the risks of disease progression and death. The vector of smoothing parameters is estimated at $\hat{\lambda} = (47.145, 41.668, 10.716)^T$.

Acknowledgments: Special Thanks to CNPq - Brazil [249308/2013-4]

References


Non-PH parametric survival modelling

Gilbert MacKenzie¹, Yasin Al-tawarah², Milisa Blagojevic³ and Kevin Burke¹

¹ University of Limerick, Ireland  
² Royal Jordanian Army Medical Corps, Jordan  
³ Keele University, UK.

E-mail for correspondence: gilbert.mackenzie@ul.ie

Abstract: Non-PH parametric survival modelling is developed within the framework of the multiple logistic function. The family considered comprises three basic models: (a) a PH model, (b) an accelerated life model and (c) a model which is non-proportional hazards and non-accelerated life. The family is generalised by means of a Gamma frailty extension which is shown to accommodate crossing hazards data. These extensions lead naturally to the concept of a Multi-Parameter Regression model described by Burke and MacKenzie (2016). The new models are used to analyse two sets of survival data and the advantages of the methods are discussed.

Keywords: Accelerated Life, Frailty, MPR model, Non-PH Model, PH Model

2 Introduction

MacKenzie (1996) introduced a family of survival models based on the multiple logistic function. The generalised time dependent logistic family (GTDL) comprised three models (a) a PH model, (b) an accelerated life (AL) model and (c) a model which is non-PH and non-AL. Development was focussed on model (c), designated the GTDL model, and a Gamma extension demonstrated its ability to deal with crossing hazards survival data (MacKenzie and Ha, 2007). This example involved the creation of a covariate dependent shape parameter. Accordingly, the idea of modelling the shape parameter more generally intrudes and this has motivated the development of multi-parameter regression survival models (MPR models) (Burke and MacKenzie, 2013, 2016). We trace these methodological developments and illustrate the main ideas using a non-PH Weibull MPR model.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
3 The GTDL family

The GTDL family was predicated on the multiple logistic function and the defining hazard functions of the three models in the family areas follows:

(a) The GTDL PH model

\[ \lambda(t; x) = \pi(t\alpha + \gamma) \exp(x'\beta) \]  

(b) The GTDL AL Model

\[ \lambda(t; x) = \lambda\phi\pi(\alpha\phi t) \]  

(c) The GTDL model

This model is not PH and not AL

\[ \lambda(t; x) = \lambda_0\pi(t\alpha + \gamma^*) \]  

In the models above \( \pi(s) = \exp(s)/[1 + \exp(s)] \), \( \lambda_0 > 0 \) and \( \lambda > 0 \) are scalars, \( \phi = \exp(x'\beta) \) and \( \gamma^* = x'\beta \). An intercept term is included in the linear predictor, \( \gamma^* \), for model (c), but not in the other two models.

4 Frailty Extensions

Standard arguments involving the multiplicative random effect, \( u_i \), on the hazard function yields, the general formulae for the marginal survivor and hazard functions

\[ S_m(t) = L[\Lambda(t)] = [1 + \phi \Lambda(t)]^{-1/\phi}, \]  

and

\[ \lambda_m(t) = \lambda(t) \frac{-L'[\Lambda(t)]}{L[\Lambda(t)]} = \frac{\lambda(t)}{1 + \phi \Lambda(t)}, \]  

respectively. Here \( S(t), \Lambda(t) \) and \( \lambda(t) \) are the basic survival quantities. Moreover, \( L[\Lambda(t)] \) is the Laplace transform and \( U \) is the random effect with density \( g(u) = [\phi^{1/\phi}\Gamma(1/\phi)]^{-1}u^{1-1}\exp(-u/\phi) \) with \( E(U) = 1 \) and \( V(U) = \phi \). In the main paper we use these formulae to generalise the three models in the family and analyse the lung cancer data.

5 MPR modelling

MPR survival models model the scale and shape parameters simultaneously as a function of covariates. We develop the Non-PH MPR Weibull model
TABLE 1. Models fitted and their marginal mles & (s.e.)

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\alpha}_0$</th>
<th>$\hat{\alpha}_1$</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\sigma}^2$</th>
<th>$\hat{\ell}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.106</td>
<td>-</td>
<td>-307.47</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>(0.223)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cox GF</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-1.146</td>
<td>1.717</td>
<td>-306.50</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>(0.675)</td>
<td>(1.024)</td>
<td>-</td>
</tr>
<tr>
<td>TDL</td>
<td>-0.832</td>
<td>-0.094</td>
<td>1.494</td>
<td>-1.380</td>
<td>-</td>
<td>-132.55</td>
</tr>
<tr>
<td></td>
<td>(0.242)</td>
<td>(0.192)</td>
<td>(0.666)</td>
<td>(0.822)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GTDL GF</td>
<td>-0.789</td>
<td>3.499</td>
<td>2.380</td>
<td>-4.612</td>
<td>0.400</td>
<td>-127.89</td>
</tr>
<tr>
<td></td>
<td>(0.326)</td>
<td>(1.408)</td>
<td>(1.413)</td>
<td>(1.676)</td>
<td>(0.176)</td>
<td>-</td>
</tr>
</tbody>
</table>

which is used (below) to fit the lung cancer data. The hazard function is given by

$$\lambda(t; x, z) = \exp(x'\beta) \exp(z'\alpha) t^{\exp(z'\alpha) - 1}$$  \hspace{1cm} (6)

where the Weibull hazard is $\lambda(t) = \lambda \gamma t^{\gamma - 1}$ ($\lambda > 0$, $\gamma > 0$) and the MPR specification is

$$\log \lambda = x'\beta \hspace{1cm} \text{(scale)} \hspace{1cm} \log \gamma = z'\alpha \hspace{1cm} \text{(shape)}$$

note that $x$ and $z$ may contain the same covariates.

The idea reprises related work on joint modelling in other arenas, for example, the developments in joint mean-covariance modelling in longitudinal prospective studies (see Pan & MacKenzie, 2003, 2006, and 2007 and by Xu & MacKenzie, 2012). This approach offers a relatively simple method of modelling changes in shape and has the added benefit of treating the two parameters symmetrically from the inferential perspective.

### 6 Applications

A classical motivating example is the crossing hazards data provided by the Gastrointestinal Tumor Study Group (GTSG)(1982), reporting the effects of chemotherapy and combined chemotherapy and radiotherapy on the survival times of gastric cancer patients.

The results of fitting several models is shown in Table 1. The most successful model is the GTDL Gamma frailty model involving separate shape parameters for the two groups. The other models (PH and non-PH) shown in the table are not successful. This shows that having covariate dependent shape parameters is sometimes useful.

We turn now to analyse the Lung Cancer data set. This was a multi-source population study of 855 incident cases in Northern Ireland diagnosed between Oct. 1st 1991 and Sept. 30th 1992. The patients were followed for
c18 months and their survival time was computed as the time from diagnosis to death or censoring. Some 693 (77%) patients had died by the censoring date (30th May, 1993). The influence of 9 covariates were analysed: Age, Sex, Treatment, WHO Status, Cell type, Sodium level, Albumen level, Metastases and Smoking category.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>PH Weibull</th>
<th>MPR Weibull</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>β</td>
<td>β, α</td>
</tr>
<tr>
<td>Age group</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WHO Status</td>
<td>β</td>
<td>β</td>
</tr>
<tr>
<td>Sex</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Smoker</td>
<td>β</td>
<td>α</td>
</tr>
<tr>
<td>Cell Type</td>
<td>β</td>
<td>β</td>
</tr>
<tr>
<td>Metastases</td>
<td>β</td>
<td>β, α</td>
</tr>
<tr>
<td>Sodium</td>
<td>β</td>
<td>β</td>
</tr>
<tr>
<td>Albumen</td>
<td>β</td>
<td>β, α</td>
</tr>
</tbody>
</table>

AIC 3723.1 3679.7  
Δ-AIC 43.4 0.0  
dim(θ) 22 30

The results are shown in Table 2. The presence of a $\beta$ indicates a statistically significant covariate in the scale parameter while the presence of an $\alpha$ indicates that the covariate is statistically significant in the shape parameter. In the PH model the shape parameter is a constant, $\gamma$. The presence of an $\alpha$ also indicates that the covariate is formally non-PH. From the AIC information, the superiority of the non-PH MPR model is apparent, even though it fits more parameters.

7 Discussion

The GTDL family is of course not the only way to model non-PH data but it was the use of a member of this family in the crossing hazards example that led, in part, to the broader idea of modelling the shape parameter symmetrically with the usual scale parameter in survival distributions. This highlights both the importance of the frailty concept and the MPR approach. In further work we have demonstrated that, unlike the lung cancer data analysed here, both concepts can be required together which suggests that the MPR model captures a form of time dependence which classical frailty models cannot.
References


Variational message passing for skew t regression

Luca Maestrini¹, Matt P. Wand²

¹ Department of Statistical Sciences, University of Padova, Italy
² School of Mathematical and Physical Sciences, University of Technology Sydney, Australia

E-mail for correspondence: luca.maestrini@phd.unipd.it

Abstract: We extend the class of variational message passing algorithms to approximate fitting and inference for skew t regression models, building on recent work concerning variational message passing on factor graph. A major advantage of a factor graph fragment approach is that calculations only need to be done once for the considered distribution family and can be easily adapted to accommodate more complex model structures. A simulation study shows how posterior dependence arising in auxiliary variable representation of a skew t response model may lead to poor performances in terms of variational message passing when using convenient factorizations of the approximating densities.

Keywords: Auxiliary variables; Factor graph fragment; Skew t; Variational approximation; Variational message passing.

1 Introduction

The notion of factor graph fragment introduced in Wand (2017) seminal paper represents a step forward towards the integration of variational approximations in mainstream statistics. This perspective provides a valid instrument to streamline algebra and computer coding when implementing variational message passing (VMP) for elaborate response regression models. A notable practical advantage is that algorithm updates only need to be derived once for a particular response model, which can be integrated in more complex structures such as those including, for instance, semiparametric components.

As shown in McLean and Wand (2018), modularity of variational message passing extends beyond the common distributions in the Bernoulli, Poisson...
and Normal families. We widen this recent body of work to include skew t regression models. Attention is reserved to an auxiliary variable representation of the model. Auxiliary variables can reduce the algebraic complexity of algorithm derivations but at the same time affect variational message passing performances, according to the assumptions on the approximating density.

2 Variational message passing

Consider a Bayesian statistical model with observed data \( D \) and parameter vector \( \theta \). Variational approximations are conceived to treat models in which the posterior density function \( p(\theta|D) \) is analytically intractable. A mean field variational approximation \( q^*(\theta) \) to \( p(\theta|D) \) is the minimizer of the Kullback-Leibler divergence

\[
\int q(\theta) \log \left\{ \frac{q(\theta)}{p(\theta|D)} \right\} d\theta
\]

subject to a product density restriction \( q(\theta) = \prod_{i=1}^{M} q(\theta_i) \), where \( \{\theta_1, \ldots, \theta_M\} \) is some partition of \( \theta \). It can be shown that the optimal \( q \)-density functions satisfy

\[
q^*(\theta_i) \propto E_{q(\theta\setminus\theta_i)} \{ p(\theta_i|D, \theta\setminus\theta_i) \}, \quad 1 \leq i \leq M,
\]

(1)

where \( \theta\setminus\theta_i \) denotes the entries of \( \theta \) with \( \theta_i \) omitted. The previous expression gives rise to an iterative scheme for obtaining the parameters of the optimal density functions \( q^*(\theta_i) \) which is known as mean field variational Bayes. VMP arrives at the same approximation via message passing on an appropriate factor graph. Among the several variants of VMP in the literature we follow the approach of Minka (2005), as summarized in Section 2.5 of Wand (2017), to derive fragment updates that allow for skew t response models to be handled within the VMP framework.

3 The skew t likelihood fragment

The skew t (Azzalini and Capitanio, 2003) likelihood fragment corresponds to the likelihood specification

\[
y_i|\theta, \sigma^2, \lambda, \nu \sim \text{Skew-t} \left\{ (A\theta)_i, \sigma^2, \lambda, \nu \right\}, \quad 1 \leq i \leq n,
\]

(2)

with \( A \) generic design matrix, \( \theta \) generic vector of coefficients and \( \nu > 0 \). If we introduce two auxiliary random variables \( a_{1i} \) and \( a_{2i} \), \( 1 \leq i \leq n \), model (2) can be expressed as

\[
y_i|\theta, \sigma^2, \lambda, a_{1i}, a_{2i} \sim N \left( (A\theta)_i + \frac{\sigma \lambda |a_{1i}| \sqrt{a_{2i}}}{\sqrt{1 + \lambda^2}}, \frac{a_{2i} \sigma^2}{1 + \lambda^2} \right),
\]

(3)

\[
a_{1i} \sim N(0, 1), \quad a_{2i} \sim \text{Inverse-\chi}^2(\nu, \nu).
\]
We then develop two VMP algorithms assuming that the optimal $q$-density admits the following alternative product restrictions:

A. \( q(\theta, \sigma^2, \lambda, \nu, a_1, a_2) = q(\theta) q(\sigma^2) q(\lambda) q(\nu) \prod_{i=1}^{n} q(a_{1i}) q(a_{2i}) \);

B. \( q(\theta, \sigma^2, \lambda, \nu, a_1, a_2) = q(\theta) q(\sigma^2) q(\lambda) q(\nu) \prod_{i=1}^{n} q(a_{1i}, a_{2i}) \).

The second partition avoids a more restrictive independence assumption on the auxiliary variables but requires an extensive use of non-standard conjugate exponential families and numerical integration when deriving a VMP algorithm. Figure 1 includes the factor graph representations of model (3) under both assumption A and B in support of VMP algorithm derivations. In detail, the shaded squares correspond to factors, which are single product components of model (3). The unshaded circles are called stochastic nodes and refer to parameters and variables expressing product dependencies in the approximating densities A and B. Edges connect factors to the stochastic nodes included in each factor.

![Factor graph](image)

**FIGURE 1.** Factor graph for Skew t likelihood specification in (3) under assumption A (left panel) and B (right panel).

4 Simulation study and application

We conduct a simulation study to assess the performances of VMP generating one hundred datasets of size \( n = 500 \) according to model (3), using a Uniform \((0, 1)\) predictor. We set the vector of location parameters \( \theta \equiv \beta = (\beta_0, \beta_1) \) to be \( \beta_0 = 1 \) and \( \beta_1 = 2 \). The scale parameter is \( \sigma = 1 \) and the shape parameters are \( \lambda = 5 \) and \( \nu = 1.5 \). The hyperparameters for \( \beta \) are fixed to \( \mu_\beta = 0 \) and \( \Sigma_\beta = 10^{10} I \) over a prior \( N(\mu_\beta, \Sigma_\beta) \). We use an Inverse-\( \chi^2(0.01, 0.01) \) prior on the squared scale. The prior for the parameter of symmetry \( \lambda \) is assumed to be \( N(0, 10^{10}) \) and that for the
degrees of freedom \( \nu \) to be \( \Gamma (1, 0.01) \). For each single parameter, we measure the accuracy of VMP approximations \( q^* (\theta_i) \) from (1) under both the assumptions A and B as

\[
\text{accuracy}(q^*) = 1 - \frac{1}{2} \int_{-\infty}^{\infty} |q^* (\theta_i) - p (\theta_i | \mathcal{D})| d\theta_i,
\]

so that \( 0 \leq \text{accuracy}(q^*) \leq 1 \). The \( L_1 \) error appearing in this last expression is a scale independent number that is invariant to monotone transformation on the parameter of interest. This implies, for instance, that the accuracy values for \( q^* (\sigma) \) and \( q^* (\sigma^2) \) coincide. Exact computation of \( p (\theta_i | \mathcal{D}) \) is replaced by MCMC samples obtained using \texttt{rstan} (Stan Development Team, 2018). MCMC samples of size 10000 were generated setting a burn-in of 5000 values and thinning the remaining 5000 by a factor of 5.

Table 1 summarizes mean and standard deviation of accuracy percentages from the simulation study, indicating a notable improvement when adopting the less restrictive assumption B.

**TABLE 1.** Average (standard deviation) accuracy from the simulation study.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Assump. A</th>
<th>Assump. B</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>0.0 (0.0)</td>
<td>37.7 (6.2)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>51.2 (17.7)</td>
<td>57.1 (4.5)</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>18.0 (16.9)</td>
<td>11.0 (3.5)</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>0.0 (0.0)</td>
<td>10.0 (3.4)</td>
</tr>
<tr>
<td>( \nu )</td>
<td>0.0 (0.0)</td>
<td>56.6 (6.1)</td>
</tr>
</tbody>
</table>

We also consider the dataset examined in Azzalini and Capitanio (2003) with the linear model

\[
y_i = \beta_0 + \beta_1 \text{CRSP}_i + \varepsilon_i, \quad \varepsilon_i \sim \text{Skew-t} \left(0, \sigma^2, \lambda, \nu\right), \quad 1 \leq i \leq n
\]

where the variables \( y_i \) and \( \text{CRSP}_i \) denote the excess rate of the Martin Marietta company and the index of return excess for the whole New York Stock Exchange respectively. Data over a period of \( n = 60 \) consecutive months from January 1982 to December 1986 are available. We estimate the posterior density functions via VMP approximation under assumption B and MCMC with the same settings of the simulation study.

The plots in Figure 2 show that VMP curves are roughly located around the modes of MCMC densities. Note also that the variance of VMP approximating densities appears lower than that of MCMC posterior densities, coherently with the theoretical results in Wang and Blei (2017). Better VMP performances could be achieved by proposing more generic assumptions on the approximating density function. However, our choice of the
product density restrictions is a compromise among algebraic complexity, feasibility and quality of the approximation.

FIGURE 2. Martin Marietta data: posterior density plots via MCMC and VMP.

Acknowledgments: The work of Luca Maestrini was carried out during a visiting period at the School of Mathematical and Physical Sciences, University of Technology Sydney, Australia. Special Thanks to Alessandra Salvan and Nicola Sartori for their assistance with this research.

References


208 VMP for skew t
Evaluating Statistical Information Retrieval Models with Different Indexing Enhancement Strategies

Kenan M. Matawie¹ and Sargon Hasso ²

¹ Western Sydney University, Australia
² Loyola University Chicago, USA

E-mail for correspondence: k.matawie@westernsydney.edu.au

Abstract: In this paper we will introduce the indexing enhancement for the Statistical Information Retrieval models (IR). To accomplish this task, we propose including an extra level of synonym and semantic classification and the distribution entropy of these synonym classifications via professional and technical resources that is determined by the query domain subject matter experts. We will evaluate this enhancement by using Rank-biased Precision (RBP) measurement after using BM25 IR model. This approach needs to re-formulate the implementations of IR models particularly the probabilistic models in an integrated system. Evaluated CRAN data will be analysed utilizing Lucene which offers a full-text search library by adding search functionality to our foundational model and system implementation.

Keywords: Indexing; Statistical Language model; Text Retrieval

1 Introduction

A full-text search application is an information retrieval (IR) application that allows users to submit ad hoc queries in an attempt to satisfy the information need of a user [Manning et al. 2008]. The search application consists of two main processes: indexing and searching. During indexing documents are analyzed and their content are stored in an inverted index [McCandless et al. 2010]. During searching process a user submits a query that consists of one or more terms and the search engine scans all the stored indexed documents and collects those documents that contain the same user’s query terms, sorts them by relevance, and returns the results to the user. Relevance is determined by the degree of similarity between the
user’s query and the matching retrieved documents containing the terms in the user’s query.

2 Statistical Scoring Functions

As we will explain in section 3, 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 functions based on some common theoretical information retrieval 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, i.e. best Match family (BM25) by Jones et al. (2000), since this is the scoring function we used for evaluation as implemented by Lucene:

\[
f(q, d) = \sum_{w \in q \cap d} c(w, q) \frac{(k + 1)c(w, d)}{c(w, d) + k(1 - b + b \frac{|d|}{avdl})} \log \frac{M + 1}{df(w)}
\]

where \(b \in [0, 1]\) is part of the normalizer term \(1 - b + b \frac{|d|}{avdl}\); \(avdl\) denotes average document length.

3 An Outline of Our Proposed Methodology

In our previous work, [Matawie, K. and Hasso, S.] (2016), we used a methodology to test various configurations of the ranking functions as implemented by Lucene with limited synonym expansion during indexing time. The data we used at that time were also limited in size. The results we obtained varied by using different language models and were improved by configuring Lucene index using WordNet as our only synonym expansion strategy. In this paper we will adopt a new methodology that focuses on:

- The use of BM25 language model only.
- New CRAN data set.
- Augment the index structure with synonym expansion using WordNet English-language lexical database.
- Use algorithmic-based stemmer, and dictionary- or language-based stemmer, i.e. lemmatizer, during analysis phase of the indexing process. Stemming reduces a word to its rood form. For example, an
algorithmic-based stemmer changes ‘administrator’, ‘administration’, and ‘administrate’ to the root ‘adminstr’ which is not a correct English root word; while, dictionary-based stemmer looks up the word in the dictionary and returns the correct root word, e.g. ‘walker’ and ‘walking’ are changed to ‘walk’.

- Use Rank-biased Precision (RBP) measurement to evaluate the efficiency and effectiveness of this new methodology.
- Build our testing platform on Elasticsearch (Elasticsearch BV.) which is based on Lucene search engine.

We built a system and evaluated the results from BM25 statistical ranking functions. Term-document matrix and probabilities for enhanced indexing models are constructed and integrated with corresponding ranking function. We also have implemented enhancements for example synonymic query expansion during indexing and searching. We noticed that running a query against an index that was generated with, say synonyms, gave us a different outcome from an index that was generated without synonyms. The same is true if we use different stemmers. We used the most popular evaluation criteria Rank-biased Precision (RBP) Moffat and Zobel (2008), a single metric, as an evaluation criterion to measure the quality across recall levels among all algorithms, i.e. relevant and non-relevant as judged by human experts. The Rank-Biased Precision (RBP) evaluation function is defined as:

$$ RBP_q = (1 - p) \sum_{i=1}^{N} r_i p^{i-1} $$

where $r_i$ is the $i$th relevance judgement, $i$ is the $i$th 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 (in this paper assumed to be 0.8). In our new methodology, we will evaluate, compare and graph the results for the well known CRAN Aviation data set.

4 Results and Performance Evaluation

Aviation document collection from the well known CRAN data containing 1400 documents and 225 queries with their associated relevance judgement were used. The BM25 relevance scoring function with six different enhancements (including default, which does not include synonym expansion nor stemming) were configured in Lucene. A combination of these configurations and field selections, i.e. search by title only, and search by title and text fields, generated eight different results, see Table 1.

The RBP values for the 225 queries for the highest (T7) and lowest (T5) enhancement with their averages are given in Figure 1. Different tests behaved differently and the maximum difference in the accuracy was about 25%.
<table>
<thead>
<tr>
<th></th>
<th>T2</th>
<th>T3</th>
<th>T5</th>
<th>T6*</th>
<th>T8</th>
<th>T7*</th>
<th>T9</th>
<th>T10*</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Possessive Stemmer</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Lower Case</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>English Stop</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Alogarithmic Stemmer</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Synonym</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Dictionary Stemmer</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

TABLE 1. Enhancement options (* title and text fields were included in search)

![Graph](image)

FIGURE 1. RBP values and averages for T7 and T5 enhancement.

Table 2 is the source data for Figure 1 showing the average of the RBP values for all the 225 queries across the eight tests we ran and compared against the judged aviation data. We also show how each test ranked among the different tests.

It’s worth noting that from each test we ran, we picked the top 20 highest scored documents, i.e. the most relevant document Lucene returned, and out of those we evaluated document relevance in relation to CRAN data set.
<table>
<thead>
<tr>
<th></th>
<th>T2</th>
<th>T3</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average RBP</td>
<td>0.31</td>
<td>0.33</td>
<td>0.12</td>
<td>0.18</td>
<td>0.34</td>
<td>0.32</td>
<td>0.12</td>
<td>0.18</td>
</tr>
<tr>
<td>Rank</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

TABLE 2. Average and Rank of RBP values for each test

5 Conclusion

In our previous work, we have seen adding synonym expansion to the index in the search application has given us improved relevance ranking. The best model was selected during our evaluation of the results using Mean Average Precision (MAP) criterion with the help of human subject matter experts. In this paper we introduced a new methodology to test, along with synonym expansion, two different stemming strategies and evaluated the results of these models using RBP measurement; this will allow us to effectively calibrate search relevance. The main focus of this paper was to show that different possible enhancements resulted in different model performance and this can lead and deserve further research and investigation to optimise and obtain better and more relevant returns for a given query especially for technical and professional data and resources. The data set we used here is restricted by size, domain and technical details but was sufficient enough to demonstrate the aim of the paper. A larger, more professional and specific data such as medical big data in a certain domain can be used and investigated with these different enhancement strategies.

References


Modelling of X-ray density (greyscale) signals from CT-scan images of rooted soil

Mario Martínez-Araya¹ and R. Murray Lark¹

¹ The University of Nottingham, School of Biosciences, Sutton Bonington Campus, Loughborough, LE12 5RD, United Kingdom

E-mail for correspondence: Mario.Martinez@nottingham.ac.uk

Abstract: In this work we analyse the X-ray density signal of computarized tomographic (CT) images of rooted soil and its relation to the location of roots in space. We used two representations of this signal: wavelets and cubic splines. Estimation of wavelet coefficients was carried out using shrinkage methods based on least asymmetric wavelet bases (symlet), while cubic spline estimation was performed via residual maximum likelihood (REML) estimation based on a mixed model representation of the cubic spline. We present analyses of CT-scan images of rooted soil from a controlled experiment with wheat, and discuss the advantages and disadvantages of cubic splines and wavelets for modelling this signal.

Keywords: Smoothing splines; Wavelets; Image analysis; Longitudinal Data; CT-scan.

1 Introduction and data

In recent years, X-ray Computed Tomography (CT) has proved valuable for studying roots in their natural undisturbed state within soil (Helliwell et al., 2012). The structural organisation of the soil is fundamental to how it functions, and hence understanding the spatio-temporal dynamics of how roots and associated soil interact is of paramount interest. We propose to study the relation between root and soil by defining the X-ray density (greyscale) signal in the CT scan as a process indexed by its location and distance with respect to the root. Hence, the integral of the signal over a region of interest defined in terms of the distance to the root can be used as a response variable which can be analysed using wavelet methods or smoothing splines. Hereafter, we refer to the X-ray (grayscale) density signal just as signal.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
The data consists of CT-scan images of rooted soil from an experiment with 96 lines of wheat pregerminated and grown for 4 weeks in a sandy loam soil sieved to <2mm. The soil was fully saturated and drained to field capacity at the beginning of experiment, and plants were kept in a controlled environment (cycle of 12 hours, 18°C day/12°C night), and were watered every week with nutrient solution. Three scans (with X-ray energy of 140kV, and current of 200µA) were obtained for each volume acquiring a total of 2400 projection images over a 360° rotation. Reconstruction of the volumes and generation of stacks of grayscale images was performed using VG StudioMAX® 2.1 Software.

For stacks of grayscale images, we model the greyscale signal using a mixed model representation of natural cubic splines (Green and Silverman, 1994; Wood, 2006) and performed estimation using REML. Alternatively, we use Symmlet wavelet bases (Daubechies, 1992) to approximate the unknown greyscale function using shrinking methods (See Vidakovic, 2009). We present results from these analyses and discuss the advantages and disadvantages of splines and wavelets for modelling the greyscale signal.

2 Model for the root/soil relation

Define \( v \equiv (x, y, z) \in \mathbb{R}^3 \) the coordinates of the centroid of a voxel of known sides size and volume. Hence, given a 3D volume \( V \subset \mathbb{R}^3 \), denote by \( R \subset V \) the set of voxels corresponding to the root and by \( S \subset V \) the set of voxels not corresponding to the root (soil, pore, water, organic matter, rhizosphere, etc.). Thus \( V = R \oplus S \) (See Figure 1 left). Following Martínez-Araya (2017), denote \( D(v, d) \) to the X-ray density signal (greyscale) at voxel \( v \in S \) which is located at a distance \( d \) with respect to the root, \( R \), defined as \( d = \arg \min_{r \in R} \sqrt{(v-r)^T(v-r)} \). The proximity to the root can affect the mean of the signal \( E(D(v, d)) = \mu(v, d) \) or the covariance \( \text{Cov}(\epsilon(v, d), \epsilon(v+h, d+u)) = C(||h||, |u|) \) with \( h \in \mathbb{R}^3 \) and \( u \in \mathbb{R} \).

Unfortunately, for these approaches traditional estimation procedures in spatial statistics require us to invert a covariance matrix of dimension \( n \times n = (n_x \times n_y \times n_z) \times (n_x \times n_y \times n_z) \) which seems infeasible as the resolution and number of voxels increases.

3 Responses, models and case study

Assume given voxels \( v_1, \ldots, v_n \in S \). To reduce the dimensionality of the data, we define the following random variables. Based on \( D(v, d) \) and given distances \( 0 < d_1 < d_2 < \ldots < d_K < L \) where \( L \) is known, define

\[
U(d_j) = \frac{\sum_{v \in S, c \leq d_j} D(v, c)}{\# \{(v, c) : v \in S, c \leq d_j\}},
\]
and
\[ V(D_j) = \sum_{v \in S, c \in D_j} D(v, c) \frac{\#{\{(v, c) : v \in S, c \in D_j}\}}{\#\{(v, c) : v \in S, c \in D_j\}}, \]
where the set \( D_j = \{d : d \in (d_{j-1}, d_j]\} \) for \( j = 1, \ldots, K \). Thus the variable \( U(d_j) \) is the average signal over a volume \( V(d_j) = \{v : v \in S, d \leq d_j\} \); and the variable \( V(D_j) \) is the average signal over the region or contour \( \mathcal{L}(d_j) = \{v : v \in S, d = d_j\} \).

Define \( m \) non-overlapped regions \( S_1, \ldots, S_m \subset S \) such that \( \bigcup_{i=1}^{m} S_i \subseteq S \). For instance, given \( m+1 \) known values of the vertical coordinates such that \( z_1^* < z_2^* < \ldots < z_{m+1}^* < G \) and \( G \) is known, we define \( m \) intervals \( Z_i = (z_{i-1}^*, z_i^*) \) and regions \( S_i \equiv S(x, y, Z_i) = \{v : v = (x, y, z) \in S, z \in (z_{i-1}^*, z_i^*)\} \) for \( i = 1, \ldots, m \). This definition of regions \( S_1, \ldots, S_m \) allows us to study \( U(d_j) \) and \( V(D_j) \) in vertical direction. For the \( i \)-th region at the \( j \)-th distance we observe \( U_i(d_j) \) and \( V_i(d_j) \), for \( i = 1, \ldots, m \) and \( j = 1, \ldots, K \). Hence, for the \( i \)-th region we observe the trajectories \( U_i = (U_i(d_1), \ldots, U_i(d_K))^T \) and \( V_i = (V_i(d_1), \ldots, V_i(d_K))^T \), and over all the regions (in the vertical direction) we can write \( U = (U_1, \ldots, U_m)^T \) and \( V = (V_1, \ldots, V_m)^T \).

We can model the signal trajectories \( U \) and \( V \) using wavelet methods such that \( \bar{U} = (N^T N)^{-1} N^T U = W \beta \) and \( \bar{V} = (N^T N)^{-1} N^T V = W \vartheta \) where \( N \) is an incidence matrix so that \( (N^T N)^{-1} N^T \) is an averaging matrix; \( W \) is a wavelet orthogonal matrix and \( \beta, \vartheta \) the wavelet coefficients. For each variable, estimation can proceed by minimising \( \bar{T} - \theta)^T (\bar{T} - \theta) + \varphi_\lambda (\theta) \) where \( \bar{T} \equiv W^T \bar{U}, \vartheta \equiv \beta \) or \( \bar{T} \equiv W^T \bar{V}, \theta \equiv \vartheta \) depending on which variable we model, and \( \varphi_\lambda (\theta) \) is a suitable penalty function and \( \lambda > 0 \) a regularization parameter. For instance \( \varphi_\lambda (\theta) \) can be soft or hard thresholding rules (See Vidakovic, 2009).

Alternatively, we can also model the trajectories by using smoothing splines such that \( U = Ng + \epsilon \) and \( V = Nt + \xi \) where \( \epsilon \sim N(0, \sigma_U^2 R_U) \) and \( \xi \sim N(0, \sigma_V^2 R_V) \) are random errors with correlation matrices \( R_U \) and \( R_V \); \( g = (g(d_1), \ldots, g(d_K))^T \), \( t = (t(d_1), \ldots, t(d_K))^T \) with \( g(\cdot) \) and \( t(\cdot) \) unknown smooth double differentiable functions such that \( J(g) < \infty \) and \( J(t) < \infty \) where \( J(f) = \int_0^1 [f''(u)]^2 du = f^T Q f \) where \( Q = \nabla \Delta^{-1} \nabla^T \) is a \( K \times K \) roughness matrix and \( \nabla, \Delta \) are banded matrices as defined by (Green and Silverman, 1994; Wood, 2006). We use the mixed model representations of natural cubic splines \( f = Xb + Za \) where \( X = [1 d] \), \( b \) are fixed effects, \( Z = \nabla (\nabla^T \nabla)^{-1} \), random effects \( a \sim N_{K-2}(0, \sigma^2 a \Delta) \), and \( \lambda = \sigma^2 / \sigma^2_a \) is the smoothing parameter. Estimation of the cubic splines proceeds using REML methods for mixed models. The linearity test \( H_0: f = Xb \) against \( H_1: f = Xb + Za \) is performed as described by Martínez-Araya (2017).

For the ranges of distances studied, a relation between proximity to the segmented root and the signal measured was observed. The higher signal values observed close to the segmented root might be attributed to different patterns of soil aggregation and porosity compared to regions further with respect to the root (See Figure 1 center and right graph).
Acknowledgments: The RootStruc Project is funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC) through grant BB/N014960/1. We acknowledge to Sonja Schmidt, Stephanie Ellis, Jonathan Atkinson, Malcolm Bennet, Sacha Mooney, and Karl Ritz for comments on this paper and for making the data available.

References

Daubechies, I. (1992) Ten lectures on wavelets. SIAM.


A classification tree for functional data

Annette Möller¹, Jan Gertheiss¹

¹ Institute of Applied Stochastics and Operations Research, Clausthal University of Technology, Germany

E-mail for correspondence: annette.moeller@tu-clausthal.de

Abstract: Many standard tools for data analysis already have their functional counterparts tailored to the specific properties of functional data. Specifically, there is a growing interest in classification models that are designed for functional data and allow to utilize information about appearance and shape of the functional observations for the classification procedure. Most of the procedures available, however, are rather inspired by “classical” statistical methods than machine learning approaches. We introduce a novel classification tree specifically designed to deal with functional predictors. The proposed tree may be used as a self-contained approach, but also as a base learner for an ensemble learning method such as a random forest or boosting. Partitioning for a chosen predictor in a specific node of the tree is based on comparing each observational curve in that node to the class-specific mean in terms of a (functional) distance measure. A curve under consideration is assigned to the class to whose mean curve it is closest in terms of the chosen metric.

Keywords: Classification; Functional Data; Functional Logistic Model; Decision Trees; Random Forests.

1 Introduction

Functional data occur frequently in various fields of applications. For many well-known statistical models, e.g. (generalized) linear/additive models, functional counterparts exist that are designed to deal with the specific properties of such data (Ramsay and Silverman, 2005). In functional data analysis the observed data are discretized measurements of an underlying function (a curve) defined on some continuous domain \( T \), such as time, space, frequency or wavelength. A curve is a realization of an infinite-dimensional random variable \( X = \{X(t), t \in T\} \) defined on some (uncountably infinite) set \( T \). A sample of size \( n \) of functional data is represented by

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
a set of random variables $X_1(t), \ldots, X_n(t)$, assumed to be independent and identically distributed as $X$.
In this paper, the aim is to investigate the relationship between a set of functional predictors $x_1(t), \ldots, x_p(t)$ and a scalar response $y$ with a suitable model. To this end we propose a novel classification tree for two-class functional data, relating class membership to information in functional predictors. Recursive binary partition of the predictor space is based on group representative mean curves and their distance to observed curves. For performing the split we investigate a variety of distance measures suitable for functional data.

2 Functional Classification Tree

Suppose we observe functional data of the form $\{y_i, x_{1i}(t), \ldots, x_{pi}(t)\}$, with $p$ functional predictor curves $x_{1i}(t), \ldots, x_{pi}(t)$, $t \in T$, and $y_i$ being the (scalar) binary response value observed at sample $i = 1, \ldots, n$. The domain $T$ of the functional predictors can be of various types, typical examples would be time, space, frequency or wavelength.

The functional tree is grown according to the schematic shown in Figure 1 and briefly explained below.

![Figure 1](image-url)

**FIGURE 1.** Illustration of a functional classification tree.

The algorithm starts with the full data set in the root node and proceeds until the terminal nodes are reached. In the node under consideration, let's say $K$ curves of each predictor $j = 1, \ldots, p$ are available, with $K_1$ curves belonging to group $A$ and $K_2$ curves belonging to group $B$, $K_1 + K_2 = K$. The group $A$ mean curve $\overline{x}_{jA}^i(t)$ of predictor $j$ and the group $B$ mean curve $\overline{x}_{jB}^i(t)$ of predictor $j$ are computed, and used as representatives for the two
Then, every available curve \( r = 1, \ldots, K \) in that node is compared to \( \bar{x}_j^A(t) \) as well as to \( \bar{x}_j^B(t) \) in terms of a (functional) distance measure. In case a curve \( x_{jr}(t) \) is closer to \( \bar{x}_j^A(t) \) in terms of the chosen distance measure it is sent to the left daughter node, in case it is closer to \( \bar{x}_j^B(t) \) the curve is sent to the right daughter node.

The above procedure is performed for the \( K \) curves of each predictor \( j = 1, \ldots, p \) in the current node and the impurity of the resulting split is measured. As with non-functional trees we may use the error rate, the Gini-index or the information criterion to measure split impurity (Hastie et al, 2009). The predictor yielding the purest daughter nodes is chosen as split variable and the split is performed as described above. Having obtained the best split in the current node, the data is partitioned into the respective regions and the process of finding the best binary partition is repeated on the resulting daughter nodes.

Tree growing is continued until a maximum depth is reached. Additionally, a node is not split further in case the number of available observations in that node is smaller than a specified value or in case the node is already pure. In a terminal node, class membership is predicted by majority voting.

To measure the distance between two curves we investigated the Euclidean and the Mahalanobis distance as well as a dissimilarity measure based on the Pearson correlation, and the angular separation which, in essence, measures the angle between two vectors.

To estimate the covariance of the curves we employed a shrinkage approach (Ledoit and Wolf, 2004). However, we plan to explore the possibility to combine shrinkage with covariance smoothing for functional data (see, e.g., Xiao et al, 2017).

In case normal distribution is assumed in the (two) groups, the split decision based on the Mahalanobis distance is strongly related to a classification boundary obtained within the predictor space by discriminant analysis.
3 Applications

We apply our functional classification tree to a data set containing speech signals of arabic digits, available in the R package \texttt{mfds} (Gorecki and Smaga, 2017).

It contains time series of 13 Mel Frequency Cepstrum Coefficients (MFCCs) corresponding to the spoken arabic digits “0”, “1”, ..., “9”. A cepstrum is a representation of the short-term power spectrum of a sound, obtained by inverse fourier transform of the log-spectrum.

Thus, the data consists of 13 functional predictors (the cepstrum coefficients), measured at 93 time points, and a scalar response variable having 10 groups.

For our tree we arbitrarily choose two of the 10 digits to perform 2-group classification. In order to avoid biased results, we investigated several combinations of 2 different digits. In the following we exemplarily present re-
sults for the digits “2”, “8”, and the digits “3”, “7”. Each data set containing 2 out of 10 digits consists of \( n = 1760 \) observations (880 per digit).

Figures 2 and 3 show a few randomly sampled curves (dashed, 10 curves from each group, respectively) belonging to 2 of the 13 cepstrum coefficients, along with the group-wise mean curves (solid). The mean curves, however, were computed from all available curves belonging to the respective group, not only from the ones appearing in the plot. The cepstrum coefficients chosen for the plots are those frequently chosen as very first split variables in the respective trees.

**FIGURE 5.** Boxplots of error rates across 50 splits into test and training data (left), and tree grown on one training set of the arabic digits data (right) when using digits 3 and 7.

For our case study we randomly partition each full (2-class) data set 50 times into a training set of size 1000 and a test set of size 760, grow the tree on the training set and predict the response on the test set. The tuning parameters controlling the maximum depth of the tree and the minimum size of observations in a node to attempt a split are both set to the (fixed) value of 3. Furthermore, we chose the Gini-index as split criterion. We compare the performance of our tree to that of the functional logistic model (Goldsmith et al, 2011) implemented in the R package **refund** (Huang et al, 2016).

Figures 4 and 5 present trees for both classification cases grown on one of the 50 training sets. The left panels in both Figures show that the proposed classification tree substantially outperforms the functional logistic model, regardless which distance metric is used for split decision.

Furthermore, from the right panel of above Figures it is apparent that the nodes become pure very early in the tree growing procedure, yielding relatively small and simple trees. Thus, for the given data sets the proposed functional classification tree seems to be able to capture the group-specific properties of the curves quite well.
4 Outlook

The current version of the proposed tree is designed for binary responses. Possible extensions to multi-class responses will be investigated in future research. Furthermore, alternative ways of finding representative curves might be investigated (Balakrishnan and Madigan, 2006).

As it is well known that a single tree exhibits high variance (see, e.g., Hastie et al., 2009), and ensembles of trees show a much better performance, our tree may be used as base learner for an ensemble of trees. The performance of this ensemble could then be compared to a recently proposed functional random forest (Möller et al., 2016), where each tree in the forest is based on a random partition of the functional curves into intervals.

References


Analysis of Dependent Pseudo-Panel Data: Proposal and Comparison of Three Different Methodological Approaches

Vicente Núñez-Antón¹, Ainhoa Oguiza¹, Jorge Virto¹

¹ Department of Econometrics and Statistics (A.E. III), University of the Basque Country UPV/EHU, Spain.

E-mail for correspondence: vicente.nunezanton@ehu.eus

Abstract: This paper discusses a model for pseudo-panel data when some but not all of the individuals stay in the sample for more than one period. We propose to model the individuals’ time dependency by using three different methodological proposals: a fixed effects longitudinal data analysis approach with specific error structure, a multiple linear regression approach adjusting for heteroscedasticity and autocorrelation, and a linear mixed effects model approach. Data correspond to the Basque labor market for the periods 1993-1999 and 2005-2012, where the latter includes an economical crisis or recession period. Data was previously treated through FORTRAN 77 programming. We construct economically reasonable cohorts based on the available covariates. Results suggest that the effect of the economic crisis on employment rates differs depending on gender and on educational levels. Results also show that the three approaches should be considered because each provides specific practical implications that are worth being assessed.

Keywords: Covariance structures; Economic crisis; Employment rates; Gender; Longitudinal data; Mixed models.

1 Introduction

When we have observations on a set of individuals along different periods of time, we say that we have a “panel of data.” However, we may have observations on sets of individuals that change from one period to another, which do not constitute a panel of data. An example are the data obtained at the Family Expenditure Surveys which are held by many countries. The main difference in our approach with respect to previous research is that we
do not consider the case of independent samples, but rather introduce time dependence between them. Thus, while using pseudo-panels, the work presented in this paper follows a different approach. We model this correlation structure time dependence by using three different methodological proposals: a fixed effects longitudinal data analysis approach with specific error structure, a multiple linear regression approach adjusting for heteroscedasticity and autocorrelation, and a linear mixed effects model approach.

2 Labor market data

The data we use come from a large data base obtained from EUSTAT (1986). The information in this survey corresponds to two time periods, the first one going from the second quarter of 1993 to the fourth quarter of 1999 (i.e., $T = 27$ time periods in the sample). The sample unit is the family and there are observations on different variables, with the following structure: between 1993-II (i.e. second quarter of 1993) and 1997-IV (i.e. fourth quarter of 1997) there are 5000 families in the sample (between 16000 and 16500 individuals) and they remain in the sample for eight time periods; between 1998-I and 1999-IV there are 3750 families (11500-12000 individuals) and they remain in the sample for six time periods. For the second period, the database includes $T = 40$ quarterly periods (i.e., ten years), going from the first quarter in the year 2005 (i.e., 2005-I) to the last quarter in the year 2012 (i.e., 2012-IV), with a rotation or household update of one eighth from one to the next quarter. For this period, there are about 5000 households per quarter in the sample (with an approximate total of about 13500 individuals). The variables selected for the analysis are: active population and employment rate, which will be the response variables for the former and latter period, and year, gender, qualification, head of household status and age, which will be used as explanatory variables.

3 Methodological issues

3.1 Construction of the data cohorts

We only illustrate this process for the period 2005-2012. Given that individuals in the data under study do not remain in the sample for the whole period, the first step in the analysis consists of the specific proposal to build the so-called “representative individuals” (i.e., individual’s cohorts) that we are able to follow along time for the complete period under study. Thus, we build the so-called pseudo panel data of individuals (Deaton, 1985). In our specific data set, cohorts were constructed as a function of the variables gender, educational level and age, which resulted in a total of 30 cohorts. In this way, as usually done when dealing with pseudo panel data settings, we have averaged the employment rates for the different individuals belonging
to each of these cohorts, so that we observe the same cohort or standard individual over time. Moreover, observations corresponding to cohorts are not independent over time and, thus, this behavior needs to be included in any methodological approach considered for their analysis.

3.2 Methodological proposals and results

We propose a model for a pseudo-panel of serially correlated samples:

$$y_{i(t)} t = \alpha_{i(t)} + x_{i(t)} t^T \beta + u_{i(t)} t, \quad t = 1, \ldots, T; \quad i(t) = 1(t), \ldots, N(t),$$

where \( \alpha_{i(t)} \) is the individual fixed effect; \( \beta \) are the slope parameters; \( y_{i(t)} t \) is the value of the dependent variable for individual \( i \) in period \( t \); \( x_{i(t)} t \) are the explanatory variables for individual \( i \) in period \( t \); \( u_{i(t)} t \) is the error term, assumed to be independent and identically distributed; \( i(t) \) denotes the \( i \)th individual and the subindex \( (t) \) is used to indicate that the \( i \)th individual is different in each period. That is, the first individual in period 1, \((i(t) = 1(1))\), is different from the first individual in period 2, \((i(t) = 1(2))\). The data represented above do not constitute a panel. Deaton proposed to group individuals in cohorts using some common characteristic, for instance, age, and calculate the mean of individuals \( i \) belonging to a given cohort (or age group), \( c \). Denoting the number of individuals in cohort \( c \) as \( n_{ct} \), we can write the model as:

$$\frac{1}{n_{ct}} \sum_{i=1}^{n_{ct}} y_{i(t)} t = \frac{1}{n_{ct}} \sum_{i=1}^{n_{ct}} \alpha_{i(t)} + \frac{1}{n_{ct}} \sum_{i=1}^{n_{ct}} x_{i(t)} t^T \beta + \frac{1}{n_{ct}} \sum_{i=1}^{n_{ct}} u_{i(t)} t$$

$$\bar{y}_{ct} = \bar{\alpha}_{ct} + \bar{x}_{ct}^T \beta + \bar{u}_{ct}, \quad c = 1, \ldots, C; \quad t = 1, \ldots, T,$$

We may assume, as in Deaton, that \( n_{ct} \approx n_{c} \) so that:

$$\bar{\alpha}_{ct} = \frac{1}{n_{ct}} \sum_{i=1}^{n_{ct}} \alpha_{i} \approx \frac{1}{n_{c}} \sum_{i=1}^{n_{c}} \alpha_{i} \approx \frac{1}{n_{c}} \sum_{i=1}^{n_{c}} (\alpha_{c} + \epsilon_{it}) = \bar{\alpha}_{c} + \bar{\epsilon}_{ct},$$

where we separate the individual effect, \( \alpha_{i} \), into two terms, \( \alpha_{c} \) and \( \epsilon_{it} \). In model (3), \( \alpha_{c} \) is a fixed effect representing the specificity of each cohort and is assumed to be constant over time, whereas \( \epsilon_{it} \sim N(0, \sigma^{2}_{\epsilon}) \) represents the random effect in the sample. Therefore, we propose a mixed model: on the one hand, \( \alpha_{c} \) represents the specific fixed effect of each cohort, on the other, \( \bar{\epsilon}_{ct} \) (the cohort mean over the individuals), is a random effect that allows for the existence of correlation between different cohorts in time. Therefore, we have:

$$\bar{y}_{ct} = \alpha_{c} + \bar{x}_{ct}^T \beta + (\bar{u}_{ct} + \bar{\epsilon}_{ct}), \quad t = 1, \ldots, T; \quad c = 1, \ldots, C,$$
where $\alpha_c$, is the cohort-specific fixed effect and $\bar{v}_{ct}$ is the model’s total random error. We assume that $\bar{u}_{ct}$ and $\bar{\epsilon}_{ct}$ are uncorrelated, and that the same is true for the explanatory variables and the error term. Results show that the three approaches should be considered because each provide specific practical implications that are worth being considered and very useful for practitioners. The three proposed methodological approaches, as well as some of their most relevant conclusions, are briefly described as follows:

- A fixed effects longitudinal data analysis approach with specific error structure previously selected by using graphical methods and goodness-of-fit criteria, where hypothesis related to nonstationary variances within cohorts, different cohort variances and nonstationary correlations within cohorts are also tested for. Final model selected an MA(3) for the covariance structures for each of the cohorts, and a fixed intercept for each cohort, as well as significant effects of the covariates gender and qualification. We could thus say that, for men in the age group between 25 and 55 years, the proportion of mean active population estimated for men with no qualification and head of the household remained between 84% and 68%, while for younger (i.e. 16-20 years), or older men (i.e. 61-65 years), this percentage was considerably reduced. With regard to the explanatory variables, both gender and qualification were relevant to explain the changes in active population. The variable status did not appear to have a significant influence on the response variable under study.

- A multiple linear regression model approach corrected for heteroscedasticity and autocorrelation, with a mean model that included main effects for year, gender, qualification and age, and interactions between gender and the other covariates, which resulted all significant. Results indicated that the economic crisis has modified the structure of the Basque labor market. Employment rate gender differences that were present in the year 2005 have decreased in the year 2012, resulting in more similar rates for male and female workers. As a final conclusion, we observed that, even though the analysis corresponds to a specific region in Spain, the Basque Country, we could not state that there were no employment rate differences for male and females, or gender equality in terms of employment rates. For the year of the end of the study, 2012, there were still some employment rate differences or gender inequalities that point to the fact that we still have some room for improvement, even for a developed country, in which there exists the conviction that there should be equal employment opportunities for male and female workers. As pointed out by Ghosh (2012), equal opportunity does not imply equal treatment; rather, it means fair treatment. In addition, we have observed that a higher qualification, in terms of holding university degrees, had a positive
effect on the reduction of employment rate gender inequalities or differences. That is, having university studies or a higher qualification significantly contributed to having an employment rate gender equality, when all the aforementioned analyzed employment characteristics were considered. An additional issue worth mentioning is that we could simultaneously study mean employment rates by gender, and for the different educational levels and ages included in the study. For example, we observed that, for the year 2005, mean male employment rates for individuals with university studies for the age interval 55-65 years old, was about 20% higher than that for females with university studies in the same age group. This difference became of 3% for the year 2012, which supported the claim that mean male and female employment rates converge during the economic crisis period, mainly due to the fact that employment rates for female workers with university studies were not really affected as those for males with university studies were.

A linear mixed effects model, with a quadratic fixed effects model for year for each cohort, main effects for gender, age and qualification, as well as the corresponding interaction between these three variables, and subject-specific random intercepts, which resulted significant for the selected model. We observed that, during the pre-crisis period, corresponding to the years 2005 to 2008, employment rates remained basically constant, and then, starting in the year 2009, they showed a clear descending trend. The proposed model illustrates the behavior and evolution of the clear existing difference between male and female employment rates over this period, which, as expected, does somehow depend on the individuals’ age and qualification through the interaction terms included in the proposed model. Employment rates behavior has been worse for males than for females for all types of individuals, although with significant differences among them. Individuals younger than 35 have been the worst hit by the crisis, especially those without university studies, with employment rates decreasing more than 20% (being always larger for males than for females) when compared to a decrease close to 5% for those individuals with university studies. For individuals between 35 and 45 years old, the decrease in employment rates was smaller, even featuring a rate increase for females without university studies, due to a possible added worker effect. For individuals older than 55, there was an employment rate increase, especially for females, also due to a possible added worker effect. As a result of the worse evolution of male employment rates, the employment rate differences between males and females decreased at the end of the crisis period. Employment rate differences were very small for individuals younger than 35 (of about 5%). Moreover, in the case of individuals with university studies, it showed better fe-
male employment rates. For individuals older than 35, although the difference in employment rates showed a significant decrease, there was still a significant disadvantage for females, especially for those without university studies (with differences of more than 10%), despite being the ones showing better employment rates throughout the crisis period.

Acknowledgments: The authors would like to thank the Instituto Vasco de Estadística (EUSTAT) for providing them with the data analyzed in this paper. 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 MTM2013-40941-P (AEI/FEDER, UE), MTM2016-74931-P (AEI/FEDER, UE), ECO2016-76884-P (AEI/FEDER, UE), IT-642-13 and UFI11/03.

References


The Hot Hand in Professional Darts

Marius Ötting¹, Roland Langrock¹, Christian Deutscher¹, Vianey Leos-Barajas¹,²

¹ Bielefeld University, Germany
² Iowa State University, USA

E-mail for correspondence: marius.oetting@uni-bielefeld.de

Abstract: We investigate the hot hand phenomenon in professional darts in a near-ideal setting with minimal to no interaction between players. Considering almost one year of tournament data, corresponding to 152,072 dart throws in total, we use state-space models to reveal moderate serial dependence in throwing performances, thus providing evidence for the existence of the hot hand.

Keywords: Hidden Markov model; Hot hand; Time series; State-space model.

1 Introduction

In sports, the concept of the “hot hand” refers to the idea that athletes may enter a state in which they experience exceptional success. In basketball, for example, the “hot hand” is manifested as a remarkable streak of successful field goal attempts. In their seminal paper, Gilovich et al. (1985) analysed basketball free-throw data to find no support for a hot hand, hence coining the notion of the “hot hand fallacy”. Since then, there has been mixed evidence, with some papers claiming to have found indications of a hot hand phenomenon and others disputing its existence.

Many of the existing studies considered data, e.g. from baseball or basketball, which we believe are hardly suitable for analysing streakiness in performances. For example, in basketball, there are several factors affecting the probability of a player to make a shot, e.g. the position (in a field goal attempt) or the effort of the defence. In particular, an adjustment of the defensive strategy to stronger focus on a player during a hot hand streak can conceal a possible hot hand phenomenon (Bocskocsky et al., 2014).

Here we investigate whether there is a hot hand effect in professional darts, a setting that does not involve the caveats outlined above due to the high level of standardisation of individual throws.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
2 Data

Data was extracted from http://live.dartsdata.com/, covering all professional darts tournaments organised by the Professional Darts Corporation (PDC) between April 2017 and January 2018. In these tournaments, players start each leg with 501 points, and the first player to reach zero wins the leg. To win the match, a player must be the first to win a pre-specified number of legs (typically between 7 and 15). To ensure reliable inference on player-specific hot hand dynamics, we limit our analysis to those players who played at least 100 legs during the time period considered.

At the beginning of legs, players consistently aim at high numbers to quickly reduce their points. The maximum score in a single throw is a triple 20 (T20), but our data indicates the outcomes triple 19 (T19), triple 18 (T18), triple 17 (T17), triple 16 (T16), triple 15 (T15), and bulls-eye (Bull) to be targeted in the initial phase of a leg as well. Thus, in the initial phase of a leg we regard any throw to land in the set $H = \{T15, T16, T17, T18, T19, T20, Bull\}$ as success.

A leg is won once a player reaches exactly 0 points, hence players do not target $H$ towards the end of legs, but rather numbers that make it easier for them to reduce to 0. To retain a high level of standardisation and comparability across throws, we truncate our time series data, excluding throws where the remaining score was less than $c = 180$ points. We thus consider the binary time series $\{y_{p,l}^{n,l}\}_{t=1,...,T_{p,l}}$, indicating the throwing success of player $p$ within his $l$–th leg in the data set, with

$$y_{p,l}^{n,l} = \begin{cases} 1 & \text{if the } t-\text{th throw lands in } H; \\ 0 & \text{otherwise}, \end{cases}$$

where the $T$-th throw is the last throw of the player in his $l$-th leg with the player’s remaining score still greater than or equal to $c = 180$. The final data set comprises $n = 152,072$ throws of a dart, by $P = 50$ players.

3 Model Formulation

We aim at explicitly incorporating any potential hot hand phenomenon into a statistical model for throwing success. Conceptually, a hot hand phenomenon naturally translates into a latent, serially correlated state process for any player considered. For average values of the state process, we would observe normal throwing success, whereas for high (low) values of the state process, we would observe unusually high (low) percentages of successful attempts. Figuratively speaking, the state process serves as a proxy for the “hotness” of the hand. The magnitude of the serial correlation then indicates the strength of any potential hot hand effect. A similar approach was indeed used by Green and Zwiebel (2017), who use a discrete-state Markov
model to measure hotness. While there is some appeal in a discrete-state model formulation, most notably mathematical convenience and ease of interpretation (with cold vs. normal vs. hot states), we doubt that players traverse through only finitely many hotness states, and advocate a continuously varying hotness state variable instead. Specifically, dropping the superscripts $p$ and $l$ for notational simplicity, we consider models of the following form:

$$y_t \sim \text{Bern}\left(\logit^{-1}(\eta_t(s_t))\right), \quad s_t = h(s_{t-1}) + \epsilon_t, \quad (1)$$

where $\{y_t\}_{t=1,...,T}$ is the observed binary sequence indicating throwing success, and $\{s_t\}_{t=1,...,T}$ is the unobserved state process indicating hotness. We thus model throwing success using a logistic regression model in which the predictor $\eta_t(s_t)$ depends, among other things, on the current hotness as measured by $s_t$. The hotness process $\{s_t\}$ is modeled using an autoregressive process, and will include the possibility to be reduced to the nested special case of independent observations, corresponding to absence of any hot hand phenomenon.

The likelihood of a model as in (1) involves integration over $s_t$ for $t = 1, \ldots, T$. An arbitrarily fine approximation of this multiple and intractable integral can be obtained by finely discretising the state space, defining a range of possible values $[b_0, b_m]$ and splitting this range into $m$ intervals $B_i = (b_{i-1}, b_i)$, $i = 1, \ldots, m$, of length $h = (b_m - b_0)/m$. The likelihood of a single throwing history can then be approximated as follows:

$$L_T = \int \cdots \int p(y_1, \ldots, y_T, s_1, \ldots, s_T) \, ds_T \cdots ds_1$$

$$\approx h^T \sum_{i_1=1}^m \cdots \sum_{i_T=1}^m p(b^*_i) p(y_1|b^*_i) \prod_{t=2}^T p(b^*_t|b^*_{t-1}) p(y_t|b^*_t), \quad (2)$$

with $b^*_i$ denoting the midpoint of $B_i$. The computational cost of evaluating the right hand side of Equation (2) is of order $O(T m^T)$. However, the discretisation of the state space effectively transforms the SSM into a hidden Markov model (HMM), with discrete-valued states, such that we can apply the corresponding efficient machinery. In particular, for this approximating HMM, the forward algorithm can be applied to calculate its likelihood at a cost of order $O(T m^2)$ only (Zucchini et al., 2016, Chapter 11). More specifically, defining $\delta = h p(b^*_i)$, $\Gamma = (\gamma_{ij})$ with $\gamma_{ij} = h p(b^*_j | b^*_i)$, and $m \times m$ diagonal matrix $P(y_t)$ with $i$-th diagonal entry equal to $p(y_t|b^*_i)$, the right hand side of Equation (2) can be calculated as a matrix product, such that

$$L_T \approx \delta P(y_1) \Gamma P(y_2) \cdots \Gamma(y_{T-1}) \Gamma P(y_T) \mathbf{1}, \quad (3)$$

with column vector $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^m$. Equation (3) applies to a single leg played by one player. Assuming independence of the individual leg's
throwing histories, the complete likelihood, for the full data set, is obtained as the product of likelihoods of the form above:

\[
L = \prod_{p=1}^{50} \prod_{t_p=1}^{L_p} \delta P(y_{1,1}^{p,t_p}) \Gamma P(y_{2,1}^{p,t_p}) \ldots \Gamma P(y_{T_p,1}^{p,t_p}) 1.
\]

We estimate the model parameters by numerically maximizing the approximate likelihood, subject to the usual technical issues as detailed in Zucchini et al. (2016).

4 Preliminary Results

Before presenting the results of the hot hand model considered, we formulate a model that corresponds to the hypothesis of no hot hand effect being present. This will serve as a benchmark for the SSM to be considered below. The relative frequency of hitting \( H \), i.e. of throwing success in the early stages of a leg, does in fact differ notably across the three throws within a player’s turn, with the empirical proportions of hitting \( H \) in our data found to be 0.359, 0.412 and 0.413 for the first, second and third throw, respectively. The substantial improvement after the first throw within a player’s turn is due to the necessary re-calibration at the start of a turn. To take this into account, in Model 1 we include the categorical covariate \( D_t \), \( D_t \in \{1,2,3\} \), indicating the position of the dart thrown at time \( t \) within the player’s current turn (first, second or third):

\[
\logit(\pi_t) = \beta_{0, p} + \beta_1 I\{D_t=2\} + \beta_2 I\{D_t=3\},
\]

with \( I\{\cdot\} \) denoting the indicator function, and \( \beta_{0, p} \) player \( p \)’s baseline level for the first dart within any given turn. For Model 1, the coefficients \( \beta_1 \) and \( \beta_2 \), which correspond to the increase of throwing success probabilities after the first throw within a player’s turn (on the logistic scale), are estimated as 0.226 and 0.233, respectively.

In Model 2, we now include a hotness state variable \( \{s_t\} \), which we assume to follow an autoregressive process of order 1:

\[
\logit(\pi_t) = \beta_{0, p} + \beta_1 I\{D_t=2\} + \beta_2 I\{D_t=3\} + s_t;
\]

\[
s_t = \phi s_{t-1} + \sigma \epsilon_t,
\]

with \( \epsilon_t \sim N(0, 1) \). Effectively this is a Bernoulli model for throwing success in which the success probability fluctuates around the players’ baseline levels — \( \beta_{0, p}, \beta_{0, p} + \beta_1 \) and \( \beta_{0, p} + \beta_2 \) for within-turn throws one, two and three, respectively — according to the autoregressive process \( \{s_t\} \). The process \( \{s_t\} \) can be interpreted as the hotness variable. For \( \phi = 0 \), the model collapses to our benchmark Model 1 (i.e. absence of a hot hand),
whereas $\phi > 0$ would support the hot hand hypothesis. For the beginning of a leg, we assume $s_1 \sim N(\mu_\delta, \sigma_\delta)$, i.e. that a player’s hotness level starts afresh in every leg according to a normal distribution to be estimated. We fit Model 3 using $m = 150$ and $-b_0 = b_m = 2.5$, monitoring the likely ranges of the process $\{s_t\}$ to ensure the range considered is sufficiently wide given the parameter estimates. Table 1 displays the parameter estimates (except the player-specific intercepts) including 95% confidence intervals based on the observed Fisher information. Crucially, the estimate $\hat{\phi} = 0.583$ strongly supports the hot hand hypothesis, with the associated confidence interval not containing 0. This result corresponds to a considerable correlation in the hotness variable underlying the players’ performances. The AIC also clearly favors the hot hand model formulation, Model 2, over the benchmark given by Model 1 ($\Delta\text{AIC} = 532$).

<table>
<thead>
<tr>
<th>parameter</th>
<th>estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$</td>
<td>0.583</td>
<td>[0.546; 0.620]</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.495</td>
<td>[0.463; 0.529]</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.243</td>
<td>[0.216; 0.270]</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.247</td>
<td>[0.219; 0.275]</td>
</tr>
<tr>
<td>$\mu_\delta$</td>
<td>-0.059</td>
<td>[-0.098; -0.020]</td>
</tr>
<tr>
<td>$\sigma_\delta$</td>
<td>0.685</td>
<td>[0.640; 0.733]</td>
</tr>
</tbody>
</table>

### 5 Outlook

Current research focuses on distinguishing between transitions within a player’s turn to throw three darts (e.g. between first and second, or between second and third throw) and those across the player’s turns (e.g. between third and fourth throw). This extension accounts for the fact that there is a short break in a player’s action between his turns, whereas within a single turn the three darts are thrown in very quick succession. Furthermore, we focus on explicitly addressing the heterogeneity across players as $\phi$ and $\sigma$ and hence the magnitude of the hot hand effect may vary across players. Modelling this variability can be achieved using covariates or, if no suitable covariates are available to explain the heterogeneity, via random effects.

### References


The Weibull MPR model for interval censored survival data

Defen Peng ¹, Kevin Burke ², Gilbert MacKenzie³

¹ BC Centre for Improved Cardiovascular Health, University of British Columbia, Canada
² Department of Mathematics, University of Limerick, Ireland
³ Centre of Biostatistics, University of Limerick, Ireland

E-mail for correspondence: gilbert.mackenzie@ul.ie

Abstract: The Weibull multi-parameter regression (MPR) model with frailty is developed for interval censored survival data. The basic MPR model which is wholly parametric with non-proportional hazards was developed by Burke and MacKenzie in their 2017 Biometrics paper. We describe the basic model, develop the interval-censored likelihood and extend the model to include Gamma frailty and structural dispersion. We present a simulation study and re-analyse data from the Signal Tandmobiel study. The MPR models are shown to be superior to their non-MPR competitors.

Keywords: Gamma frailty; Interval Censoring; MPR survival models; Structural dispersion; Weibull MPR model.

1 Introduction

The concept of multi-parameter regression (MPR) survival modelling was introduced by Burke & MacKenzie (2013) and developed in Burke & MacKenzie (2017). MPR survival models model the scale and shape parameters simultaneously: they are parametric and more flexible than classical proportional hazards (PH) survival models. In their first papers the technique was developed for right-censored survival data. Burke & MacKenzie (2017) reanalysed the survival of 855 incident cases of lung cancer (Wilkinson, 1995) and the MPR Weibull provided a better fit to the data. Here we develop MPR models for interval censored survival data arising in longitudinal studies and use a MPR model with frailty to analyse data from the Signal Tandmobiel study (Gomez et al., 2009).

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
2 MPR Models

There is a wide class of two-parameter parametric survival models with scale and shape parameters. Here, we specialize to the MPR Weibull model because it has proved useful in other work and has the advantage of defaulting to a standard proportional hazards model when the shape parameter is a constant.

2.1 Weibull MPR survival model

The Weibull multi-parameter survival regression model is defined by
\[ \lambda(t_i; \beta, \alpha) = \lambda \gamma t_i^{\gamma - 1}, \]
where, \( \lambda \) is the scale parameter and \( \gamma \) is the shape parameter (Collett, 2003). It follows that the basic survival function is \( S(t_i; \beta, \alpha) = \exp[-\lambda t_i^\gamma] \). Let
\[ \lambda_i = \exp(x_i^T \beta) \quad \text{and} \quad \gamma_i = \exp(z_i^T \alpha) \]
where \( x_{0i} = 1 = z_{0i} \forall i \) are intercept terms in each regression and \( x \) and \( z \) are same covariates, in the same order, but labelled differently in the two linear predictors, for ease of exposition.

2.2 Interval Censoring

From MacKenzie and Peng (2013) a general likelihood for interval censored data, utilizing the actual times at which the \( i \)th patient presents for examination and setting \( \theta = (\beta^T, \alpha^T)^T \) from equation (2), is
\[ L_1(\theta) = \prod_{i=1}^n \left[ S(t_{i,k-1}; \theta) - S(t_{ik}; \theta) \right]^{\delta_i} \left[ S(t_{ci}; \theta) \right]^{1-\delta_i}. \]

Typically the \( i \)th patient fails in the interval \( (t_{i,k-1}, t_{ik}] \) such that there are \( n_{ic} \) interval censored patients and \( n_c \) censored or withdrawn at specific times such that \( n_{ic} + n_c = n \), the total sample size. In (3), the interval censored observations play the role of “failures”.

2.3 Frailty Extension

The Weibull multi-parameter survival regression model with Gamma frailty is defined by \( \lambda(t_i; \alpha, \beta, u) = u\lambda \gamma t_i^{\gamma - 1} \), whence, \( S(t_i; \alpha, \beta, u) = \exp[-u\lambda t_i^\gamma] \), where, \( \lambda = \exp(x^T \beta), \gamma = \exp(z^T \alpha) \), and \( u \) is an unobserved frailty term and the random variable \( U \sim \text{Gamma}(a, b) \). When \( a = b = 1/\phi \), \( E(U) = 1 \) and \( V(U) = \phi \). Then, after some algebra, we may show that \( S(t_i; \alpha, \beta) = [1 + \phi \Lambda(t_i)]^{-1/\phi} \), where, \( \Lambda(t_i) = \lambda t_i^\gamma, \lambda = \exp(x^T \beta), \gamma = \exp(z^T \alpha) \), whence
\[ L(\theta|t_i, \delta_i, x_i, z_i) = \prod_{i=1}^n \left\{ [1 + \phi \Lambda(t_{i,k-1})]^{-1/\phi} - [1 + \phi \Lambda(t_{ik})]^{-1/\phi} \right\}^{\delta_i} \times \left\{ [1 + \phi \Lambda(t_{ic})]^{-1/\phi} \right\}^{1-\delta_i}. \]
TABLE 1. Simulation: MPR Weibull Model: 50% censoring

<table>
<thead>
<tr>
<th>n = 200</th>
<th>Covariates</th>
<th>( \hat{\beta} )</th>
<th>SE</th>
<th>%bias</th>
<th>( \hat{\alpha} )</th>
<th>SE</th>
<th>%bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td>2.0612</td>
<td>0.1658</td>
<td>3.0612</td>
<td>2.0483</td>
<td>0.1050</td>
<td>2.4140</td>
</tr>
<tr>
<td></td>
<td>( x_1 )</td>
<td>0.5167</td>
<td>0.2533</td>
<td>3.3469</td>
<td>0.2496</td>
<td>0.1469</td>
<td>-0.1485</td>
</tr>
<tr>
<td></td>
<td>( x_2 )</td>
<td>0.3263</td>
<td>0.2600</td>
<td>8.7530</td>
<td>-0.0883</td>
<td>0.1498</td>
<td>-11.7465</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>2.0330</td>
<td>0.0981</td>
<td>1.6487</td>
<td>2.0188</td>
<td>0.0626</td>
<td>0.9397</td>
</tr>
<tr>
<td></td>
<td>( x_1 )</td>
<td>0.5072</td>
<td>0.1577</td>
<td>1.4425</td>
<td>0.2532</td>
<td>0.0924</td>
<td>1.2604</td>
</tr>
<tr>
<td></td>
<td>( x_2 )</td>
<td>0.2999</td>
<td>0.1484</td>
<td>-0.0223</td>
<td>-0.1014</td>
<td>0.0873</td>
<td>1.4095</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>2.0099</td>
<td>0.0720</td>
<td>0.4965</td>
<td>2.0101</td>
<td>0.0461</td>
<td>0.5066</td>
</tr>
<tr>
<td></td>
<td>( x_1 )</td>
<td>0.5067</td>
<td>0.1097</td>
<td>1.3426</td>
<td>0.2498</td>
<td>0.0649</td>
<td>-0.0749</td>
</tr>
<tr>
<td></td>
<td>( x_2 )</td>
<td>0.3062</td>
<td>0.1076</td>
<td>2.0646</td>
<td>-0.0979</td>
<td>0.0631</td>
<td>-2.1102</td>
</tr>
</tbody>
</table>

True values: scale=(2, 0.5, 0.3), shape=(2, 0.25, -0.1)

2.4 Structural Dispersion

In the structural dispersion (SD) paradigm we allow the frailty variance to be person-specific via another regression model. Thus

\[
\phi_i = \exp(w_i^T \psi),
\]

where the vector \( w_i \) contains the same covariates as \( x \) and \( z \) and the \( w_{0i} = 1 \nabla i \) is the intercept term.

When the frailty variance is unstructured the model defaults to the standard SPR or MPR Weibull gamma frailty model with \( \phi = \exp(\psi_0) \)

3 Simulation study

We conducted a simulation study to assess the performance of the interval-censored Weibull MPR model. Emergence times were generated from the Weibull regression model (with or without frailty) with two covariates: \( x_1 \), a binary covariate (1 = New treatment and 0 = Old treatment: 50% split) mimicking the treatment effect, and \( x_2 \) a continuous baseline covariate distributed as \( N(0, 0.25) \). The results in Table 1 are based on \( \lambda = 2.0, \beta_1 = 0.5, \beta_2 = 0.3; \alpha = 2.0, \alpha_1 = 0.25, \alpha_2 = -0.1; \phi = 1.0 \), for sample sizes \( n = 200, 500, 1000 \) and a censoring rate of 50%. The number of replications is 1000. The results show that the bias of the mle estimators is acceptably low. The results for the frailty model were similar (not shown). There was some bias in the frailty variance for \( n = 200 \) which disappeared at the larger sample sizes. Overall, the results suggest that the underlying MPR model is recoverable for reasonable sample sizes.
TABLE 2. Selected models fitted and information criteria

<table>
<thead>
<tr>
<th>Model</th>
<th>Scale</th>
<th>Shape</th>
<th>SD</th>
<th>ℓ(θ)</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M5 SPR</td>
<td>c,full</td>
<td>c</td>
<td>-</td>
<td>-5520.08</td>
<td>11048.34</td>
</tr>
<tr>
<td>M6 SPR +GF</td>
<td>c,full</td>
<td>c</td>
<td>c</td>
<td>-5486.08</td>
<td>10982.15</td>
</tr>
<tr>
<td>M7 SPR+GF+SD</td>
<td>c,full</td>
<td>c</td>
<td>c,full</td>
<td>-5483.46</td>
<td>10982.91</td>
</tr>
<tr>
<td>M11 MPR</td>
<td>c,main</td>
<td>c,main</td>
<td>-</td>
<td>-5501.69</td>
<td>11011.39</td>
</tr>
<tr>
<td>M13 MPR</td>
<td>c,full</td>
<td>c,full</td>
<td>-</td>
<td>-5493.68</td>
<td>11003.36</td>
</tr>
<tr>
<td>M14 MPR+GF</td>
<td>c,full</td>
<td>c,full</td>
<td>c</td>
<td>-5466.07</td>
<td>10950.13</td>
</tr>
<tr>
<td>M15 MPR+GF+SD</td>
<td>c,full</td>
<td>c,full</td>
<td>c,full</td>
<td>-5465.62</td>
<td>10943.24</td>
</tr>
</tbody>
</table>

Separate Sexes

<table>
<thead>
<tr>
<th>Model</th>
<th>Scale</th>
<th>Shape</th>
<th>SD</th>
<th>ℓ(θ)</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M18 MPR+GF+SD Girls</td>
<td>c,dfm</td>
<td>c,dfm</td>
<td>c,dfm</td>
<td>-2742.16</td>
<td>5494.32</td>
</tr>
<tr>
<td>M6 SPR+GF Boys</td>
<td>c,dfm</td>
<td>c</td>
<td>c</td>
<td>-2740.17</td>
<td>5490.35</td>
</tr>
</tbody>
</table>

NB: c=intercept, full=gender+dfm+interaction, main=gender+dfm
GF=Gamma Frailty, SD= Structural Dispersion
Some 24 models fitted overall

4 Data Analysis

The Signal Tandmobiel study is a longitudinal prospective oral health study conducted in Flanders (Belgium) from 1996 to 2001. The response was time (yrs.-5) to the emergence of the permanent upper left first premolars. Two covariates were analysed gender (sex): 0 = boy (52%), 1 = girl (48%) and dfm coded: 0 (57%) if the primary predecessor was sound and 1 (43%) if it was decayed, missing or filled. The data were analysed using R programmes and packages including nlm and icfit (R Development Team, 2012; Fay and Shaw, 2010).

Table 2 shows a selection of models fitted to the data. In general, as judged by the AIC, the MPR models were superior to the equivalent SPR models. Gamma frailty models were superior to the equivalent non-frailty models and the SD models sometimes outperformed the equivalent Gamma frailty model. Overall, Model 15 (M15) provided the best fit as judged by the AIC. This is the “full” model containing three separate regressions - scale, shape and structural dispersion - with gender, dfm and their interaction in each model. The fit obtained by Model 15 is shown in Figure 1.

Thus, it transpires that within each sex, the time to emergence does not follow a proportional hazard model. It is fortunate that the toothdata are sufficiently extensive to allow these additional investigations and tests of model fit against non-parametric alternatives to be undertaken. We conclude that both sex and dfm are important and that time to emergence of the permanent upper left first premolars is significantly earlier in girls and those children in whom the predecessor was decayed, missing, or filled and that, likely, there are other unmeasured covariates.
5 Discussion

MPR models are relatively new and are of increasing interest to statisticians working in survival analysis. To the best of our knowledge this is the first time that the effect of interval censoring has been investigated in the MPR survival model setting and the first time that MPR models have been used with Gamma frailty and SD. The models performed well yielding an interesting analysis of the Signal Tandmobiel study data.

References


Coupled Markov-switching regression: inference and a case study using electronic health record data

Jennifer Pohle\textsuperscript{1}, Ruth King\textsuperscript{2}, Mihaela van der Schaar\textsuperscript{3}, Roland Langrock\textsuperscript{1}

\textsuperscript{1} Bielefeld University, Germany
\textsuperscript{2} University of Edinburgh, United Kingdom
\textsuperscript{3} University of Oxford, United Kingdom

E-mail for correspondence: jennifer.pohle@uni-bielefeld.de

\textbf{Abstract}: Coupled hidden Markov models (HMMs) are designed to capture the structure of multivariate time series whose underlying latent state variables interact, but do not evolve synchronously. Here we extend coupled HMMs to allow for covariates in the observed time series, which leads to the class of coupled Markov-switching regression models. The method is applied to electronic health record data of 702 patients from an intensive care unit at the UCLA, where the aim is to gain a better understanding of the course of a disease as well as early-warning signs of potentially critical developments.

\textbf{Keywords}: hidden Markov model; Markov chain; medical statistics; time series

\section{Introduction}

Hidden Markov models (HMMs) are time series models which assume the observations to depend on an underlying unobserved Markov chain with finitely many states. They have been applied in many different areas, e.g. speech recognition, finance, medicine, and ecology (for an overview, see Zucchini \textit{et al.}, 2016). In the case of multivariate time series, within a basic HMM formulation, the variables would be expected to evolve synchronously in the sense that they are driven by the same underlying state sequence. However, in some applications, e.g. in medicine, the observed variables do not necessarily evolve in lockstep, although they may be correlated. For example, a substantial change in a patient’s rate of breathing may or may not be accompanied by immediate visible changes in other vital signs. Coupled...
hidden Markov models (CHMMs) overcome this limitation by assuming separate but correlated state sequences to underlie the different variables observed, hence “coupling” the state processes of multiple HMMs (Brand, 1997).

However, the observations often depend not only on the underlying state, but also on external factors, e.g. the blood pressure of a patient might depend on its general health state as well as certain medications. Therefore, we extend CHMMs to allow for covariates in the observation processes, which leads to the flexible class of coupled Markov-switching regression models (CMSR). We apply this method to electronic health record data collected for 702 patients within the medical intensive care unit at the University of California in Los Angeles (UCLA).

2 Coupled Markov-switching regression models

2.1 Basic formulation of coupled hidden Markov models

We consider an \( M \)-dimensional observed time series of length \( T \), denoted by \( \{Y_t\}_{t=1}^{T} \), with \( Y_t = (Y_{1,t}, \ldots, Y_{M,t}) \). A CHMM for \( \{Y_t\} \) involves \( M \) state sequences, summarized in the vector \( \{S_t\}_{t=1}^{T} \), with \( S_t = (S_{1,t}, \ldots, S_{M,t}) \), where \( S_{m,t} \in \{1, \ldots, N_m\} \). The dependence structure of a CHMM with two underlying state processes is displayed in Figure 1.

![Figure 1. Basic structure of a CHMM with two underlying Markov chains.](image)

Similar to HMMs, given the state sequences, the distribution of \( Y_{m,t} \) is fully determined by the current state \( S_{m,t} \) of its associated state sequence. However, for each state sequence, the future state \( S_{m,t+1} \) depends not only on its current state \( S_{m,t} \) — as would be the case if we were to consider \( M \)
separate HMMs — but on the current states of all $M$ state sequences:

$$\text{Pr}(S_{m,t+1}|S_1, \ldots, S_t) = \text{Pr}(S_{m,t+1}|S_t) \neq \text{Pr}(S_{m,t+1}|S_{m,t}).$$

For simplicity of notation, from now on we assume that the number of states is the same for each state sequence, i.e. $N_m = N$ for all $m$. A CHMM can be written as an HMM with an $N^M \times N^M$ transition probability matrix $\Gamma = (\gamma_{ij}), i, j = 1, \ldots, N^M$, with each state representing an $M$–tuple corresponding to the possible states of $\{S_t\}$. The state space thus is simply the Cartesian product of the $M$ individual state spaces. Using the formulation with extended state space, the parameters can be estimated using standard HMM machinery, in particular by conducting a numerical maximization of the likelihood, which is evaluated using the forward algorithm. For model checking, pseudo residuals can be used, and the states can be decoded using the standard Viterbi algorithm.

### 2.2 Coupled Markov-switching regression models

To allow for covariates in the observation processes, we assume the expectations $\mathbb{E}(Y_{m,t})$ of the state-dependent distributions to be state-specific functions of (variable-specific) covariate vectors $X^m_t = (X^m_{1,t}, \ldots, X^m_{p_m,t})$:

$$\mathbb{E}(Y_{m,t}|s_{m,t}) = \beta_0^{(m,s_{m,t})} + \beta_1^{(m,s_{m,t})} X^m_{1,t} + \ldots + \beta_{p_m}^{(m,s_{m,t})} X^m_{p_m,t}.$$

Thus, the CMSR model comprises state-specific regression functions for each variable $Y_m$ while also taking into account possible interactions in the state processes. Again, the forward algorithm and numerical optimization can be used to find the maximum likelihood estimates. Further extensions, for instance to generalized CMSR models with response distributions from the exponential family, are straightforward.

### 3 Case study on UCLA electronic health record data

Our dataset contains hourly information on vital signs, treatments and also personal data of 702 intensive care unit (ICU) patients at the UCLA, with a total number of 114927 observed time points. All of the patients considered underwent dialysis, which reduces the heterogeneity between the patients in the sense that individuals in a comparable situation are considered. Furthermore, these patients stayed in the ICU for at least 24 hours. Our aim is to model the observed heart rate, respiratory rate, systolic and diastolic blood pressure. The time series of observed vital signs all exhibit substantial changes over time, however, these changes do not always occur synchronously, which motivates the use of CMSR models rather than multivariate HMMs.
FIGURE 2. Estimated (marginal) state-dependent densities of the fitted CMSR model. The black lines correspond to the first state of the model, and the gray lines to the second state. Dashed lines indicate the presence of ventilation, dotted lines indicate the presence of vasopressors. The estimated correlations between systolic and diastolic blood pressure are \( \rho^{(1)} = 0.48 \) and \( \rho^{(2)} = 0.43 \) for state 1 and state 2, respectively.

We fitted a CMSR model with state-dependent normal distributions and \( N = 2 \) possible states per variable, considering the presence of ventilation and vasopressors as dummy covariates for all four vital signs. Since the blood pressure variables are highly correlated, we assume these two variables to depend on the same state sequence, with state-dependent bivariate normal distributions. The resulting marginal state-dependent distributions of our CMSR model are displayed in Figure 2. The estimated distributions are distinct, the model is thus able to capture changes in all four vital signs. However, only the presence of vasopressors is estimated to highly affect the state-dependent distributions, and only for the blood pressure variables (see Table 1).
TABLE 1. Estimated parameters of the state-dependent distributions for the four observed vital signs and both states, respectively.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0^{(1)}$</td>
<td>75.15</td>
<td>16.33</td>
<td>102.41</td>
<td>53.71</td>
</tr>
<tr>
<td>$\beta_0^{(2)}$</td>
<td>100.39</td>
<td>24.49</td>
<td>141.39</td>
<td>69.36</td>
</tr>
<tr>
<td>$\beta_{ven}^{(1)}$</td>
<td>3.16</td>
<td>0.93</td>
<td>5.38</td>
<td>1.79</td>
</tr>
<tr>
<td>$\beta_{ven}^{(2)}$</td>
<td>2.80</td>
<td>1.86</td>
<td>-1.64</td>
<td>-2.16</td>
</tr>
<tr>
<td>$\beta_{vaso}^{(1)}$</td>
<td>0.28</td>
<td>-0.17</td>
<td>-5.35</td>
<td>-1.14</td>
</tr>
<tr>
<td>$\beta_{vaso}^{(2)}$</td>
<td>6.59</td>
<td>-0.70</td>
<td>-17.74</td>
<td>-12.28</td>
</tr>
<tr>
<td>$\sigma^{(1)}$</td>
<td>9.83</td>
<td>3.83</td>
<td>13.81</td>
<td>10.68</td>
</tr>
<tr>
<td>$\sigma^{(2)}$</td>
<td>12.44</td>
<td>5.18</td>
<td>19.60</td>
<td>18.04</td>
</tr>
</tbody>
</table>

4 Conclusion

Our preliminary results suggest that CMSR models are promising tools for analyzing the interplay of latent variables which do not evolve in lockstep, in particular such that are related to the health state of a patient. To decrease the computational costs, sparser parametrization of the transition probabilities could be used, as proposed for example by Saul and Jordan (1999). Current work focuses on model selection, on refinements of the particular model applied to the UCLA data, and on drawing meaningful medical inference.

Acknowledgments: We would like to thank Dr. Scott Hu (Division of Pulmonary and Critical Care Medicine, David Geffen School of Medicine, UCLA) for providing us with the clinical data.

References


Covariate-Dependent Spectral Estimation of Multiple Nonstationary Time Series

Ori Rosen¹, Michael Bertolacci², Edward Cripps², Sally Cripps³

¹ University of Texas at El Paso, U.S.A.
² University of Western Australia, Australia
³ University of Sydney, Australia

E-mail for correspondence: orosen@utep.edu

Abstract: We propose methodology for spectral analysis of multiple nonstationary univariate time series, each associated with a vector of covariates. The method is based on a Bayesian mixture model with weights that depend on the covariates, and the estimation relies on Markov chain Monte Carlo methods. The methodology is illustrated with simulated and real data.

Keywords: Logit-Stick-Breaking Prior; MCMC; Nonstationary Time Series; Whittle Likelihood.

1 Introduction

We propose methodology for fitting local spectra corresponding to multiple nonstationary time series, each associated with multiple covariates. This method extends the AdaptSPEC technique (Rosen et al., 2012) and the methodology described in Krafty et al. (2017). Specifically, Krafty et al. (2017) propose a method that flexibly models the power spectra of multiple stationary multivariate time series as nonparametric functions of frequency and a single covariate observed on each subject. The present paper extends this methodology to multiple nonstationary univariate time series, allowing for multiple covariates on each subject. The model is a mixture of AdaptSPEC components with a logit stick-breaking prior (Rigon and Durante, 2018) on the mixing probabilities, facilitating the dependence on subject-specific covariates. The methodology is applied to data from sleep studies, where it is of interest to study the relationships between signals such as electroencephalography (EEG) from multiple subjects and self-reported sleep

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
outcomes, such as the amount of time slept during the night.

2 AdaptSPEC

This section provides a brief summary of the AdaptSPEC method (Rosen et al., 2012). Let a time series consist of an unknown number of segments, $s$, and let $\xi_{k,s}$ be the unknown location of the end of the $k$th segment, $k = 1, \ldots, s$, where $\xi_{0,s}$ and $\xi_{s,s}$ are $t = 0$ and $t = n$, respectively. Then conditional on $s$ and $\xi_s = (\xi_{0,s}, \ldots, \xi_{s,s})'$, it is assumed that the process $\{X_t\}$ is piecewise stationary. That is,

$$X_t = \sum_{j=1}^{s} X_{tj} \delta_{j,s}(t),$$

(1)

where, for $j = 1, \ldots, s$, the processes $X_{tj}$ are independent and stationary, and $\delta_{j,s}(t) = 1$ if $t \in [\xi_{j-1,s}+1, \xi_{j,s}]$ and 0 otherwise. Consider a realization $\mathbf{x} = (x_1, \ldots, x_n)'$ from process (1), where the number and locations of the stationary segments are unknown. Given a partition $\xi_s$ of the time series $\mathbf{x}$, the $j$th segment consists of the $n_{j,s}$ observations $\mathbf{x}_{j,s} = \{x_t : \xi_{j-1,s}+1 \leq t \leq \xi_{j,s}\}$, $j = 1, \ldots, s$, with underlying spectral densities $f_{j,s}$. Let $M_{j,s} = \lfloor (n_{j,s} - 1)/2 \rfloor$ be the integer part of $(n_{j,s} - 1)/2$, and let $\omega_{m_j} = m_j/n_{j,s}$, $m_j = 1, \ldots, M_{j,s}$ be the Fourier frequencies for the $j$th segment. The discrete Fourier transform (DFT) at frequency $\omega_{m_j}$ for the $j$th segment is then given by

$$y(\omega_{m_j}) = n_{j,s}^{-1/2} \sum_{t=1}^{n_{j,s}} x_{j,s}^{(t)} \exp(-2\pi i \omega_{m_j} t),$$

where $x_{j,s}^{(t)}$ is the $t$th value of $\mathbf{x}_{j,s}$. The periodogram at frequency $\omega_{m_j}$ is then given by $I_{n_{j,s}}(\omega_{m_j}) = |y(\omega_{m_j})|^2$. The likelihood of the $j$th segment is approximated via the Whittle likelihood (Whittle, 1953, Whittle, 1954).

$$L(\mathbf{x}_{j,s} \mid f_{j,s}) \propto \prod_{m_j=1}^{M_{j,s}} f_{j,s}(\omega_{m_j})^{-1} \exp\{ -I_{n_{j,s}}(\omega_{m_j})/f_{j,s}(\omega_{m_j}) \}. \quad (2)$$

We assume that $n_{j,s} \geq t_{\min}$, where $t_{\min}$ is taken to be large enough in order for the local Whittle likelihood to provide a good approximation to the local likelihood. For a given partition $\xi_s$, using (2), the approximate likelihood of the time series is

$$L(\mathbf{x} \mid f_{1,s}, \ldots, f_{s,s}, \xi_s) \propto \prod_{j=1}^{s} \prod_{m_j=1}^{M_{j,s}} \exp\left\{ -\left[ \log f_{j,s}(\omega_{m_j}) + I_{n_{j,s}}(\omega_{m_j})/f_{j,s}(\omega_{m_j}) \right] \right\}. \quad (3)$$

Prior distributions are placed on all the parameters, including the number of segments, $s$, and the partition, $\xi_s$. 
3 Covariate-Dependent Mixture

The proposed model is a mixture of AdaptSPEC components with weights that depend on covariates. Given \( N \) subjects, each with a time series \( x^k = (x_1^k, \ldots, x_n^k)' \) and a covariate vector \( u_k \), the contribution of the \( k \)th subject to the likelihood is given by

\[
L(x^k) \propto \sum_{\ell=1}^r \pi_{\ell}(u_k) L_{k\ell},
\]

where \( L_{k\ell} = L(x^k \mid f_{1, s_\ell}, \ldots, f_{s_\ell, s_\ell}, \xi_{s_\ell}) \) is given in Equation (3), and \( \pi_{\ell}(u_k) \) is the weight attached to component \( \ell \), conditional on \( u_k \). In (4), the partitions, \( \xi_{s_\ell} \), are given, but as in AdaptSPEC, priors are placed on the partitions and the number of segments in each mixture component.

To model \( \pi_{\ell}(u_k) \), we use the logit stick-breaking prior (Rigon and Durante, 2018), given by

\[
\pi_{\ell}(u_k) = v_{\ell}(u_k) \prod_{h=1}^{\ell-1} (1 - v_h(u_k)),
\]

where \( \text{logit}(v_h(u_k)) = \beta_h' u_k \), and \( \beta_h \) are unknown regression parameters. In principle, the number of mixture components is infinite, but in practice, this number is truncated to \( r \) components.

4 Simulated Example

One hundred piecewise stationary time series, each of length 256, are generated from a mixture of 5 components. Let \( z_k \in \{1, 2, 3, 4, 5\} \) indicate, for each \( k = 1, \ldots, 100 \), its cluster membership. Each time series consists of two segments of equal lengths drawn from the autoregressive processes

\[
x_{k,t} = \phi_{z_k, t, 1} x_{k, t-1} + \phi_{z_k, t, 2} x_{k, t-2} + \epsilon_{k,t}
\]

with coefficients given in the following table.

<table>
<thead>
<tr>
<th>( j )</th>
<th>( t \leq 128 )</th>
<th>( t &gt; 128 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p = 1 )</td>
<td>( p = 2 )</td>
<td>( p = 1 )</td>
</tr>
<tr>
<td>( j = 1 )</td>
<td>1.5</td>
<td>-0.75</td>
</tr>
<tr>
<td>( j = 2 )</td>
<td>-0.8</td>
<td>0</td>
</tr>
<tr>
<td>( j = 3 )</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>( j = 4 )</td>
<td>1.5</td>
<td>-0.75</td>
</tr>
<tr>
<td>( j = 5 )</td>
<td>-0.8</td>
<td>0</td>
</tr>
</tbody>
</table>

The \( 2 \times 1 \) covariate vectors, \( u_k \), are distributed evenly over a \( 10 \times 10 \) grid in \([0, 1] \times [0, 1] \).

The model is fit with the number of components truncated to 20. Fits of the time varying log spectra for one time series from each of the clusters are displayed in Figure 1.
5 Application: MESA

The application data come from the Multi-Ethnic Study of Atherosclerosis (MESA) related to sleep. For the analysis, we use data on 49 females and 44 males between the ages 56–89. The dependent variable is EEG for five minutes of the first non-REM period, recorded at 64 Hz, resulting in 19,200 observations per patient. The covariates are age and gender. Figure 2 displays the predicted time varying log spectra for six combinations of the covariates corresponding to the ages: 60, 70 and 80. These predictions are obtained using the mixture weights evaluated at these age values for each gender. From this initial analysis, it appears that EEG is not affected much by gender and age.

Acknowledgments: This work was supported in part by NSF grant DMS-1512188.

References

FIGURE 2. Predicted time varying log spectra for six combinations of the covariates.


Structured additive multiple-output noncrossing Bayesian quantile regression models

Bruno Santos¹, Thomas Kneib¹

¹ Georg-August-Universität Göttingen, Germany

E-mail for correspondence: bruno.santos@uni-goettingen.de

Abstract: Quantile regression models are a powerful tool for studying different points of the conditional distribution of univariate response variables. Their multivariate counterpart extension though is not straightforward, starting with the definition of multivariate quantiles. Here in this work, we propose a flexible Bayesian quantile regression model when the response variable is multivariate, where we are able to define a structured additive framework for all predictor variables. We build on previous ideas considering a directional approach to define the quantiles of a response variable with multiple-outputs (Guggisberg, 2017). We combine this approach with a proposal in the literature to define non-crossing quantiles in every directional quantile model (Rodrigues & Fan, 2017). We define a Markov Chain Monte Carlo procedure for model estimation, where the non-crossing property is obtained considering a Gaussian process design to model the correlation between several quantile regression models. We illustrate the results of these models using German data from the Socio Economic Panel, where the interest lies in explaining more dimensions of inequality in the population, such as income and health, taking into account the dependence between these two variables.

Keywords: Multiple-output response variable; Noncrossing Bayesian quantile regression; Inequality dimensions.

1 Multiple-output Bayesian quantile regression

While univariate quantile regression is already a mainstay in data analysis in different areas, ranging from Ecology to Economics, for instance, multivariate quantile regression models have not attracted the same attention. A first complication for these models is the definition of a multivariate...
quantile, which can be done in different ways. One possible definition to study these multivariate conditional quantiles was proposed by Hallin et al. (2010), where a directional approach is used to define these quantities. If the response variable is defined as \( Y \in \mathbb{R}^k \), then a directional index can be defined by \( \tau \in \mathcal{B}^k := \{ v \in \mathbb{R}^k : 0 < ||v||_2 < 1. \} \), which is a collection of vectors encompassed in the unit ball of \( \mathbb{R}^k \). This directional index can be split into two parts, \( \tau = \tau u \), where \( u \in \mathcal{S}^{k-1} := \{ z \in \mathbb{R}^k : ||z|| = 1 \} \), represents the direction and \( \tau \in (0, 1) \) constitutes the magnitude. If we define \( \Gamma_u \), an arbitrary \( k \times (k-1) \) matrix of unit vectors, where \( (u : \Gamma_u) \) establishes an orthonormal basis of \( \mathbb{R}^k \), then we can determine the \( \tau \)th quantile of \( Y \) as the \( \tau \)th quantile hyperplane that we obtain from the regression of \( Y_{iu} := u^t Y \) on the marginals of \( Y^\perp := \Gamma_u^t Y \) with an intercept term. Moreover, the extension to the more general regression setting, with covariates \( X \), is achieved adding properly this term to the directional regression problem mentioned before. Considering a sample of \( n \) observations of \( (Y, X) \) one can find the directional multivariate quantile regression parameters as the solution of the minimization problem

\[
\min_{\alpha_{\tau}, \beta_{\tau Y}, \beta_{\tau X}} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau}(Y_{iu} - \beta_{\tau Y} Y_{iu}^\perp - \beta_{\tau X} X_i - \alpha_{\tau}),
\]

where \( \rho_{\tau}(u) = u(\tau - I(u < 0)) \) is the check loss function, \( I(.) \) is the indicator function, \( X_i \) is the vector of predictor variables, \( \alpha_{\tau} \) is an intercept term, \( \beta_{\tau Y} \) is a directional quantile regression coefficient and \( \beta_{\tau X} \) is the usual quantile regression coefficient vector.

A Bayesian procedure for this problem was proposed by Guggisberg (2017) and can be formulated considering the equation

\[
Y_{iu} = \beta_{\tau Y} Y_{iu}^\perp + \beta_{\tau X} X_i + \alpha_{\tau} + \epsilon_i,
\]

where \( \epsilon_i \) is distributed according to an asymmetric Laplace distribution with location equal to zero, scale parameter \( \sigma \) and fixing \( \tau \). If we consider a normal prior distribution for \( \theta \), combined with a location scale mixture representation of the asymmetric Laplace distribution, then it is possible to use a Gibbs sampler to draw samples from the posterior distributions of all parameters of interest (see Guggisberg, 2017).

When these conditional quantiles are estimated separately, often one obtains estimates where quantiles cross, violating the monotonicity property of these functions. Taking into consideration multivariate quantiles, it is possible to use the method proposed by Rodrigues and Fan (2017) in each direction of interest in this framework as well. For this, we need to consider that for each fixed direction, we assume there is a correlation between the estimated quantiles. Then one needs to find a value for the bandwidth parameter of an exponential correlation function, for which there are crossing conditional quantiles. Additionally, another feature that is also able to
take into consideration in these models is structured additive predictors, which would add more flexibility in the modelling structure (see Lang et al., 2014). These type of predictors would not pose any greater difficulty for the noncrossing innovation, as this technique rearranges just the estimated quantiles based on the parameter estimates.

2 Application to inequality data in Germany

In order to illustrate how this directional approach can shed new light to the analysis of quantile regression models, we consider data from the Socio Economic Panel collected in Germany in the year of 2012. For this exercise, we are interested in two dimensions of inequality in the population, namely health and income. For the former, we use a standardized health score which gives a good approximation of the physical well being of each person, while for the latter we consider the logarithm of the income of the household.

In order to explain the conditional variation of these two variables, we select a few variables to estimate these directional bivariate quantile regression models. These are age, sex, educational level (edu: split into 4 levels), family status (FamStd: married [1], separated or divorced [2], single [3] or widowed [4]), and a dummy variable to differentiate the West from the East part of Germany (OW). The following model was estimated for 32 equally spaced vectors in the unit circle in $\mathbb{R}^2$ starting with $u = (1,0)$, where the $x$ coordinate is the health score and the $y$ coordinate is log income,

$$
Q_{Y_u}(\tau|X) = \beta_{0u}^u + \beta_{1u}^u Y_u + f_{\tau}^u(\text{age}) + \beta_{2u}^u, \text{edu2} + \beta_{3u}^u, \text{edu3} + \beta_{4u}^u, \text{edu4} + \beta_{5u}^u, \text{FamStd2} + \beta_{6u}^u, \text{FamStd3} + \beta_{7u}^u, \text{FamStd4} + \beta_{8u}^u, \text{OW},
$$

where the indexes $u$ and $\tau$ are used to indicate the dependence of the regression parameters on the direction and the quantile for every model and for $f_{\tau}^u(.)$ we considered cubic P-splines with 20 equidistant knots.

When we analyze this directional method, a first challenge is to visualize all the results, taking into considerations all these chosen directions. In Figures 1, 2 we make a first effort in doing that, where for a given variable and given quantile, we observe the variation of the posterior estimates and its respective credible intervals for different directions. We use colors to differentiate between positive (blue) and negative (red) estimates, and shapes to discriminate between point estimates (squares) and interval estimates (triangles). With this combination of colors and shapes, one can check whether the quantile regression parameter is greater or smaller than zero for a given direction. For instance, considering a given direction where one triangle is blue and the other is red, then we can conclude that the respective credible interval for that parameter contains the zero. On the other hand, if both triangles are blue, then one can conclude that this parameter should be greater than zero, given its credible interval.

Analyzing the results of the plots, we are able to compare the estimates for different directions and different quantiles. For example, in Figure 1 one
can analyze how the estimates change for variable Family status (married), for \( \tau = 0.1 \) and \( \tau = 0.5 \). For the same direction, one could find negative estimates for one quantile and then not different than zero for the other. Generally these estimates change drastically when we compare all 32 directions for these two quantiles, 0.1 and 0.5. And for the dummy variable West region of Germany, Figure 2, one can see how the magnitude of the parameters estimates also change when we compare different quantiles.

As for the nonlinear effect of age, we must take into account not only the different quantiles but also its respective different directions for which we have estimates. In Figure 3, we show the posterior estimates for two

FIGURE 1. Posterior estimates and credible intervals for variable Family status (separated or divorced), considering 32 equally spaced directions in the unit circle, for \( \tau = 0.1, 0.5 \).

FIGURE 2. Posterior estimates and credible intervals for dummy variable West Germany, considering 32 equally spaced directions in the unit circle, for \( \tau = 0.1, 0.5 \).
different directions, \( u_1 = (1/\sqrt{2}, 1/\sqrt{2}) \) and \( u_2 = (-1/\sqrt{2}, 1/\sqrt{2}) \). We can define \( u_1 \) as the models where we give equal weight to the health and income, while \( u_2 \) also gives both variables equal weight, though considers an inverted orientation for health. For the former direction, we say that age has an almost constant effect between 25 to 55 years old and a downwards trend for all other ages. Comparing the effect of age for health and income separately, for instance, using directions \((1, 0)\) and \((0, 1)\), we have that age has a negative effect for health status, while it has a mostly positive effect for income. This means that when we combine these two variables using \( u_1 = (1/\sqrt{2}, 1/\sqrt{2}) \), the positive outcomes of income are not able to surmount the negative effects of the health status, specially for older people. On the other hand for \( u_2 = (-1/\sqrt{2}, 1/\sqrt{2}) \), the effect of age is mainly positive with a noteworthy descending trend around 60 years old and also in the early ages.

![Graph](image.png)

**FIGURE 3.** Nonlinear effect of age using \( \tau = 0.4 \) considering different directions.

Moreover, the combination of the posterior estimates for different quantiles enables the construction of quantile contours or Tukey depth regions. In Figure 4, one can compare these estimates in order to measure the differences between men and women and also between East and West regions of Germany. These plots are able to show for which directions there are differences between these two set of variables. In order to produce these figures, we consider average values for all other predictor variables. Then the quantile region can be defined as the intersection of the upper halfspaces for each quantile when we vary the directions (Hallin et al., 2010). This way one can have a more complete comparison of the effect of these predictors in the bivariate response variable simultaneously. For instance, we can say that the differences between men and women are not as evident as the differences between East and West of Germany. And even when we take a closer look in the plot contrasting these two regions, we can notice that the income dimension produces more divergences between East and West than health status.
3 Final remarks

Quantile regression models are able to provide answers regarding the conditional quantile functions of the response variable given a set of predictors. A multivariate version using a directional approach should be capable of giving more insight of the multivariate structure of data, while taking into consideration the dependence of the set of response variables of interest. For this directional technique, one can take advantage of previous developments for univariate Bayesian quantile regression models, such as structure additive modelling, as well as noncrossing quantile adjustments.

References


A Regularization Approach for the Detection of Differential Item Functioning in Generalized Partial Credit Models

Gunther Schauberger\textsuperscript{1,2}, Patrick Mair\textsuperscript{3}

\textsuperscript{1} Department of Statistics, LMU Munich, Germany
\textsuperscript{2} Chair of Epidemiology, Department of Sport and Health Sciences, Technical University of Munich, Germany
\textsuperscript{3} Harvard University, Massachusetts, USA

E-mail for correspondence: gunther.schauberger@tum.de

Abstract: Most common analysis tools for the detection of differential item functioning (DIF) in item response theory are restricted to the use of single covariates. We propose a penalized likelihood approach for the detection of uniform DIF which is applicable to a broad range of polytomous item response models. A joint model is specified where the possible DIF effects for all items and all covariates are explicitly parameterized. The model automatically detects DIF effects and provides trait estimates that correct for the detected DIF effects from different covariates simultaneously. An application is presented using data from the children’s depression inventory (CDI).

Keywords: Item Response Theory; Differential Item Functioning; GPCM; GPCM-lasso.

1 Introduction

Whenever items are found that show different endorsement probabilities with respect to certain characteristics of the participants, this phenomenon is termed \textit{differential item functioning} (DIF). We present an approach which overcomes several limitations of the DIF methods proposed in the literature. It works for dichotomous as well as polytomous items and it allows researchers to include several covariates of potentially mixed scale levels simultaneously. Technically, it extends the idea proposed by Tutz and Schauberger (2015) where DIF detection in Rasch models is translated into a parameter selection problem. A simultaneous treatment of several

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
covariates becomes important in multicollinear situations, where applying a method repeatedly to different variables is not appropriate. Also, if one is interested in latent trait estimates which are corrected for possible DIF effects, simultaneous treatment of several variables is necessary. While the approach of Tutz and Schaubberger (2015) is restricted to DIF detection based on the Rasch model, the current work aims at the more general generalized partial credit model (GPCM) proposed by Muraki (1992). Therefore, their approach is generalized both to polytomous items and to the incorporation of different item discrimination parameters. Also, besides regular DIF effects of differential step functioning (DSF) as described by Penfield (2007) can be incorporated. DSF implies differences at the level of each step of a polytomous item with respect to different covariate values.

2 A DIF Model for Generalized Partial Credit Models

Let us formally introduce the GPCM as the most general model to be considered in our DIF approach. Let \( Y_{pi} \in \{0, 1, \ldots, k_i\} \) denote the response of person \( p, p \in \{1, \ldots, P\} \), to item \( i, i \in \{1, \ldots, I\} \). Each item \( i \) can differ with respect to the number of response categories \( k_i \). Then, the GPCM can be denoted by

\[
\log \left( \frac{P(Y_{pi} = r)}{P(Y_{pi} = r - 1)} \right) = \beta_i (\theta_p - \delta_{ir}), \quad r = 1, \ldots, k_i.
\]

The model consists of three different types of parameters: the trait parameters \( \theta_p \), the item (location) parameters \( \delta_{ir} \), and the item discrimination parameters \( \beta_i \). To ensure identifiability, for the sequel of the paper we will assume \( \theta_p \sim N(0, 1) \).

The idea of the proposed GPCMlasso model is to have parameters representing DIF instead of examining DIF with various testing approaches. It can be denoted by

\[
\log \left( \frac{P(Y_{pi} = r)}{P(Y_{pi} = r - 1)} \right) = \beta_i (\theta_p - \delta_{ir} - x_p^T \gamma_i), \quad r = 1, \ldots, k_i.
\]

(1)

The GPCM is extended by (possibly) \( m \times \sum_{i=1}^{I} k_i \) additional parameters contained in \( \gamma_i \) which represent DIF and DSF. If the analysis of DSF is not of interest, Eq. (1) can be reduced to

\[
\log \left( \frac{P(Y_{pi} = r)}{P(Y_{pi} = r - 1)} \right) = \beta_i (\theta_p - \delta_{ir} - x_p^T \gamma_i).
\]

(2)

By restricting \( \gamma_i = \gamma_i \), only regular DIF is parameterized (with \( m \times I \) additional parameters).
In both cases, the additional parameters $\gamma$ represent changing item parameters for different values of the respective covariates $x_p$. Therefore, if such a parameter is unequal to zero it indicates uniform DIF or DSF. The new person-specific item parameters can be denoted by $\delta_{irp} = \delta_{ir} + x_p^T \gamma_{i(r)}$.

### 3 Penalized Likelihood Estimation

In our model, parameter selection is equivalent to DIF detection because whenever one of the additional $\gamma$ parameters is unequal to zero it represents DIF. Therefore, we propose to maximize the penalized likelihood

$$l_p(\cdot) = l(\cdot) - \lambda J(\cdot)$$

instead of the regular (marginal) likelihood $l(\cdot)$ where $J(\cdot)$ represents a specific penalty term and $\lambda$ is the tuning parameter. Depending on whether Eq. (2) (i.e., DIF only) or Eq. (1) (i.e., DIF and DSF detection) is assumed, we use the penalty terms

$$J(\theta, \delta, \beta, \gamma) = \sum_{i=1}^{I} \sum_{j=1}^{m} |\gamma_{ij}|$$

or

$$J(\theta, \delta, \beta, \gamma) = \sum_{i=1}^{I} \sum_{j=1}^{m} \sum_{r=1}^{k_i} |\gamma_{ij(r)}| + \sum_{i=1}^{I} \sum_{j=1}^{m} \sum_{r<s} |\gamma_{ij(r)} - \gamma_{ij(s)}|,$$

respectively. In Figure 1, the parameter paths of the DIF/DSF effects along the tuning parameter $\lambda$ resulting from these two penalty terms are depicted for a single exemplary item (5 response categories, 3 covariates). Here, the upper plot illustrates penalty (3) which represents a simple selection penalty for DIF detection while the lower plot illustrates penalty (4) which represents a fusion and selection penalty for DIF/DSF detection. In both cases, the dashed grey lines represent an exemplary model where the item suffers from DIF with respect to variable Var1. DSF would be observed, if in the right plot the parameters corresponding to one covariate would be different (instead of a cluster of identical parameters which represents regular DIF). In general, the optimal value of the tuning parameter is determined by BIC or, alternatively, by cross-validation.

### 4 Application to CDI Data

For illustration, we apply the method to data from the children’s inventory (CDI) originally used in Vaughn-Coaxum et al. (2016). The data set contains data from 1471 participants on 26 items of the CDI. The variables Age, Gender (0: female; 1: male), Race (1: white; 2: black; 3: asian;
FIGURE 1. Parameter paths of DIF parameters (upper plot) and DIF/DSF parameters (lower plot) along log(λ + 1) for an exemplary item to illustrate penalty terms (3) and (4), respectively.

4: hispanic) and Educ_Parents (1: both parents graduated from College; 0: otherwise) will be treated as possibly DIF-inducing covariates.

We focus on the detection of DIF (instead of DSF) and apply GPCMlasso to the CDI data using the penalty (3). Figure 2 shows the coefficient paths for the first two items for all variables along the tuning parameter λ. The red dashed lines represent the optimal model according to BIC. In both items at least one DIF effect is identified by GPCMlasso in the BIC-optimal model. For example, the first item functions differently for males and females.
Overall, only two out of 26 items are diagnosed to be completely DIF-free. The method also provides estimates for the trait parameters $\theta_p$ which are automatically adapted for all DIF effects.
References


An extended random-effects framework for meta-analysis

Francesco Sera12, Antonio Gasparrini12

1 Department of Public Health, Environments and Society, London School of Hygiene & Tropical Medicine, London, United Kingdom,
2 Centre for Statistical Methodology, London School of Hygiene & Tropical Medicine, London, United Kingdom

E-mail for correspondence: francesco.sera@lshtm.ac.uk

Abstract: Extensions of the standard meta-analytical model have been proposed for analyses characterized by complex correlations structures among summary estimates within and between studies. In this contribution, we illustrate a unified framework for meta-analysis based on linear mixed-effects (LME) models, where the correlation between effect sizes is modelled through an extended and flexible random-effects structure. The definition of a unified framework for meta-analysis, complemented with the implementation in a freely-available and fully documented software, will provide researchers with a flexible tool for addressing non-standard pooling problems.

Keywords: Meta-analysis; Linear Mixed Effects; Distributed Lag Non-linear Models.

1 Background

Meta-analysis has become a standard method to summarize evidence in several scientific fields. Traditional applications commonly consider a single outcome estimated from independent studies. However, extensions to deal with more complex meta-analytical problems have been presented, for instance multivariate models for pooling multiple outcomes (Jackson 2011) or multi-parameter associations (Gasparrini 2012), and multilevel versions for hierarchically-structured studies (Stevens 2009). Although these and other extensions were presented separately, all can be described as cases where non-independence among observations within and between studies creates more complex correlation structures that need to be modelled or...
accounted for. In this contribution, we illustrate a unified framework for meta-analysis based on linear mixed-effects (LME) models, where the correlation between effect sizes is modelled through an extended and flexible random-effects structure.

2 A general extended framework for meta-analysis

Following the formulation for LME, a general extended framework for meta-analysis can be written in a single level equation as:

\[ y_i = X_i \beta + Z_i b_i + \epsilon_i, \quad i = 1, \ldots, k, \]

\[ b_i \sim N(0, \Sigma_i), \quad \epsilon_i \sim N(0, S_i), \]

where \( y_i \) is the vector of effect-size estimates, \( X_i \) is the design matrix of the fixed effects, \( \beta \) is the vector of the fixed effects, \( Z_i = (Z_{i1}^{(m)}, \ldots, Z_{i1}^{(2)}) \) is the design matrix of the random effects at \( m-1 \) levels, \( b_i^T = (b_{i1}^{(m)} T, \ldots, b_{i1}^{(2) T}) \) are the random effects at \( m-1 \) levels, and \( \epsilon_i \) is the vector of the level-1 sampling error with known covariance matrix \( S_i \).

The covariance matrix of the random-terms \( \Sigma_i \) is a block diagonal matrix:

\[ \Sigma_i = \text{diag} \left\{ \Sigma_i^{(m)}(n_{i,(m-1)}, \ldots, n_{i,2}), \bigoplus_{i=1}^{n_{i,(m-1)}} \Sigma_i^{(m-1)}, \ldots, \bigoplus_{i=1}^{n_{i,2}} \Sigma_i^{(2)} \right\} \]

are the numbers of units within level \( m-1, \ldots, 2 \), and \( \Sigma_i^{(m)}, \ldots, \Sigma_i^{(2)} \) are the covariance matrices of the random effects at \( m-1 \) levels. The model have marginal distribution \( y_i \sim N(X_i \beta, \Psi_i) \), where \( \Psi_i = S_i + Z_i \Sigma_i Z_i^T \).

2.1 Particular applications of the extended framework for meta-analysis

The standard random-effect meta-analytical model is defined by:

\[ y_i = \beta_0 + b_i + \epsilon_i, \quad i = 1, \ldots, k, \]

\[ b_i \sim N(0, \sigma_b^2), \quad \epsilon_i \sim N(0, \sigma_i^2), \]

where \( b_i \) is the study-specific random effect distributed with between-study variance \( \sigma_b^2 \), and \( \epsilon_i \) is the error term. The univariate meta analytic random effect model can be written in terms of the general extended framework for meta-analysis in (1), with \( X_i = Z_i = 1, \Sigma = [\sigma_b^2], \) and \( S_i = [s_i^2] \).

In the multivariate meta-analysis, \( p \) outcomes or treatments are reported on each study. The multivariate meta-analytic model can be written in terms of the united framework in (1), with \( X_i = Z_i = I_p, \beta^T = [\beta_1 \beta_2 \ldots \beta_p], b_i^T = [b_{i1} b_{i2} \ldots b_{ip}], \) and unstructured between-groups \( \Sigma_i \) and within-groups \( S_i \) covariance matrices.

In same applications, summary effect sizes have natural levels of hierarchy (Stevens 2009). For example, two or more study coming from the same
research group or laboratory. Higher-level units could also be firms, hospital, school districts, or cities. A multilevel meta-analytic model can be written as:

\[
y_{ij} = \beta_0 + b_i + b_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, k \quad j = 1, \ldots, p_i
\]

\[
b_i \sim \text{N}(0, \sigma_{b1}^2) \quad b_{ij} \sim \text{N}(0, \sigma_{b2}^2) \quad \epsilon_{ij} \sim \text{N}(0, s_{ij}^2)
\]

Here $\sigma_{b1}^2$ measures the heterogeneity of the effect sizes across higher levels units (e.g. countries or school districts), while $\sigma_{b2}^2$ measure the variability of the effects estimates across inner levels units (e.g. cities or schools).

The multilevel model can be written in term of the unified framework in (1), with $X_i = I_{pi}$; $Z_i = [1_{pi} \quad I_{pi}]$; $\beta^T = [\beta_0]$; $b_i^T = [b_i \quad b_{i1} \quad b_{i2} \quad \ldots \quad b_{ipi}]$; $\Sigma_i = \text{diag} \{\sigma_{b1}^2, \sigma_{b2}^2, \ldots, \sigma_{b2}^2\}$; $S_i = \text{diag} \{s_{i1}^2, s_{i2}^2, \ldots, s_{ipi}^2\}$.

The degree of complexity could increase further. For example, in dose-response meta-analysis, longitudinal meta-analysis with multiple outcomes collected at various time points, spatial meta-analysis with clustered spatial structure, or combinations of any of the models above. Still, all of these extensions can be defined using the general framework in (1).

3 Inferential procedures and software

We have defined (restricted) maximum likelihood (ML and REML) estimators, with efficient computational strategies based on profiled methods that alternate (restricted) iterative generalized least squares ) IGLS and RIGLS) and Newton-Raphson procedures. In the current implementation, inference is based on asymptotic multivariate normal distribution of (RE)ML estimates, although corrections based for instance on Kenward&Roger or Knapp&Hartung methods are under consideration. The analytic framework and the inferential procedures are implemented in an extended version of the R package mvmeta.

4 Example

The modelling framework and its software implementation will be illustrated using an example of a multi-city analysis evaluating the complex non-linear temperature-mortality associations. Daily mortality and mean temperature data collected through 1985 and 2015 were analysed in 519 locations within 29 countries. A two-stage analysis was performed. In the first stage, non-linear dependencies were modelled using multi-parameter spline functions. The derived set of coefficients (with estimated covariance matrix) were pooled using the extended meta-analytical framework, in this case used to represent a multivariate multilevel meta-analytical problem, where the sets of multiple coefficients are nested within locations and countries. Figure 1 illustrates the temperature-mortality association estimated...
in each country, represented using best linear unbiased predictions (BLUPs) of the multivariate outcomes at the country level. In Figure 2, we present the pooled association together with those estimated for each location in one country (Canada), using BLUPs estimates at the city level.

FIGURE 1. Countries BLUPs temperature-mortality association curves

FIGURE 2. Cities (Canada) BLUPs temperature-mortality association curves

Overall, a high degree of the heterogeneity was observed between both countries and locations within country. Such residual variability can be investigated with the addition of fixed-effects variables in an extended meta-regression.
5 Conclusions

The definition of a unified framework for meta-analysis, complemented with the implementation in a freely-available and fully documented software, provide researchers with a flexible tool for addressing non-standard pooling problems.

References


Non-homogeneous dynamic Bayesian networks with edge-wise coupled parameters

Mahdi Shafiee Kamalabad¹ and Marco Grzegorczyk¹

¹ Johann Bernoulli Institute, Groningen University, Netherlands

E-mail for correspondence: m.shafiee.kamalabad@rug.nl

Abstract: Non-homogeneous dynamic Bayesian networks (NH-DBNs) have become a popular statistical tool for reverse-engineering networks with time-varying parameters. We propose a novel NH-DBN model which infers for each individual network edge if the associated interaction parameter stays similar over time (and should be enforced to be coupled) or if it changes from segment to segment (and should better stay uncoupled). Our empirical results on a benchmark data set from synthetic biology show that the new model yields a higher network reconstruction accuracy than six state-of-the-art NH-DBN models.

Keywords: Dynamic Bayesian network; Non-homogeneous; Edge-wise coupled.

1 Introduction

A challenging statistical problem is to infer networks of interacting variables from time series data. Topical applications stem from systems biology where one objective is to learn the structures of gene regulatory networks from gene expression data. The traditional homogeneous dynamic Bayesian networks (DBNs) are inappropriate when the network parameters are time-dependent. Therefore non-homogeneous DBNs (NH-DBNs) were proposed in the literature. NH-DBNs infer the network structure along with a segmentation of the data into disjoint segments with segment-specific interaction parameters. To avoid model over-flexibility, more recently proposed NH-DBNs feature coupling mechanisms that enforce the segment-specific interaction parameters to stay similar among segments; see, e.g., Grzegorczyk and Husmeier (2012). However, enforcing the parameters to stay similar over time can also have counter-productive effects. To fix this deficit, a more flexible partially segment-wise coupled NH-DBN and a generalized

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
A fully coupled NH-DBN with segment-specific coupling strengths were proposed; see Shafiee Kamalabad and Grzegorczyk (2016, 2017, 2018). In this paper, we propose a new NH-DBN with partially edge-wise coupled parameters. Unlike all earlier proposed coupled NH-DBNs, the new model operates edge-wise. It infers for each individual edge if the corresponding parameter should be coupled or not. That is, the new model infers the best trade-off between an uncoupled and a coupled NH-DBN from the data.

2 Methods

Consider a piece-wise linear regression model with response $Y$ and a covariate set $\pi = \{X_1, \ldots, X_k\}$. We assume that the data points have a temporal order and can be divided into disjoint segments $h = 1, \ldots, H$, where each $h$ has specific regression coefficients, $\beta_h = (\beta_{h,0}, \beta_{h,1}, \ldots, \beta_{h,k})^T$. Let $y_h$ be the response vector and $X_h$ be the design matrix for segment $h$, where each $X_h$ has a first column of 1's for the intercept ($\beta_{h,0}$). We assume:

$$y_h \sim N(X_h \beta_h, \sigma^2 I)$$

where $\sigma^{-2} \sim GAM(0.005, 0.005)$, and $\beta_1 \sim N(0, \sigma^2 \lambda_u I)$ with hyperparameter $\lambda_u^{-1} \sim GAM(2, 0.2)$. For segments $h > 1$ we use a vector of binary variables, $\delta = (\delta_0, \ldots, \delta_k)$, to define the new prior:

$$\beta_h \sim N(\delta \odot \tilde{\beta}_{h-1}, \sigma^2 \cdot diag\{\lambda_c \delta + \lambda_u (1 - \delta)\})$$

where $\tilde{\beta}_{h-1}$ is the posterior expectation of $\beta_{h-1}$, $\lambda_c^{-1} \sim GAM(2, 0.2)$, $\odot$ is the Kronecker product, and $diag\{x\}$ denotes a diagonal matrix whose elements are the elements of the vector $x$. The new prior yields a consensus:

- For $\delta = 0$ we have $\beta_h \sim N(0, \sigma^2 \lambda_u I)$ for all $h$, and the model is fully uncoupled; i.e. every regression coefficient has prior expectation 0.
- For $\delta = 1$ we have $\beta_h \sim N(\tilde{\beta}_{h-1}, \sigma^2 \lambda_c I)$ for $h > 1$, and the model is fully coupled. The whole regression coefficient vector $\beta_h$ is enforced to stay similar to $\tilde{\beta}_{h-1}$ (and hence to stay similar to $\tilde{\beta}_{h-1}$).
- The new model infers $\delta = (\delta_0, \ldots, \delta_k)$ from the data, so as to search for the best trade-off between an uncoupled and a fully coupled model. $\delta_i = 0$ indicates that the coefficients $\beta_{1,i}, \ldots, \beta_{H,i}$ for $X_i$ are uncoupled, while they are coupled (enforced to stay similar) for $\delta_i = 1$.

A priori we assume the $\delta_i$ to be i.i.d. Bernoulli distributed, $\delta_i \sim BER(0.5)$. For the posterior distribution, we have:

$$p(\beta_1, \ldots, \beta_H, \sigma^2, \lambda_u, \lambda_c, \delta|y_1, \ldots, y_H) \propto \prod_{h=1}^{H} p(y_h|\sigma^2, \beta_h) \cdot p(\lambda_u) \cdot p(\lambda_c)$$

$$\cdot p(\sigma^2) \cdot p(\delta) \cdot P(\beta_1|\lambda_u, \sigma^2) \cdot \prod_{h=2}^{H} P(\beta_h|\lambda_u, \lambda_c, \sigma^2, \delta, \tilde{\beta}_{h-1})$$
For the model the marginal likelihood \( p(y_1, \ldots, y_H | \lambda_u, \lambda_c, \delta) \) can be computed in closed form and the parameters can be sampled from their full conditional distributions. Making use of the marginal likelihood, we get:

\[
p(\lambda_u, \lambda_c, \delta | y_1, \ldots, y_H) \propto p(y_1, \ldots, y_H | \lambda_u, \lambda_c, \delta) \cdot p(\lambda_u) \cdot p(\lambda_c) \cdot p(\delta)
\]

When the covariate set \( \pi \) and the data segmentation are unknown, we infer both from the data. We assume all covariate sets with up to 3 covariates to be equally likely, while \( p(\pi) = 0 \) if \( |\pi| > 3 \). As prior on the no. of segments \( H \) we take a Poisson distribution, \( H \sim \text{Poi}(1) \), and we identify \( H \) with \( H - 1 \) changepoints, \( \tau = \{\tau_1, \ldots, \tau_{H-1}\} \). Data point \( t \) is in segment \( h \) if \( \tau_{h-1} < t \leq \tau_h \). We assume the changepoints to be distributed like the even-numbered order statistics of \( 2(H-1) + 1 \) uniformly distributed points.

We use Markov Chain Monte Carlo (MCMC) simulations to take samples \( \{\pi^{(w)}, \tau^{(w)}, \lambda_u^{(w)}, \lambda_c^{(w)}, \delta^{(w)}\}_{w=1,\ldots,W} \) from the posterior \( p(\pi, \tau, \lambda_u, \lambda_c, \delta | D) \), where \( D \) is the data, and the changepoint set \( \tau \) implies the segmentation of \( D \) into \( H = |\tau| + 1 \) segments with response vectors \( y_1, \ldots, y_H \). For sampling \( \pi \) and \( \tau \) we implement Metropolis-Hastings MCMC moves: Changepoint birth, death and re-allocation moves for \( \tau \), and covariate addition, deletion and exchange moves for \( \pi \).

To learn a network with \( n \) variables \( Y_1, \ldots, Y_n \) we apply the model to each \( Y_i \) separately. The potential covariate sets \( \pi_i \) for response \( Y_i \) are all subsets of \( \{Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n\} \). As the interactions are subject to a time delay, the segment-specific design matrices \( X_{ij} \) are built from the covariate values at the preceding time points. For each possible network edge, \( Y_j \rightarrow Y_i \), we estimate the score (marginal edge posterior probability) \( \hat{e}_{j,i} = \frac{1}{W} \sum_{w=1}^{W} I_{j \rightarrow i}(\pi_i^{(w)}) \), where \( I_{j \rightarrow i}(\pi_i^{(w)}) = 1 \) if \( Y_j \in \pi_i^{(w)} \), and \( I_{j \rightarrow i}(\pi_i^{(w)}) = 0 \) else. When the true edges are known, the prediction accuracy can be quantified via the area under the precision-recall curve (AUC). The network reconstruction accuracy increases in the AUC.

**Overview to competing NH-DBNs:**
Six competing NH-DBN models are given by:

- **DBN:** The conventional homogenous DBN has no changepoints, \( H = 1 \). One regression coefficient vector \( \beta_1 \) applies to all data points.

- **UNCOUPLED NH-DBN:** Our model reduces to an uncoupled model for \( \delta = 0 \). The priors are: \( \beta_h \sim N(0, \sigma^2 \lambda_u I) \) for all \( h \).

- **FULLY (SEQUENTIAL) COUPLED NH-DBN:** This model from Grzegorczyk and Husmeier (2012) automatically couples all neighbouring regression coefficients with the same strength. Our new model reduces to a fully coupled model when setting \( \delta = 1 \). The priors are: \( \beta_1 \sim N(0, \sigma^2 \lambda_u I) \) and \( \beta_h \sim N(\beta_{h-1}, \sigma^2 \lambda_c I) \) for \( h \geq 2 \).
FIGURE 1. PCA plot. Every MCMC simulation outputs scores $\hat{e}_{i,j}$ for all edges. We arrange the scores of each simulation vector-wise and z-score-standardize all vectors. We project the vectors onto the first two principle components (PCs).

- **SWITCH (UNCOPLED/COUPLED) NH-DBN**: This model switches between an uncoupled and a fully coupled NH-DBN. The priors are:

  $$
  \beta_h \sim \begin{cases} 
  N(0, \lambda_u \sigma^2 I) & \text{if } \delta^* = 0 \text{ or } h = 1 \\
  N(\tilde{\beta}_{h-1}, \lambda_c \sigma^2 I) & \text{if } \delta^* = 1 \text{ and } h > 1
  \end{cases}
  $$

  where $\delta^* \sim BER(0.5)$, indicates the model. For $\delta^* = 0$ the model is uncoupled. For $\delta^* = 1$ the model switches to a coupled model.

- **PARTIALLY SEGMENT-WISE COUPLED NH-DBN**: This model from Shafiee Kamalabad and Grzegorczyk (2016) uses a partially segment-wise coupling concept. It infers for each segment $h > 1$ whether it is uncoupled from or coupled to the preceding one. The coupling (uncoupling) always applies to all covariates. The priors are:

  $$
  \beta_h \sim \begin{cases} 
  N(0, \lambda_u \sigma^2 I) & \text{if } \delta_h^* = 0 \text{ or } h = 1 \\
  N(\beta_{h-1}, \lambda_c \sigma^2 I) & \text{if } \delta_h^* = 1 \text{ and } h > 1
  \end{cases}
  $$

  where $\delta_1^* := 0$, and $\delta_h^* \sim BER(0.5)$ for $h > 1$. $\delta_h^* = 1$ indicates that segment $h$ is coupled to segment $h - 1$, while $\delta_h^* = 0$ indicates that it is uncoupled. At each changepoint all regression coefficients stay either similar ($\delta_h^* = 1$) or get dissimilar ($\delta_h^* = 0$). The vector $\delta = (\delta_1, \ldots, \delta_{k+1})^T$ of length $k+1$ is replaced by the vector $\delta^* = (\delta_2^*, \ldots, \delta_H^*)^T$ of length $H - 1$.

- **GENERALIZED FULLY COUPLED NH-DBN**: This model from Shafiee Kamalabad and Grzegorczyk (2017,2018) generalizes
the fully coupled NH-DBN. It introduces segment-specific coupling parameters $\lambda^h_c$:

$$\beta_h \sim \begin{cases} N(0, \lambda_u \sigma^2 I) & \text{if } h = 1 \text{ or } h = 1 \\ N(\bar{\beta}_{h-1}, \lambda^h_c \sigma^2 I) & \text{if } h > 1 \text{ and } h > 1 \end{cases}$$

where $\lambda^h_c \sim GAM(a_c, b_c)$ for $h = 2, \ldots, H$. Coupling applies to all regression coefficients, but the coupling strengths are segment-specific.

3 Empirical study and results

By means of synthetic biology, Cantone et al. (2009) designed a network in *S. cerevisiae* (yeast), and then measured in vivo gene expressions under galactose- and glucose-metabolism. As the true network is known and the data are real wet-lab gene expression data, we use this data set to compare the new NH-DBN with its competitors. For each of the 7 models, we run 10 MCMC simulations with $100k$ iterations, i.e. 70 simulations in total. For each simulation we then arrange the simulation-specific edge scores in a vector, and we use a principle component analysis (PCA) to visualize similarities between the 70 vectors. To this end, we z-score-standardize all vectors and project them onto the first two principle components (PCs). It can be seen from Figure 1 that the model-specific simulations are grouped together (convergence) and that our new model is nearest to the true network. Figure 2 shows that the new model also yields the highest average AUC (prediction accuracy) among all models. Figure 3 shows the true yeast network and a network predicted with the new NH-DBN model.

4 Conclusions

We have proposed a new NH-DBN with partially edge-wise coupled parameters. Unlike all earlier proposed coupled NH-DBNs, our new model
Figure 3. Yeast network. Left: The true network. Right: Network predicted with our new model. As the true network has 8 edges, we extracted the 8 edges with the highest average scores. Grey edges refer to false positive edges. The edge labels give the edge scores.

operates edge-wise. It infers for each individual edge if the corresponding parameter should be coupled or not. On a benchmark yeast gene expression data set, our new NH-DBN model has reached a higher network reconstruction accuracy than six competing NH-DBN models.

References


Accelerated estimation of mixture models for rating variables

Rosaria Simone

1 Department of Political Sciences, University of Naples Federico II, Italy

E-mail for correspondence: rosaria.simone@unina.it

Abstract: For incomplete problems like those involving mixture models, Louis’ identity for the observed information matrix allows for its direct computation within the Expectation-Maximization algorithm. For a class of mixture models introduced for rating data, the technique is applied together with the corresponding acceleration strategy: as a consequence, a direct formulation of standard errors can be possibly obtained and best-subset variable selection can be applied with reasonable computational efforts.

Keywords: Louis’ identity; Information matrix; CUB models

1 Introduction

Recently, increasing efforts are paid to take into consideration measurement errors also when collecting ordinal scores, like rating or ranking data. For discrete outcomes conveying latent perception, the concept of error reminds to the fuzziness of the choice. In order to combine the expression of the true scores with such measurement errors, a mixture paradigm can be applied: this is the case of CUB models (D’Elia and Piccolo, 2005). This rationale assumes that the choice among \( m \) ordered alternatives is driven by two main components: the feeling, which is modelled by means of a shifted Binomial distribution \( b_j(\xi_i) \), and the uncertainty (indecision, fuzziness), which is modelled by means of a discrete Uniform distribution. Then, for a random sample \( (R_1, \ldots, R_n) \):

\[
Pr(R_i = j \mid y_i, w_i, \theta) = \pi_i b_j(\xi_i) + (1 - \pi_i) \frac{1}{m},
\]  

(1)
for \( i = 1, \ldots, n \) and \( j = 1, \ldots, m \). Selected covariates can be possibly specified with a logit link to model parameters \( \theta = (\beta', \gamma')' \):

\[
\text{logit}(\pi_i) = y_i \beta; \quad \text{logit}(\xi_i) = w_i \gamma. \tag{2}
\]

### 1.1 The EM algorithm

Being an incomplete problem, estimation of model parameters can be carried out by the implementation of the Expectation-Maximization algorithm (Piccolo, 2006): then, analytical derivation of the (observed) information matrix \( I(\theta; r) \) has been pursued. In full generality, given the model equations:

\[
\text{logit}(\pi_i) = y_i \beta, \quad \text{logit}(\xi_i) = w_i \gamma
\]

for a full CUB model specification with covariates, the estimable parameter vector \( \theta' = (\beta, \gamma)' \) characterizes the probability distribution of the outcome \( R_i \) of the \( i \)-th subject as follows:

\[
Pr(R_i = r_i | \theta, y_i, w_i) = \pi_i b_{r_i}(\xi_i) + (1 - \pi_i) \frac{1}{m},
\]

being \( y_i, w_i \) (row) vectors of covariates’ values. Then, let us consider the complete log-likelihood:

\[
l_c(\theta) = \sum_{g=1}^{2} \sum_{i=1}^{n} \left[ z_{gi} \log(a_{gi}) + z_{gi} \log(p_g(r_i; \eta_g)) \right], \tag{3}
\]

where \( a_g \) denotes the mixing proportion of the \( g \)-th component in the mixture, governed by the probability distribution \( p_g(r_i; \eta_g) \) for given parameter vector \( \eta_g \), and the \( z_{gi} \) are the missing data:

\[
z_{gi} = \begin{cases} 
1 & \text{if the } i\text{-th observation is drawn from the } g\text{-th component} \\
0 & \text{otherwise.}
\end{cases}
\]

Then, at the \( k \)-th iteration:

\[
\mathbb{E}[z_{gi}|\theta^{(k)}] = \tau_{gi}^{(k)} = \frac{a_{gi}(\psi_g^{(k)}) p_g(r_i; \eta_g^{(k)})}{\sum_{j=1}^{2} a_{ji}(\psi_j^{(k)}) p_j(r_i; \eta_j^{(k)})}. \tag{4}
\]

In particular, in our case:

\[
\tau_{1i}^{(k)} = \frac{\pi_i^{(k)} b_{r_i}(\xi_i^{(k)})}{Pr(R_i = r_i | \theta^{(k)}, y_i, w_i)}, \quad \tau_{2i}^{(k)} = 1 - \tau_{1i}^{(k)},
\]

where one sets:

\[
\text{logit}(\pi^{(k)}) = y_i \beta^{(k)} \quad \text{logit}(\xi^{(k)}) = w_i \gamma^{(k)}.
\]
Thus, at the \( k \)-th step, the expectation of the complete log-likelihood (3) to be maximized over \( \theta \) is given by:

\[
Q(\theta; \theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}[l_c(\theta)] = \sum_{i=1}^{n} \tau_{1i}^{(k)} \log(\pi_i) + \sum_{i=1}^{n} (1 - \tau_{1i}^{(k)}) \log(1 - \pi_i) + \sum_{i=1}^{n} \tau_{1i}^{(k)} \log(b_{r_i}(\xi_i)) + \sum_{i=1}^{n} (1 - \tau_{1i}^{(k)}) \log\left(\frac{1}{m}\right).
\]

yielding to the updated estimate \( \theta^{(k+1)} \).

2 The Louis Method: the complete likelihood

The observed information matrix of an incomplete problem can be expressed as:

\[
I(\theta; r) = I_c(\theta; r) - I_m(\theta; r)
\]

where \( r = (r_1, \ldots, r_n) \) denotes the observed sample, \( \theta \) is the estimable parameter vector and \( I_c(\theta; r) \) denotes the likelihood of the complete problem. Louis’ identity (Louis, 1982) provides a precise representation of the missing information \( I_m(\theta; r) \), yielding to:

\[
I(\theta; r) = I_c(\theta; r) - V_c(\theta; r) + V(\theta; r)
\]

where

\[
I_c(\theta; r) = \mathbb{E}[I_c(R, Z; \theta) | R = r]
\]

\[
V_c(\theta; r) = \mathbb{E}[S_c(R, Z; \theta) \cdot S_c(R, Z; \theta)^T | R = r]
\]

\[
V(\theta; r) = S(\theta; r) \cdot S(\theta; r)^T.
\]

Once the complete information matrix and the complete score function \( S_c(R, Z; \theta) \) are derived, at each step of the EM procedure the observed information matrix can be evaluated (see also Oakes, 1999). In particular, at the last iteration it will return the observed information matrix computed at the ML parameter estimates \( I(\hat{\theta}; r) \). The advantages of Louis’ identity stem from the fact that the second derivatives of the complete log-likelihood are in general easier to compute than those of the incomplete problem and this task can be pursued with elements computed within the EM steps.

2.1 Derivation of standard errors

At that point, \( I(\hat{\theta}; r) \) should be inverted to obtain the (asymptotic) variance-covariance matrix and thus derive estimates of the standard errors. Given the particular form that the components in the Louis identity assume for
CUB models, it is possible to obtain a direct expression of the standard errors by applying the following result from matrix algebra (Miller, 1981):

Let $G$ and $H$ be matrices, such that $G + H$ is not singular, with $\text{rank}(H) = r$. Let $H = E_1 + \cdots + E_r$, with $\text{rank}(E_i) = 1$ and such that:

$$C_{k+1} = G + E_1 + E_2 + \cdots + E_k$$

is non-singular for every $k = 1, \ldots, r$. Then, if $C_1 = G$,

$$C_{k-1}^{-1} = C_k^{-1} - \nu_k C_k^{-1} E_k C_k^{-1}$$

where

$$\nu_k = \frac{1}{1 + \text{tr}(C_k^{-1} E_k)}.$$  

In particular, $G + H = C_{r+1}$ and

$$(G + H)^{-1} = C_r^{-1} - \nu_r C_r^{-1} E_r C_r^{-1}$$

where

$$\nu_r = \frac{1}{1 + \text{tr}(C_r^{-1} E_r)}.$$  

### 2.2 Acceleration strategy

Within the EM algorithm, it is natural to consider the mapping $\mathcal{M}: \Theta \to \Theta$, $\Theta$ being the parameter space, such that:

$$\mathcal{M}(\theta(t)) = \theta(t+1).$$

Due to convergence, the ML estimate $\hat{\theta}$ is a fixed point $\mathcal{M}(\hat{\theta}) = \hat{\theta}$, and thus a first-order Taylor expansion around $\hat{\theta}$ provides the following relation:

$$\theta(t+1) - \hat{\theta} \approx J(\hat{\theta})[\theta(t) - \hat{\theta}]$$  

(10)

where $J(\theta)$ is the Jacobian matrix of $\mathcal{M}$:

$$J(\theta) = \left( \frac{\partial M_i(\theta)}{\partial \theta_j} \right)_{i,j=1,\ldots,d}$$

where $\theta = (\theta_1, \ldots, \theta_d)'$. By re-arranging terms and for sufficiently large $t$, one obtains:

$$\hat{\theta} \approx \theta(t) + (I - J(\hat{\theta}))^{-1}(\theta^{t+1} - \theta(t))$$  

(11)

where $I$ denotes the $d$-dimensional identity matrix. Thus, after a certain number of steps, say $\text{iter}_c$, Louis proposes to accelerate the convergence by updating the estimate $\theta^{(k)}$ with its approximation:

$$\tilde{\theta}^{(k)} \approx \theta^{(k-1)} + (I - J(\theta^{(k)}))^{-1}(\theta^{(k)} - \theta^{(k-1)})$$  

(12)

and by using it to initialize the $k + 1$ step. For the computation of $(I - J(\theta^{(k)}))^{-1}$, Louis found the following convenient representation:

$$(I - J(\theta^{(k)}))^{-1} = \mathcal{I}^{-1}(\theta^{(k)}; r) \mathcal{I}_c(\theta^{(k)}; r).$$  

(13)
2.3 A case study: best-subset variable selection

Among the possible variable selection techniques, the best-subset search is in principle the most time-consuming one, especially in cases the optimization procedure is iterative, as for the EM algorithm. On the other hand, for CUB models the best-subset approach is particularly appealing since in full generality one should combine the search for covariates for the uncertainty component with the search for covariates for the feeling component. Then, given an appropriate accelerated estimation procedure, the best-subset selection for CUB models gains feasibility. For illustrative purposes, this approach is shown on a case study involving data taken from the GESIS ALLBUS German survey 2016. The chosen variable from the 2012 wave is the importance attributed to politics and public life: measurements were collected on a \( m = 7 \) point scale, ranging from 1 = ‘unimportant’ to 7 = ‘very important’. Imagine one wishes to understand if and how age, gender (\( \text{Gender} = 1 \) for women), marital status (\( \text{Marital} = 1 \) for people in a relationship) and having the German citizenship are affecting such evaluation and to that purpose one wishes to implement a best-subset strategy. With only 4 covariates, this approach requires the testing and fitting of 225 models: according to the BIC index, the best explanatory features are granted by the model equations (\( 
\begin{align*}
\text{logit}(\pi_i) &= 0.580 - 0.042 \text{Age}_i + 2.062 \text{Gender}_i + 0.869 \text{Marital}_i \\
\text{logit}(\xi_i) &= 0.122 - 0.009 \text{Age}_i + 0.181 \text{Gender}_i
\end{align*}
\) )

The standard EM procedure for this model requires 35 iterations and 4.519 seconds, whereas the accelerated procedure yielded the same results with 14 iterations and with 1.906 seconds. Table 1 reports the mean number of iterations and mean duration time for estimation of all 225 models.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Number of Iterations</th>
<th>Mean time</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>29.07</td>
<td>3.033</td>
</tr>
<tr>
<td>Fast EM</td>
<td>13.57</td>
<td>1.327</td>
</tr>
</tbody>
</table>

3 Conclusions and on-going developments

The output of \( \mathcal{I}_c(\hat{\theta}; \mathbf{r}) \) from the EM algorithm should be inverted to obtain the (asymptotic) variance-covariance matrix. Given the particular form that the components in the Louis’ identity assumes for CUB models, it is
possible to obtain a direct expression of the standard errors by applying some results from matrix algebra. Then, \( I^{-1}(\theta^{(k)}; r) \) can be promptly and directly computed also within the acceleration phase. In this respect, from both simulation and empirical evidence it turns out that some efforts should be devoted to the tuning of the parameter \( \text{iter}_c \).

**Acknowledgments:** Special Thanks to Cristina Pinto, Master of Science in Mathematics, for preliminary computation on the topics.

**References**


Mining signals of astronomical sources via Bayesian nonparametric mixture modelling

Andrea Sottosanti¹, Mauro Bernardi¹², Alessandra R. Brazzale¹

¹ University of Padua, Department of Statistical Sciences - Padua, Italy
² Istituto per le Applicazioni del Calcolo “Mauro Picone”, CNR - Roma, Italy

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

Abstract: We propose a new hierarchical Bayesian nonparametric model to discover and locate the signal of astronomical sources in a map of gamma-ray photons detected by the Fermi telescope. Our model simultaneously induces clustering on the photons and gives an estimate of the number of sources, while separating them from the irregular signal of the background contamination that extends over the entire map.

Keywords: Astrostatistics, Bayesian Nonparametrics, B-splines, Clustering.

1 Introduction

The detection of astronomical sources is an interdisciplinary field which includes both statistical and astronomical methods. This type of analysis starts from collections of photon counts, for which information about directions and energies is known, and aims to determine the number of sources in the map and to pool individual events into the corresponding clusters. In this work, we propose a new model to analyse the collection of gamma-ray photons detected by the LAT on-board the Fermi satellite, which is the telescope designed to record high energy particles. One of the main goals of the Fermi mission is to detect new gamma-ray sources, such as Active Galactic Nuclei, in the extra-galactic space, that is the region in the sky outside the band of the Milky Way. These sources are typically point-like, but there are also diffuse sources which make up the so-called isotropic diffuse gamma-ray background (IGRB) that is a class of various contaminating events, both isotropic and not, that can hide the signal of the sources.

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).
In literature, the most common approaches for source extraction are discussed in Section 7.5 of *Hobson et al. (2010)*: these are called single-source models, as they segment the map into subregions of fixed dimension for testing the presence of a new source in each. *Jones et al. (2015)* propose instead to explore simultaneously the entire map and infer on the directions of the sources using a Bayesian finite mixture model. *Sottosanti et al. (2017)* extend the latter to account for the background contamination of the Fermi LAT data. Here, we propose a novel approach based on a mixture of two Bayesian nonparametric models to discover and locate new gamma-ray sources and simultaneously account for the high background contamination of the data.

## 2 Bayesian nonparametric mixture modelling

We consider a collection of $n$ photons characterised by their galactic longitude and latitude, $x_i$ and $y_i$, defined on a plane $(x_{\text{min}}, x_{\text{max}}) \times (y_{\text{min}}, y_{\text{max}})$. We define a statistical model to study how photons are scattered on the map, to infer the directions of the sources. As a specific observation may either come from a specific source or can be due to some background phenomena, we classify each photon $i$ as signal or background using its spatial directions $x_i = (x_i, y_i)$ according to the mixture model

$$f(x_i|\Theta) = \delta s(x_i|\psi) + (1 - \delta)b(x_i|\varphi),$$

where $0 < \delta < 1$ and $s(\cdot|\cdot)$ and $b(\cdot|\cdot)$ are density functions that model the direction of a photon $i$ when it comes respectively from the sources and from the background. *Sottosanti et al. (2017)* suggests to use a parametric density for $b(\cdot|\cdot)$ that catches the shape of the analysed region of the Fermi LAT data, while $s(\cdot|\cdot)$ is taken to be a finite mixture of King’s PSFs, according to *Jones et al. (2015)*.

This modelling approach has two strong limitations. First, the finite mixture representation does not account for the presence of new clusters in case new data become available. Furthermore, the parametric formulation of the background model is not appropriate and may lead to the detection of fake clusters.

We adopt here a novel and more flexible approach based on Bayesian nonparametric methods and in particular we consider the Dirichlet Process (DP) mixture modelling approach. This class of methods allows to reconstruct a generic density function through an infinite mixture of base densities, inducing clustering on the observations. Formally, in our model we assume that

$$s(x_i|\psi) = \sum_{j=1}^{\infty} \pi^s_j K(x_i|\psi_j), \quad b(x_i|\varphi) = \sum_{j=1}^{\infty} \pi^b_j G(x_i|\varphi_j),$$

where $\pi^s_j$ and $\pi^b_j$ are the mixing coefficients.
where $K(\cdot|\cdot)$ and $G(\cdot|\cdot)$ are density functions and $\pi^s$ and $\pi^b$ are mixing weights, having the form $\pi_1 = V_1$, $\pi_j = V_1 \prod_{l=1}^{j-1}(1-V_l)$, $V_l \sim \text{Beta}(1, \alpha)$, $\forall l$, given by the stick-breaking representation of the Dirichlet Process (Sethuraman 1994). Model (1) simultaneously discovers and locates the sources in the map and accounts for the heavy and irregularly shaped background contamination which characterizes this kind of data. The weight $\delta$ is a mixing parameter which represents the amount of information attributable to the astronomical sources; we assume that, a priori, it is randomly distributed according to $\text{Beta}(\alpha, \alpha)$.

### 2.1 Modelling the signal of sources

We first define the Dirichlet Process mixture used to estimate $s(\cdot|\cdot)$, the distribution of the signals from sources in the map. From a modelling perspective, the Dirichlet Process prior allows to add a possible infinite number of clusters to the model; this property makes it suitable for discovering and locating an unknown number of astronomical sources in a map starting from the directions of the photons.

Let us consider a single photon $i$ with direction $x_i$ and suppose we know a priori the origin of $i$, that is a certain source $j$ with directions $\mu_j$. According to Ackermann et al. (2013), the direction of the photon $i$ distributes around its source $j$ as a bivariate random variable with density function

$$
    p(x_i; \mu_j, \tilde{\sigma}^2_i, \tilde{\nu}_i, \rho_i) = \rho_i t(x_i; \mu_j, \tilde{\sigma}^2_{i,\text{core}}I, \tilde{\nu}_{i,\text{core}})
    + (1 - \rho_i) t(x_i; \mu_j, \tilde{\sigma}^2_{i,\text{tail}}I, \tilde{\nu}_{i,\text{tail}}),
$$

where $t(\cdot; \gamma, \delta, \epsilon)$ is the density function of the bivariate Student $t$ distribution with non centrality vector $\gamma$, scale matrix $\delta$ and $\epsilon$ degrees of freedom, $I$ is a $2 \times 2$ identity matrix and $\tilde{\sigma}^2_i = (\tilde{\sigma}^2_{i,\text{core}}, \tilde{\sigma}^2_{i,\text{tail}})$, $\tilde{\nu}_i = (\tilde{\nu}_{i,\text{core}}, \tilde{\nu}_{i,\text{tail}})$ and $\rho_i$ are known quantities which describe the precision of the telescope in detecting the photon $i$. With respect to the notation of the DP mixture described in (1), the cluster variable $K(\cdot|\cdot)$ is here taken as (2), and so $\psi_j \equiv \mu_j$. Finally, since any source $j$ can be anywhere in the sky, we assume that a priori the location parameters $\mu_j$ are uniformly distributed over the analysed map.

### 2.2 Modelling the background

In the gamma-ray framework, even if the probabilistic structure of a source is well known, the background contamination is complex and irregular, since it is due to different phenomena. We recall here the definition of normalized B-spline basis function of order $m$ (De Boor 2001), defined on a set of $m+1$ knots that establish both its location and shape. Furthermore, since they are normalized, they can be used as density functions. These characteristics make them a suitable tool for modelling irregular data. We propose here a
novel approach to model the background based on Dirichlet Process mixtures of B-splines. More specifically, we estimate the background density function $b(\cdot|\cdot)$ using a Dirichlet Process mixture with cluster density $\mathcal{G}(\cdot|\cdot)$ equal to the product of two B-spline basis functions of degree 4, to model respectively the longitude and the latitude of the signals independently. We write then $\mathcal{G}(x_i|\varphi_j) = B^n_4(x_i|l_k)B^n_4(y_i|b_k)$, where $B^n_4(\cdot|\cdot)$ is a normalized basis function of order 4, and $l_k = (l_{k1}, \ldots, l_{k5})$ and $b_k = (b_{k1}, \ldots, b_{k5})$ are the unknown knots for the longitude and latitude. We complete the Bayesian specification of the model defining the prior distribution for the set of unknown parameters $\varphi_k = (l_k, b_k)$. For the longitude $l_k$, we assume

$$l_{k1} \sim \mathcal{U}(x_{\min}, x_{\max}), \quad l_{ki}|l_{k(i-1)} \sim \mathcal{U}(l_{k(i-1)}, x_{\max}), \quad i = 2, \ldots, 5,$$

while for $b_k$ we take the same prior structure, just substituting $(x_{\min}, x_{\max})$ with $(y_{\min}, y_{\max})$.

### 3 Algorithm

We propose an efficient MCMC algorithm to simulate from the posterior distribution of the model parameters. By recalling here the truncated stick-breaking representation of the DP proposed by Ishwaran and Zarepour (2000), which assumes that the number of components in the mixture is upper bounded, we build a computationally fast algorithm for posterior inference which does not require any tuning parameter. We define as $\mathcal{S}^{(t)}$ the set of photons assigned to the source during iteration $t$ and $\mathcal{B}^{(t)}$ the ones pulled into the background. We let the two processes admit at most $k_s$ and $k_b$ components, both set equal to 100. The algorithm iteratively:

1. pulls observations into $\mathcal{S}^{(t)}$ and $\mathcal{B}^{(t)}$ and updates $\delta^{(t)}|\cdots \sim \text{Beta}(\alpha + |\mathcal{S}^{(t)}|, \alpha + |\mathcal{B}^{(t)}|)$;

2. a) splits the observations assigned to $\mathcal{S}^{(t)}$ into $k_s$ clusters and updates the source model parameters, and b) splits the observations assigned to $\mathcal{B}^{(t)}$ into $k_b$ clusters and updates the background model parameters.

New values from the posterior distributions of the weights $\pi^s$ and $\pi^b$ can be easily obtained using the simulation scheme presented in Ishwaran and James (2001). A closed Gibbs-sampling scheme for the location parameters $\mu_j$ can be set up through a double data-augmentation step on the density function (2): the first introduces a latent variable to avoid the mixture representation, and the second rewrites the multivariate Student t distribution as the product of a bivariate Gaussian and a gamma distribution.

Regarding the set of parameters of the background model $\{l_j, b_j\}_{j=1}^{k_b}$, a closed form of the full-conditional distribution is not easily identifiable; however, we can update each knot once at a time conditioning on the values
Bayesian mining of astronomical sources

FIGURE 1. Posterior predictive distribution of the background model (left) and of the source model (right).

<table>
<thead>
<tr>
<th>cluster</th>
<th>$\mu_{jx}$</th>
<th>$\mu_{jy}$</th>
<th>$\pi_j^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-108.504</td>
<td>52.814</td>
<td>0.447 (0.351,0.54)</td>
</tr>
<tr>
<td>2</td>
<td>-95.800</td>
<td>57.374</td>
<td>0.231 (0.158,0.312)</td>
</tr>
<tr>
<td>3</td>
<td>-115.386</td>
<td>70.211</td>
<td>0.122 (0.066,0.184)</td>
</tr>
<tr>
<td>4</td>
<td>-119.199</td>
<td>52.635</td>
<td>0.139 (0.075,0.207)</td>
</tr>
<tr>
<td>5</td>
<td>-98.347</td>
<td>76.734</td>
<td>0.050 (0.001,0.095)</td>
</tr>
<tr>
<td></td>
<td>-113.962</td>
<td>75.929</td>
<td></td>
</tr>
</tbody>
</table>

TABLE 1. Posterior modes and 95% HPD intervals of the parameters ($\mu_1, \ldots, \mu_5$) of the source model, together with the relative intensity of the first five components. The last two rows show that the posterior distribution of $\mu_5$ is bimodal.

assumed by the others: $\forall j = 1, \ldots, k_b$ and $i = 1, \ldots, 5$ refer to

$$b_{ji}^{(t)} | \cdots \sim \pi(b_{ji} \mid b_{j,-i}, B^{(t)}), \quad b_{ji}^{(t)} | \cdots \sim \pi(b_{ji} \mid b_{j,-i}, B^{(t)}).$$

The support of the distributions defined by (3) is closed; this allows us to build a simulation algorithm based on rejection sampling with uniform distributions in the role of proposals.

4 Results

We apply our model to a portion of the sky covered by the Fermi LAT. In particular, we consider the square of coordinates $(-120, -90) \times (50, 80)$ in the extra-galactic space. The background contamination is here more concentrated on the lower part due to the proximity to the galactic plane and so it can not be assumed to be uniform.

We first set $\alpha$, $\alpha_s$ and $\alpha_b$, respectively the parameter of the Beta prior over $\delta$ and the concentration parameters of the DPs, equal to 1.

We run the simulation algorithm for 20,000 iterations and we use the first 15,000 as burn-in period. The intensity of the source component is 0.077
and its 95% credibility interval is (0.061, 0.091). Table 1 shows the posterior estimates of the parameters in the mixture model for the sources. Clusters 1, 2 and 4 agree with already known sources in this region of the sky; in addition, Cluster 3 embeds the signal of two overlapping sources. The intensity of the fifth component is around 5% and the posterior distribution of $\mu_5$ shows two modal values, as we can see from the last two rows in Table 1. Both of them do not correspond to already known sources, and for this reason their nature should be investigated in more depth.

Figure 1, on the left, shows the reconstruction of the background component using the proposed B-spline DP mixture model, while on the right we see the posterior predictive distribution for the sources. As expected, the intensity of the noise is much higher on the lower part of the map. We conclude that our model succeeds in separating the sources from the background and allows us to detect both already discovered sources and new clusters.

References


Correcting Under-Recording of Eruptions in Historical Volcano Data

Oliver Stoner

1 University of Exeter, United Kingdom

E-mail for correspondence: ors203@exeter.ac.uk

Abstract: Historical volcano data can suffer from under-recording of eruption occurrence, which can vary with time and magnitude. A Bayesian hierarchical framework is employed, to model simultaneously the true eruption rate and the under-recording mechanism, in order to obtain a more reliable inference on the relationship between eruption magnitude and frequency.

Keywords: Hierarchical; Bayesian; Natural Hazards.

1 Introduction

The LaMEVE dataset is a record of historic eruptions, with each entry including an estimated eruption year and an estimate of the magnitude of the eruption. As in Rougier et al., this is defined by:

\[ \text{magnitude} = \log_{10}(\text{erupted mass in kg}) - 7 \]

Unfortunately, the recording of volcano eruptions is not complete; some entries rely on historical records, while many rely on geological analyses, where the likelihood of an eruption leaving a discoverable trace depends on the location, time and magnitude of the eruption (Rougier et al.). This means that any inference on the temporal profile of eruption rates which assumes complete recording is likely biased. It is therefore desirable to quantify the under-recording, such that the frequency of eruptions can be more reliably investigated.

2 Methodology

A framework for correcting under-recording in count data has been recently developed and presented in Stoner et al.. In this framework, the recorded
count $z$ is modelled as a Binomial quantity, where the number of trials is an unobserved Poisson quantity $y$, representing the true value of the count that has been incompletely recorded. The true count generating process is then modelled through the mean of the Poisson quantity $\lambda$, while the under-recording mechanism can be modelled through the Binomial probability $\pi$, analogous to the recording probability, to mitigate the bias involved in inference on both processes. The basic structure of the model is hence given by:

$$z \mid y \sim \text{Binomial}(\pi, y); \quad y \sim \text{Poisson}(\lambda) \quad (1)$$

In order to use this framework, the data must be aggregated into counts of eruptions over a chosen time interval and to achieve this the data were aggregated over 1000 intervals of 100 years.

![FIGURE 1. Total eruption counts for each of the last 1000 centuries, separated by magnitude.](image)

As discussed in Rougier et al., a large portion of eruption magnitude estimates have been rounded to the nearest integer, which could introduce issues if eruption magnitude is treated as a continuous variable, such as potential overestimation of the rate of near-integer value magnitudes and underestimation elsewhere. To overcome this problem, we follow Rougier et al. by classifying the data into four bins based on magnitude: Low $[4.5,5.5)$, Medium $[5.5,6.5)$, High $[6.5,7.5)$ and Very High $[7.5,8.5)$. The resulting set of eruption counts can be seen in Figure 1. Ignoring under-recording, the data appear to suggest the rate of eruptions has been dramatically increasing in recent centuries.
Correcting Under-Recording of Eruptions

For an eruption in century $t \in T = 1, 2, ..., 1000$, where $T = 0$ represents the 21st century, and of magnitude $m \in M = \{\text{Low, Medium, High, Very High}\}$ the model is structured as follows:

\[ z_{t,m} | y_{t,m} \sim \text{Binomial}(\pi_{t,m}, y_{t,m}) \]  \hspace{1cm} \text{(2)}

\[ \log \left( \frac{\pi_{t,m}}{1 - \pi_{t,m}} \right) = \beta_{0,m} + \sum_{k=1}^{3} \beta_{k,m} w_{t,m}^{k} \]  \hspace{1cm} \text{(3)}

\[ y_{t,m} \sim \text{Poisson}(\lambda_{m}) \]  \hspace{1cm} \text{(4)}

\[ \log(\lambda_{m}) = \alpha_{0} + \alpha_{1} x_{m} + \epsilon_{m} \]  \hspace{1cm} \text{(5)}

\[ \epsilon_{m} \sim \text{Normal}(0, \sigma^{2}) \]  \hspace{1cm} \text{(6)}

Here $w_{t,m}$ is the transformed century $t$ ($w_{t,m} = \log(t + 1)$ for $m = \text{Low}$, $w_{t,m} = t/1000$ otherwise), and $x_{m}$ is defined by the midpoint of the magnitude bin, minus the mean of the midpoints. The change in the recording probability $\pi_{t,m}$ is characterised by a different cubic polynomial for each magnitude bin in (3). A log-linear relationship between the eruption rate and magnitude is introduced in (5), such that information is pooled from the different bins into parameters $\alpha_{0}$ and $\alpha_{1}$. This is intended to aid in estimating the rate of Very High eruptions, of which there are very few observations. Additional flexibility is introduced by allowing the eruption rate for each bin to deviate from this line according to a Normal distribution, to allow for potential biases in the estimation of eruption magnitude.

A key concept in Stoner et al. is that the presence of under-recording means the information provided by the data is only partial, and must be supplemented to ensure parameter identifiability between $\lambda_{m}$ and $\pi_{t,m}$. The simplest way to achieve this is to specify an informative prior distribution for $\beta_{0,m}$. Noting that the linear predictor of $\pi_{t,m}$ reduces to $\beta_{0,m}$ when $t = 0$, a Normal distribution with mean 4 and precision 2 was specified for each $\beta_{0,m}$, to represent a hypothetical belief that present day recording of eruptions is near complete. Finally, the number of non-zero observations for the Very High bin was determined to be too low to provide any meaningful inference on its change in recording probability over time, so the recording probability was fixed at 1 for this bin.

3 Results

The estimated change in the recording probabilities, for the first three magnitudinal bins, can be seen in Figure 2. All three curves show near monotonic decreasing trends going backwards in time, with a pattern of increasing recording probability for higher magnitudes generally holding. Figure 3 shows the posterior mean estimates of the eruption rate for each magnitudinal bin, with associated 95% credible intervals. The solid black line represents the median predicted relationship between magnitude and rate, as defined in (5).
FIGURE 2. Posterior median estimates of the effect of time on the probability of recording a volcano eruption, for the first three magnitude bins, with associated 95% credible intervals.

FIGURE 3. Plot showing the estimated eruption rate (on the log-scale) for the different magnitudinal bins, with associated 95% credible intervals. The solid black line shows the posterior median estimate of $\alpha_0 + \alpha_1 x_m$, with associated 95% credible interval.

Finally, the return period $R$ for an eruption in a given magnitude classification $m$ can be calculated as:

$$ R_m = \frac{1}{\lambda_m} $$

In Table 1, it can be noted that the return period estimates for a Very High eruption is similar to the estimate in Rougier et al. for the return period of a volcano exceeding magnitude 8, which has a median of 17000 years, with associated 95% credible interval (5200, 48000), though it is not yet clear if the two quantities are directly comparable.
TABLE 1. Return period (years) approximate predictive quantiles for an eruption in each of the four magnitudinal bins.

<table>
<thead>
<tr>
<th>Magnitude</th>
<th>Lower 95%</th>
<th>Median</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>4.4</td>
<td>6.2</td>
<td>8.3</td>
</tr>
<tr>
<td>Medium</td>
<td>77</td>
<td>110</td>
<td>150</td>
</tr>
<tr>
<td>High</td>
<td>360</td>
<td>530</td>
<td>730</td>
</tr>
<tr>
<td>Very High</td>
<td>9100</td>
<td>19000</td>
<td>50000</td>
</tr>
</tbody>
</table>

4 Summary and Conclusions

In this article the challenges posed by under-recording were explored in the context of historic volcano eruptions. A dataset of volcanic eruptions was aggregated both by century and into four bins, based on eruption magnitude, which resulted in a dataset of counts. A general framework for correcting under-recording in count data was borrowed from Stoner et al., the flexibility and generality of which meant that little adaption was necessary to design and implement an appropriate model for this problem. By accounting for the relationship between eruption magnitude and time and the under-recording mechanism, a more reliable inference for the relationship between eruption magnitude and rate of occurrence was made possible. This inference relies on the assumption that the rate of eruptions was constant in time over the period analysed, and the results could also be sensitive to the way in which the eruptions were aggregated both into centuries and into four magnitudinal bins. Both of these sensitivities are worthy of future investigation, though it can be noted that these results were similar to those in Rougier et al., a study of the same data set with a substantially different approach.

Acknowledgments: Special Thanks to Professor Jonathan Rougier for introducing the author to the dataset and his existing work on the subject.

References


An Augmented Likelihood Ratio Index for Categorical Response Models

Ejike R. Ugba\textsuperscript{1}, Jan Gertheiss\textsuperscript{1}

\textsuperscript{1} Institute of Applied Stochastics and Operations Research, Clausthal University of Technology, Germany

E-mail for correspondence: ejike.richard.ugba@tu-clausthal.de

Abstract: Several studies on measures of predictive power of categorical response models pay more attention to binary outcome models than ordinal models. We propose a response-adjusted likelihood ratio index that takes cognizance of the effect of increasing number of response categories on the likelihood ratio index and that also applies to both binary and polytomous outcome models estimated via maximum likelihood. The proposed measure is considered an augmentation and a better alternative to McFadden’s measure. It also competes favorably with McKelvey & Zavoina’s measure known for its approximation of the $R^2$ from an underlying continuous model. The results from both the simulation and the real-life data examples attest to these.

Keywords: Goodness-of-fit; Latent variable; Likelihood ratio; Ordinal model; Pseudo-$R^2$.

1 Introduction

Whereas the coefficient of determination ($R^2_{\text{ols}}$) provides an intuitive and a well defined summary measure of a model’s predictive power in ordinary linear regression, there is no exact replica of such for most discrete limited dependent variable models. In this study, we intend to work out a very compact generalization of the likelihood-based measures from the literature and then propose a simple but very useful index from the same generalization. The proposed measure could be partly considered an augmentation of the likelihood ratio index proposed in McFadden (1974). Given an ordinal response variable $y_i$, and $x_i$ a vector of covariates, a popular modeling approach is to consider a continuous underlying latent variable $y_i^* = x_i^T \beta + \epsilon_i$, $i = 1, 2, \ldots, n$, with $\beta$ a vector of regression parameters and $\epsilon_i$ an error...
term. Suppose \( \tau_0 < \tau_1 < \ldots < \tau_k \) are cut-points on the continuous scale of \( y_i^* \) such that the observed response \( y_i \) satisfies the following threshold model:
\[
y_i = j \Leftrightarrow \tau_{j-1} < y_i^* < \tau_j, \quad \text{where } j = 1, 2, \ldots, k \text{ and } \tau_0 = -\infty, \tau_k = \infty.
\]
For error \( \epsilon_i \), typically a normal or logistic distribution is assumed, leading to a probit or logit model, respectively; see, e.g., Agresti (2002) for details. As a consequence of the latent variable approach, a common criterion when assessing the goodness-of-fit of the corresponding ordinal model is that the pseudo-\( R^2 \) measuring its predictive power be as close as possible to the \( R^2_{\text{ols}} \) of the underlying latent variable model (Hagle & Mitchell, 1992). Several studies on the goodness-of-fit measures for categorical models seem to favor McFadden’s \( R^2_{\text{mf}} \) on account of its intuitive interpretation and base rate stability in binary models (Menard, 2000). Notwithstanding, \( R^2_{\text{mf}} \) is known to have some very remarkable drawbacks. Apart from underestimating the \( R^2_{\text{ols}} \), its value depreciates monotonically with increasing number of categories of ordinal models. These would necessitate an augmentation of the likelihood ratio index.

2 The Generalized Likelihood Ratio Index

Given that the individual formulations of several likelihood-based goodness-of-fit measures are primarily anchored on the null and full model log-likelihoods and also considering that their respective outcomes for models under inspection largely depend on the information furnished by these same likelihoods, it would be appropriate to first provide a generalization of these measures and thereby study their relative performance. Hence, assuming the following simple generalized likelihood ratio index (GLRI):
\[
GLRI = 1/\phi \left\{ 1 - \frac{l(M_\beta)}{l(M_\alpha)} \right\}^\gamma,
\]
where \( l(M_\alpha) \) and \( l(M_\beta) \) are the fitted and null model log-likelihoods respectively, \( \phi \geq 1 \) being a factor potentially depending on the two log-likelihoods and \( \gamma \) a fixed positive integer. The GLRI attains a maximum as \( l(M_\beta) \to 0 \) and \( \phi \to 1 \); it attains a minimum when \( l(M_\alpha) = l(M_\beta) \). A well defined class of likelihood ratio measures could be derived from (1). Of course, with \( \phi = 1 \) and \( \gamma = 1 \), the GLRI collapses to the simplest case, being the McFadden’s measure \( R^2_{\text{mf}} \). To adjust for the effect of increasing number of response categories on the likelihood ratio measure, we propose a specification of the GLRI with \( \phi = 1 \) and \( \gamma = \sqrt{2k} \), where \( k = 2, 3, \ldots \) is the number of response categories. This automatically yields a sort of response-synchronizing likelihood ratio index. For instance, with \( k = 2 \) (i.e., the binary case), the following squared likelihood ratio index for assessing the fit of binary models is obtained:
\[
R^2_{k=2} = 1 - \left\{ \frac{l(M_\beta)}{l(M_\alpha)} \right\}^2.
\]
Suppose one writes $\psi(M^+ + \beta) = \{l(M^\alpha) + l(M^\beta)\}/l(M^\alpha)$, one has $R^2_{k=2} = \psi(M^+ + \beta)R^2_{mf}$, so that, $0 < R^2_{mf} < 1$, $1 < \psi(M^+ + \beta) < 2$ and $0 < R^2_{k=2} < 1$. In general, for $k \geq 2$, the following applies,

$$R^2_{k \geq 2} = 1 - \left\{l(M^\beta)/l(M^\alpha)\right\}^{\sqrt{2k}},$$

with $0 < R^2_{k \geq 2} < 1$ and the factor $\sqrt{2k}$ adjusting for the effect of increasing number of categories. This of course makes sense going by the invariance property of cumulative models motivated by a latent variable. The same parameters occur for the effect regardless of how the cut-points discretize the continuous scale (Agresti, 2002). Thus, an adequate summary measure of predictive power ought therefore to yield practically the same value for models with different choice of response categories. To keep things simple, we write $R^2_{ug}$ instead of $R^2_{k \geq 2}$ for the proposed Ugba & Gertheiss measure and go on to compare the same with similar measures in a simulation and real data study.

3 Simulation Study & Real Data Analysis

The performance of the proposed measure is tasted out in a simulation study and with the wine data set adopted from Randall (1989). Hence, given $y^*_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ where $x_{i1} \sim U(0, 1), x_{i2} \sim N(0, 1)$, $\epsilon_i \sim N(0, 1)$, $\alpha = 0$ and $\beta_1 = \beta_2 = 1$. A thousand replications of $y^*$ having all variables and parameters fixed except $\epsilon_i$ are obtained for $n = 500$ observations. Different sets of probit models are also fitted for $k \geq 2$ categories of $y^*_i$. The $R^2_{ug}$ together with other similar measures are calculated and compared with the latent measure ($R^2_{ols}$). As shown in Figure 1, $R^2_{ug}$ provides a good approximation of ($R^2_{ols}$) as well as competes favorably with the McKelvey & Zavoina (1976) measure in binary models. Moreover, the performance of the obtained measures are also compared against increasing number of response categories of an ordered probit model (see Figure 2). Apparently, all the measures except the $R^2_{mf}$ tend towards the underlying measure ($R^2_{ols}$) as $k$ increases and apart from being the closest to the underlying measure, the proposed measure is very much stable across all levels of response category.

With the wine data set, the variable ‘response’ is the scorings of wine bitterness on a 0–100 (quasi) continuous scale, while ‘rating’ is an ordered factor with 5 levels: a grouped version of response. Both the continuous and ordered models are obtained using as predictors the two treatment factors, ‘temperature’ and ‘contact’. Different $R^2$’s from these models are obtained and compared in Table 1 against the latent $R^2$ of the continuous model. Again, in a sharp contrast to the other measures and $R^2_{mf}$ in particular, the augmented $R^2_{ug}$ comes closest to the latent $R^2_{ols}$ and also relatively stable across categories of an ordered model.
FIGURE 1. The MSE of $R^2$s from the simulation study. The used initials are coiled from the likelihood measures in Table 1, with the last two non-likelihood measures being the McKelvey & Zavoina (1976) and the Tjur (2009) measures.

FIGURE 2. The proximity of likelihood-based goodness-of-fit measures to the underlying measure ($R^2_{\text{ols}}$) and stability across different numbers of response levels.

4 Discussion

Whether or not to use the R-squared measure for discrete limited dependent variable models has been an issue of debate in the literature for decades. For the school of thought in favor of, it is also the debate on which of the
measures abundantly proposed in the literature ‘best’ captures the predictive strength of fitted categorical models. Popular opinion supports McFadden’s measure on the account of its simple formulation, base rate stability in binary models as well as its intuitive interpretation as the proportional reduction in the log-likelihood statistics of fitted models. However, as also observed in this study, the $R^2_{mf}$ has some very remarkable drawbacks that makes it inappropriate for categorical models motivated by an underlying latent model. As shown in Figure 2, apart from underestimating the underlying measure, $R^2_{mf}$ seems to suggest that the larger the number of response categories the poorer the model, which certainly is not the case as reflected by the other measures which all appreciate towards the underlying measure. The $R^2_{ug}$ among all the measures considered provides the best alternative to the $R^2_{mf}$, as it redresses the key limitations of the $R^2_{mf}$ and also shares similar interpretation. Furthermore, the $R^2_{ug}$ also approximates the Mckelvey & Zavoina’s measure, which is known to approximate the underlying measure pretty close in many situations, although it is rarely being used due to its complexity in computation and interpretation. For an easy computation of the $R^2_{ug}$ and many more measures of predictive strength of categorical models, the CatgoF package available at https://github.com/ejikeugba/CatgoF proves useful.

References


TABLE 1. R-squared measures of ordered models from the bitterness of wine data, with the latent R-squared highlighted below. The first two columns refer to collapsed versions of the original 5-level response

<table>
<thead>
<tr>
<th>Measure</th>
<th>Number of Response Levels</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox &amp; Snell (1989)</td>
<td>0.362</td>
<td>0.388</td>
<td>0.393</td>
<td></td>
</tr>
<tr>
<td>Nagelkerke (1991)</td>
<td>0.408</td>
<td>0.422</td>
<td>0.417</td>
<td></td>
</tr>
<tr>
<td>McFadden (1974)</td>
<td>0.207</td>
<td>0.194</td>
<td>0.174</td>
<td></td>
</tr>
<tr>
<td>Ugba &amp; Gertheiss</td>
<td>0.433</td>
<td>0.457</td>
<td>0.453</td>
<td></td>
</tr>
</tbody>
</table>

*Latent $R^2 = 0.461$


## Author Index

<table>
<thead>
<tr>
<th>Author</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abegaz, F</td>
<td>40</td>
</tr>
<tr>
<td>Adam, T</td>
<td>30</td>
</tr>
<tr>
<td>Al-tawarah, Y</td>
<td>199</td>
</tr>
<tr>
<td>Aston, J</td>
<td>29</td>
</tr>
<tr>
<td>Augustin, NH</td>
<td>69</td>
</tr>
<tr>
<td>Barnes, C</td>
<td>36</td>
</tr>
<tr>
<td>Behrouzi, P</td>
<td>40</td>
</tr>
<tr>
<td>Bellio, R</td>
<td>63</td>
</tr>
<tr>
<td>Berger, M</td>
<td>44</td>
</tr>
<tr>
<td>Bernardi, M</td>
<td>50, 282</td>
</tr>
<tr>
<td>Bertolacci, M</td>
<td>247</td>
</tr>
<tr>
<td>Blagojevic, M</td>
<td>199</td>
</tr>
<tr>
<td>Brady, A</td>
<td>54</td>
</tr>
<tr>
<td>Brazzale, AR</td>
<td>282</td>
</tr>
<tr>
<td>Bremhorst, V</td>
<td>163</td>
</tr>
<tr>
<td>Brewer, MJ</td>
<td>189</td>
</tr>
<tr>
<td>Burke, K</td>
<td>199, 237</td>
</tr>
<tr>
<td>Carlan, M</td>
<td>157</td>
</tr>
<tr>
<td>Cazzaro, M</td>
<td>90</td>
</tr>
<tr>
<td>Chandler, RE</td>
<td>36</td>
</tr>
<tr>
<td>Colombi, R</td>
<td>59</td>
</tr>
<tr>
<td>Costain, DA</td>
<td>96</td>
</tr>
<tr>
<td>Cripps, E</td>
<td>247</td>
</tr>
<tr>
<td>Cripps, S</td>
<td>247</td>
</tr>
<tr>
<td>Currie, I</td>
<td>17</td>
</tr>
<tr>
<td>Dan, N</td>
<td>63</td>
</tr>
<tr>
<td>Deutscher, C</td>
<td>231</td>
</tr>
<tr>
<td>Dupont, E</td>
<td>69</td>
</tr>
<tr>
<td>Durbán, M</td>
<td>174</td>
</tr>
<tr>
<td>Ebrahimpoor, M</td>
<td>73</td>
</tr>
<tr>
<td>Eilers, PHC</td>
<td>78</td>
</tr>
<tr>
<td>Elayouty, A</td>
<td>84</td>
</tr>
<tr>
<td>Emsley, R</td>
<td>134</td>
</tr>
<tr>
<td>Faraway, J</td>
<td>54</td>
</tr>
<tr>
<td>France, A</td>
<td>96</td>
</tr>
<tr>
<td>Friedl, H</td>
<td>151</td>
</tr>
<tr>
<td>Gasparrini, A</td>
<td>265</td>
</tr>
<tr>
<td>Geraci, M</td>
<td>102</td>
</tr>
<tr>
<td>Gertheiss, J</td>
<td>219, 293</td>
</tr>
<tr>
<td>Giordano, S</td>
<td>59</td>
</tr>
<tr>
<td>Goeman, J</td>
<td>73</td>
</tr>
<tr>
<td>Gong, M</td>
<td>106</td>
</tr>
<tr>
<td>Gregory, SD</td>
<td>112</td>
</tr>
<tr>
<td>Grilli, L</td>
<td>118</td>
</tr>
<tr>
<td>Groll, A</td>
<td>122</td>
</tr>
<tr>
<td>Grzegorczyk, M</td>
<td>270</td>
</tr>
<tr>
<td>Hasso, S</td>
<td>209</td>
</tr>
<tr>
<td>Hastie, T</td>
<td>122</td>
</tr>
<tr>
<td>Held, L</td>
<td>8</td>
</tr>
<tr>
<td>Heller, G</td>
<td>128</td>
</tr>
<tr>
<td>Hellier, J</td>
<td>134</td>
</tr>
<tr>
<td>Hohberg, M</td>
<td>139</td>
</tr>
<tr>
<td>Hu, S</td>
<td>145</td>
</tr>
<tr>
<td>Hu, W</td>
<td>128</td>
</tr>
<tr>
<td>Karasek, S</td>
<td>151</td>
</tr>
<tr>
<td>Kauermann, G</td>
<td>168</td>
</tr>
<tr>
<td>King, R</td>
<td>242</td>
</tr>
<tr>
<td>Klein, N</td>
<td>157</td>
</tr>
<tr>
<td>Kneib, T</td>
<td>30, 122, 139, 157, 253</td>
</tr>
<tr>
<td>Lambert, P</td>
<td>163</td>
</tr>
<tr>
<td>Lang, S</td>
<td>157</td>
</tr>
<tr>
<td>Langrock, R</td>
<td>30, 231, 242</td>
</tr>
<tr>
<td>Lark, RM</td>
<td>215</td>
</tr>
<tr>
<td>Lebacher, M</td>
<td>168</td>
</tr>
<tr>
<td>Lee, D</td>
<td>179</td>
</tr>
<tr>
<td>Lee, D-J</td>
<td>174</td>
</tr>
<tr>
<td>Leitner, M</td>
<td>183</td>
</tr>
<tr>
<td>Leos-Barajas, V</td>
<td>231</td>
</tr>
<tr>
<td>Lorenzo-Arribas, A</td>
<td>189</td>
</tr>
<tr>
<td>Möller, A</td>
<td>219</td>
</tr>
</tbody>
</table>
Machado, RJM, 195
MacKenzie, G, 199, 237
Maestrini, L, 204
Mair, P, 259
Marino, MF, 118
Marra, G, 195
Martínez-Araya, M, 215
Matawie, KM, 209
Matias, C, 7
Mayr, A, 30
Miller, C, 84, 106
Murphy, TB, 145

Núñez-Antón, V, 225
Nicolussi, F, 90

O’Hagan, A, 145
Oguiza, A, 225
Ötting, M, 231
Overstall, AM, 189

Pütz, P, 139
Paccagnella, O, 118
Peng, D, 237
Pickles, A, 134
Pohle, J, 242
Prosdocimi, I, 54, 69
Putter, H, 1

Rampichini, C, 118
Reutterer, T, 63
Rosen, O, 247

Santos, B, 253
Schauberger, G, 259
Scheibelhofer, P, 151
Schmid, M, 44
Scott, M, 84, 106
Sera, F, 265
Shafiee Kamalabad, M, 270
Simone, R, 276
Simpson, D, 69
Sottosanti, A, 282
Spitali, P, 73
Spitoni, C, 1
Stolfi, P, 50
Stoner, O, 288
Swanson, B, 128

Thurner, PW, 168
Titman, A, 1
Titman, AC, 96
Tsonaka, R, 73
Tutz, G, 122

Ugba, ER, 293

van den Hout, A, 195
van der Schaar, M, 242
Virto, J, 225

Wagner, H, 157, 183
Wagner, M, 44
Waldron, S, 84
Wand, MP, 204
Wit, EC, 40