Proceedings of the

29th International Workshop on Statistical Modelling

Volume 1

July 14 – 18, 2014

Göttingen, Germany

Thomas Kneib, Fabian Sobotka,
Jan Fahrenholz, Henriette Irmer

(editors)

Proceedings of the 29th International Workshop on Statistical Modelling,
Volume 1,
Göttingen, July 14 – 18, 2014,
Thomas Kneib, Fabian Sobotka, Jan Fahrenholz, Henriette Irmer
(editors),
Göttingen, 2014.

Editors:
Thomas Kneib, tkneib@uni-goettingen.de
Fabian Sobotka, fabian.sobotka@wiwi.uni-goettingen.de
Jan Fahrenholz, jfahren@uni-goettingen.de
Henriette Irmer, iwsm2014@uni-goettingen.de

Centre for Statistics
Georg-August-University
Platz der Göttinger Sieben 5
37073, Göttingen, Germany

Printed by:
Druckerei Brüggemann GmbH
Violenstraße 23
28195 Bremen
Germany

Scientific Programme Committee

- Thomas Kneib (Chair)
  Georg-August-University Göttingen, Germany
- Jim Booth
  Cornell University, Ithaca, USA
- Maria Durban
  University Carlos III of Madrid, Spain
- Gillan Heller
  Macquarie University, Sydney, Australia
- Arnošt Komárek
  Charles University in Prague, Czech Republic
- Stefan Lang
  Universität Innsbruck, Austria
- Vito Muggeo
  University of Palermo, Italy
- Mikis Stasinopoulos
  London Metropolitan University, UK
- Lola Ugarte
  Universidad Pública de Navarra, Spain
- Florin Vaida
  University of California, San Diego, USA
- Helga Wagner
  Johannes Kepler Universität Linz, Austria

Local Organizing Committee

- Thomas Kneib (Chair)
- Heike Bickeböller
- Jan Fahrenholz
- Jan Gertheiss
- Henriette Irmer
- Nadja Klein
- Tatyana Krivobokova
- Juliane Manitz
- Julia Meskauskas
- Patrick Michaelis
- Oleg Nenadić
- Hauke Rennies
- Holger Reulen
- Benjamin Säfken
- Fabian Sobotka
- Alexander Sohn
- Elmar Spiegel
- Elisabeth Waldmann

Statistics is coming home

Dear Participants,

we welcome you to the 29th International Workshop on Statistical Modelling (IWSM) taking place in Göttingen, being both in the very heart of Germany and the hometown of statistics. The 29th IWSM will coincidentally be held in the 29th week of 2014 and – although you can follow traces of Carl Friedrich Gauß at various places in Göttingen – not every part of the workshop will be Gaussian since this is not a standard normal conference.

The well-established key features of the workshop (no parallel sessions, a limited number of distinguished invited talks, a focus on interdisciplinarity and young contributing authors, a short course preceeding the conference, a supplementing social program) can all be found in this year's program again. These elements keep the IWSM unique in the landscape of statistical conferences. 145 participants from all over the world will have the chance to attend the 56 contributed oral presentations and take their time in the separate extended session to stroll around the 45 poster presentations. All abstracts are also provided as a PDF from the conference website. We are glad that Antoine de Falguerolles from Toulouse, Alejandro Jara from Santiago de Chile, Sophia Rabe-Hesketh from Berkeley, Gerhard Tutz from Munich and Simon Wood from Bath have accepted the invitation to give a one hour presentation in order to inform and entertain us more extensively. For the short course, Benjamin Hofner and Andreas Mayr from Erlangen provided an introduction to "Boosting for statistical modelling" to 25 participants.

Still, this is only the setup. The high standards of the conference and the quality of all presentations – oral and poster – is ensured by the scientific committee devoting a considerable amount of work to the review process of submitted abstracts. They were also able to gather an international group of renowned experts to give the keynote presentations. Thanks to this hard work, the 29th week of 2014 will hopefully be, once again, very fruitful for ongoing research, potential collaborations and statistical modelling itself.

However, the IWSM does not only bring together people from different countries, but also from different ages and experience levels. As per tradition, this is encouraged by awarding the best student paper, the best student oral presentation and the best student poster at the conference dinner Thursday night. Furthermore, two student travel grants have been kindly provided by the Statistical Modelling Society.

Finally, we thank all authors who contributed to these proceedings for participating in the workshop and for carefully preparing their manuscripts and talks or posters. We hope that you can find inspiration in the home of statistics and also enjoy the historic city, the university and the historic observatory of Gauß himself at the welcome reception on Monday night.

Thomas Kneib, Fabian Sobotka, Jan Fahrenholz and Henriette Irmer
on behalf of the local organizing committee
Göttingen, May 2014

Contents

Part I - Invited Papers

Antoine de Falguerolles: "La statistique à Goettingue": a tentative tribute by an IWSM participant from Toulouse 3

Alejandro Jara, Andrés F. Barrientos: Bayesian nonparametric approaches for the analysis of compositional data based on Bernstein polynomials 15

Sophia Rabe-Hesketh, Anders Skrondal : Consistent estimation of mixed models by ignoring non-ignorable missingness 27

Gerhard Tutz: Regularization and Sparsity in Discrete Structures 29

Simon N. Wood: General smooth additive modelling 43

Part II - Contributed Papers

Haakon Bakka: She'll be coming 'round the mountain: Simple models of complex spatial behaviour 51

Andreas Bender, Fabian Scheipl, Wolfgang Hartl, Helmut Küchenhoff: Modeling Exposure-lag-response associations in Survival Models with application to nutrition of critically ill patients ... 57

Sarah Brockhaus, Fabian Scheipl, Torsten Hothorn, Sonja Greven: The Functional Linear Array Model and an Application to Viscosity Curves 63

Carlo G. Camarda: Reconstructing Mortality Series by Cause of Death: Two alternative approaches 69

Jona Cederbaum, Sonja Greven, Marianne Pouplier, Phil Hoole: Functional linear mixed model for irregularly spaced phonetics data 75

Sammy Chebon, Helena Geys, Ann De Smedt, Christel Faes: A multivariate random component model for simultaneously interval censored and right truncated data. 81

Iain D. Currie: Smooth mixed models for balanced longitudinal data 87

Fernanda De Bastiani, Mikis Stasinopoulos, Robert Rigby, Audrey H.M.A. Cysneiros: Spatial Modelling using GAMLSS ... 93

Vera Djordjilović, Monica Chiogna, Maria Sofia Massa, Chiara Romualdi: Refining the structure of a pathway with a view to prediction of gene silencing effects 99

Mark W. Donoghoe', Ian C. Marschner': Smooth semi-parametric adjustment of rate differences, risk differences and relative risks 105

Elisa Duarte, Bruno de Sousa, Carmen Cadarso-Suarez, Vitor Rodrigues, Thomas Kneib: Structured Additive Regression (STAR) models applied in the analysis of breast cancer risk in central Portugal 111

Paul H. C. Eilers, Marco C. A. M. Bink: RNA Sequencing and Zipf's Law 117

J. Etxeberria, M.D. Ugarte, T. Goicoa, A. Militino: Forecasting cancer mortality figures in Spanish provinces with an ANOVA-type P-spline model 123

Salvatore Fasola, Vito M.R. Muggeo: Change-point estimation in piecewise constant regression models with random effects 127

Gianluca Frasso, Paul H.C. Eilers: Composite smooth estimation of the state price density implied in option prices 133

Jan Gertheiss, Verena Maier, Engel F. Hessel, Ana-Maria Staicu: Modeling Binary Functional Data with Application to Animal Husbandry 139

Ipek Guler, Christel Faes, Francisco Gude, Carmen Cadarso-Suárez : Comparing the predictive performance of different regression models for longitudinal and time-to-event data 145

Karl Heiner , John Hinde : Measures of Interaction for Relationships Among Dichotomous Variables 151

Gillian Heller, Lindsay Dunlop: A latent variable approach to digit preference 157

Jesus Crespo–Cuaresma, Martin Feldkircher, Bettina Grün, Paul Hofmarcher , Stefan Humer : A latent variable approach to derive consensus GDPs 163

Nadja Klein, Francisco Gude, Carmen Cadarso-Suárez,, Thomas Kneib: Bivariate Gaussian Distributional Regression: An Application on Diabetes 167

Arnošt Komárek, María José García-Zattera, Alejandro Jara: Regression modelling of misclassified correlated interval-censored data 173

Matieyendou Lamboni, Renate Koeble, Adrian Leip: Bayesian spatial disaggregating of shares: application to land use shares in EU 179

Emmanuel Lesaffre , Baoyue Li, Luk Bruyneel: Joint modeling of a multilevel factor analytic model and a multilevel covariance regression model 185

Gianfranco Lovison, Christian Schindler: Separate regression modelling of the Gaussian and Exponential components of an EMG response from respiratory physiology 189

Marica Manisera, Paola Zuccolotto: New perspectives on rating data modelling: the Nonlinear CUB 195

Juliane Manitz, Jonas Harbering, Marie Schmidt, Thomas Kneib, Anita Schöbel: Network-based Source Detection: From Infectious Disease Spreading to Train Delay Propagation 201

Louise Marquart, Michele Haynes, Peter Baker: Impact of misspecified random effect distributions on models for panel survey data 207

Sebastian Meyer, Leonhard Held: Flexible estimation of spatio-temporal interaction in a point process model for infectious disease spread 213

Annette Möller, Gerhard Tutz, Jan Gertheiss: Random Forests for Functional Covariates 219

Abdolreza Mohammadi, Fentaw Abegaz, Ernst Wit: Efficient Bayesian inference for Copula Gaussian graphical models 225

Pedro A. Morettin, Chang Chiann, Michel H. Montoril: Wavelet Estimation of Functional-coefficient Regression Models 231

C.K. Mutambanengwe, C. Faes, M. Aerts: A restricted composite likelihood approach to modelling Gaussian spatial data 235

Pathé Ndao, Aliou Diop, Jean-François Dupuy: Estimating conditional extreme quantiles under random censoring 239

Margret-Ruth Oelker, Fabian Sobotka, Thomas Kneib: On (Semiparametric) Mode Regression 243

Helen Ogden : Inference for generalized linear mixed models with sparse structure 249

Yi Pan, Jianxin Pan: Joint modeling of between-subject and within-subject covariance matrices for financial data 255

Daniela Pauger, Helga Wagner: Bayesian Effect Fusion for Categorial and Ordinal Predictors 261

Jean Peyhardi, Catherine Trottier, Yann Guédon: Partitioned conditional generalized linear models for categorical data 269

Eugen Pircalabelu, Gerda Claeskens, Sara Jahfari, Lourens Waldorp: Nodewise graphical modeling using the Focused Information Criterion for 'p larger than n' settings. 273

Wolfgang Pössnecker, Gerhard Tutz: Regularization and Selection of Proportional Versus Nonproportional Effects in Sequential Logit Models 279

Pedro Puig: Beyond the Gauss' principle 285

Trias Wahyuni Rakhmawati, Geert Molenberghs, Geert Verbeke, Christel Faes: Local Influence Diagnostics for Generalized Linear Mixed Models With Overdispersion 291

Shahla Ramzan, Gerhard Tutz, Christian Heumann : Improved Methods for Nearest Neighbor Imputation 297

Christian Röver, Tim Friede: Bayesian inference in random-effects meta-analysis ... 303

C. M. Russo, S. P. Willemsen, D. Leão, E. Lesaffre: Nonlinear mixed-effects models for bioequivalence on pharmacokinetic data ... 307

Carl Scarrott: Statistical Estimation of Pollution Reductions for Meeting Government Targets ... 313

Gunther Schauberger, Gerhard Tutz: DIFboost: A Boosting Method for the Detection of Differential Item Functioning ... 319

Fabian Scheipl, Ana-Maria Staicu, Sonja Greven: Functional Additive Mixed Models ... 325

Sabine K. Schnabel, Paul H.C. Eilers : Expectile smoothing for big data ... 331

Fabian Sobotka, Thomas Kneib: Semiparametric quantile regression using a mixed models representation ... 337

James Sweeney, Finbarr O'Sullivan, David Hawe: Improved quantitative analysis of tissue characteristics in PET studies with limited uptake information ... 341

Ardo van den Hout: Time-dependency in multi-state models: specification and estimation ... 347

Tobias Voigt, Roland Fried: Modeling Cherenkov Telescope images for Variable Construction in Classification ... 353

Matthieu Wilhelm , Laura M. Sangalli: Spatial Splines for Generalized Additive Models ... 359

Paul Wilson: The Misuse of The Vuong Test For Non-Nested Models to Test for Zero-Inflation ... 363

Author Index ... 369

Part I - Invited Papers

"La statistique à Goettingue": a tentative tribute by an IWSM participant from Toulouse

Antoine de Falguerolles[1]

[1] Université de Toulouse (retired), France

E-mail for correspondence: antoine@falguerolles.net

Abstract: The historical background of the academic world in Göttingen in the early 19th century and its main institutions are outlined. A tentative sample of figures whose works contributed to the statistical reputation of Göttingen in the period 1750-1900 is considered. A French perspective is adopted.

Keywords: History of statistics, rise of statistical modelling in Göttingen, statistical visualization, generalized linear models

1 Introduction

What does the name Göttingen suggest to a typical man (or women) in the street in Toulouse? When asked, Barbara's song comes first, Friedrich Gauss is second, and David Hilbert is rarely mentioned. Barbara's memorable song does not fit the scientific scope of an International Workshop on Statistical Modelling (IWSM) meeting. Bernard Bru, Anders Hald, Oscar Sheynin, Stephen M. Stigler, ... have done justice to the *mathematicorum principes* Gauss and Hilbert (and to many more!). What more can be said on "La statistique à Goettingue"?
German universities in general and Göttingen university in particular with their curricula, libraries, and links with related Academies generated a genuine interest in France after the Treaty of Basle (1795). Two former alumni of Göttingen, the French born Charles de Villers and the Swiss born Philipp Albert Stapfer, played an important role in this dissemination (Décultot, 2008). The contrast with the French situation was striking: the French universities had been closed under the Revolution in 1793 and were to be recreated by Napoléon as a centralised imperial university (1806, 1808). The French interest in the German system lasted in learned circles notwithstanding tumultuous relationships. A testimony of the scientific recognition

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

of Göttingen university can be found in the journal kept by Maurice Janet, a French visiting graduate student of Hilbert in 1912 (Mazliak, 2014).
What follows echoes my own personal perception of the rise of statistical modelling which benefited from the academic climate in Göttingen. Section 2 discusses the emergence of the word statistics which Göttingen seems to be credited for. Section 3 describes some of the historical background of the academic world in Göttingen and Section 4 its main institutions in the early 19th century. Section 5 presents a tentative and biased sample of figures who, before the First World War, contributed to the statistical reputation of Göttingen. Excerpts of documents and reanalyses of data will illustrate my oral presentation.

2 The emergence of the word statistics

In 1819, when elected to the French Academy, Pierre-Édouard Lémontey (1762-1826) had to deliver an eulogy to his predecessor, the well known Abbé André Morellet (1727-1819). In his speech, Lémontey asserted that Morellet had contributed to a "singular novelty" of his epoch. The singularity was the birth to two positive sciences, one allegedly established in Germany under the name of Statistics (*statistique*) and the other in England, also allegedly, under the name of political economy (*économie politique*). Lémontey further emphasised that Morellet had clearly established their undisputable French origin! Ignoring the French pretence, Theodore M. Porter (Porter, 1986), less committed to claiming credit *pour la France*, appropriately recalls that

> [it is] William Petty, who invented the phrase "Political arithmetic" and is thought by many to have a hand in the composition of [John] Graunt's [*Observation upon the Bills of Mortality of 1662*] (p. 19)

and that

> "Statistics" derives from a German term, *Statistik*, first used as a substantive by the Göttingen professor Gottfried Achenwall in 1749 (p. 23).

2.1 Earliest appearance?

In most papers or books on the history of statistics, and whatever the language used, there is a section devoted to the first use of the words statistics and statistician. This can be exemplified in French and Italian. Maurice Block (Block, 1878), the Prussian-born French statistician, acknowledges the priority of Gottfried Achenwall (1719-1772) but mentions (p. 6) that Guéry (*sic*) had quoted in his "famous work" the use of both words (in latin) as early as 1672 by Heleno Politano (`http://archive.thulb.uni-jena.de/hisbest/receive/HisBest_cbu_00020823`). Note that an earlier-than-Politano French use of the word *statistique*, widely accepted in France, is

inaccurate (Pépin, 2005). The Italian Giovanni-Battista Salvioni (Salvioni, 1879) recognizes in Hermann Conring (1606-1681), from Helmstedt University, a forerunner of modern government statistics but also quotes Heleno Politano.
Whatever the truth, statistics became rapidly a fashionable word although its meaning remained unclear. There are numerous instances of bizantine discussions on what statistics is about and how other fields are not statistics. In the early 19th century, these mostly addressed the respective domains of political economy, state administration, history and geography. But there were also interesting intuitions: the naturalist Jean-Baptiste Lamarck (1744-1829) coined the phrase *Météorologie-statistique* in an early French statistical journal (*Annales de Statistique*, 1802).

2.2 And nowadays?

Is statistics a well defined domain? This is still a matter of taste and discussions. Afficionados of clustering, machine learning, data analysis, data visualisation, big data, ... are often uncompromising in their claim for autonomy. Still, I assume that, for most participants of the IWSM, statistics is the interplay of data and probabilistic models. Of course, data and models are both constructions, the more abstruse being sometimes the data rather than the model.

3 Setting historical and geopolitical backgrounds

Göttingen is seated in the German Federal state (*Bundesland*) of Lower Saxony (*Niedersachsen*) and hosts an internationally renowned university. This is the interesting result of a complicated evolutionary process, an oversimplified slice of its trajectory being summarized below.

3.1 The Electorate: 1708-1706

The territories of the actual Land include historic sub-principalities of the medieval Duchy of Brunswick-Lüneburg or, later, of the Electorate of Hanover (formally the Electorate of Brunswick-Lüneburg) (1708). The House of Hanover headed the Electorate. This house was a cadet branch of a prominent and intricate family, the Guelphs (*Welf*), with a complicated history of divisions, acquisitions and mergers of principalities. In addition, in 1714, the Elector, *Georg Ludwig* (1660-1727), became king of Great Britain and Ireland (with a claim on France) under the name of George I, so that the Electorate and Great Britain and Ireland were ruled in personal union.

3.2 The Napoleonic interlude: 1807-1813

During this short interval, Napoléon created the Kingdom of Westphalia (*Königreich Westphalen*) ruled by his younger brother Jérôme. The Kingdom was bilingual (*Deutsch* and *Französisch*) and replicated the French

"model" (law, administration ...); its Capital was Cassel (now *Kassel)*. A tentative assessment of the impact of this Napoleonic enterprise is given in Knopper (2008).
Göttingen became the seat of the *préfecture* (*Präfektur*) of the Leine Department (*Departement der Leine*), named after the Leine river which flows through it. In 1810, the Kingdom was remodelled and incorporated three more departments, among them the *Departement der Aller*, with Hanover as seat of the *préfecture*. The short lived Kingdom supported three universities: Halle, Göttigen and Marburg. In 1810, King Jérôme had closed the Helmstedt University (another institution connected with the Guelph family, the *Braunschweig-Wolfenbüttel*). Note that a similar situation had occured in the Netherlands where Napoléon had closed the university of Franeker in 1811.

3.3 The Kingdom of Hanover

In 1814 the Congress of Vienna elevated the Electorate to a kingdom, with Hanover as its Capital. A Higher Vocational College/Polytechnic Institute (*Höhere Gewerbeschule/Polytechnische Schule*) was founded in 1831 which evolved into a reputed technical university recently renamed the *Gottfried Wilhelm Leibniz Universität Hannover*. The joint ruling ended in 1837 with the death of William IV Henry: the Guelphs did comply to the German semi-Salic law (his brother Ernest Augustus (1771-1851) ascended to the Hanover throne) and to the British male-preference (his niece Victoria (1819-1901) to the throne of the United Kingdom of Great Britain and Ireland).
In 1866 the Kingdom of Hanover was annexed by Prussia and became the Prussian Province of Hanover. In 1871 it became part of the German Empire along with Alsace and the Moselle region of the Lorraine (*Reichsland Elsaß-Lothringen*) following the Franco-Prussian War.

4 Some academic institutions in Göttingen

The Napoleonic *Almanach Royal de Wesphalie* (Cassel: Imprimerie royale, 1812 and 1813) provides a detailed picture in French of the academic institutions in Göttingen. This is obviously biased information. But it can be supplemented. For example, the Göttingen academic paradigm at the turn of the 18th century has been thoroughly investigated in a recent Franco-German book edited by Bödeker *et al.* (2010). My own view follows.

4.1 The Georgia Augusta

The university, founded in 1734-1737, was innovatively designed by Gerlach Adolph Baron Münchhausen (1688 - 1770) to be in the spirit of the Enlightenment. It is named after George II Augustus (*Georg August*) the second ruler in personal union of the kingdoms of Great Britain and Ireland, and of the Electorate of Hanover.

During the Napoleonic interlude, the *Université de Gœttingen* comprised four Faculties (Theology, Law, Medicine and Philosophy). The latter consisted of Philosophy *per se*, Mathematics in general, and in particular ... Astronomy, ..., **History and Statistics**, ..., Music. As expected Carl Friedrich Gauss (1777-1855), professor of astronomy at Göttingen, held a position of ordinary professor in the Faculty of Philosophy. He is designated as *le chevalier* Gauss since he had been knighted to the Napoleonic *Ordre de la Couronne de Westphalie*.

4.2 Göttingen Academy

The city also hosts the Göttingen Academy of Sciences and Humanities (*Akademie der Wissenschaften zu Göttingen*), a learned society, founded by George II Augustus in 1751 and still active.
During the Napoleonic interlude, the *Société Royale des sciences, de l'histoire et de la littérature, à Goettingue* used to meet once a month. The Society, organised in four disciplinary classes, counted ordinary members (*circa* 20), honorary members (*circa* 10), foreign associate members (*circa* 90), and corresponding associates (*circa* 200). The second class dealt with **pure** (*pures*) and **applied** (*mixtes*) **mathematics** and **astronomy**; it counted 4 members: Messrs Mayer, Thibaut, Harding, and *le chevalier* Gauss.
The foreign associate members were mostly Europeans (with the exception of Spanish or Portugese members). Among them:

- several leading astronomers: John Herschel (1792-1871) (the son of Wilhelm Herschel, the Hanoverian-born British astronomer), Franz-Xaver von Zach (1754-1832), Count (Pierre-Simon de) Laplace (1749-1827), Jean-Baptiste Delambre (1749-1822), (Father) Giuseppe Piazzi (1746-1826), (Father) Barnaba Oriani (1752-1832), Guillaume Olbers (1758-1840).
- two prominent personal counselors to the Prussian King: Wilhelm von Humboldt and his younger brother Alexander von Humboldt (1769-1859).
- a leading French polymath who pioneered official statistics in France (the so-called *statistique des préfets*): le comte Chaptal (1756-1832).

The corresponding associates were less recognized personalities, some of them having a real connection with different conceptualisations of statistics: Emmanuel Étienne Duvillard de Durand (1755-1832) (Swiss-born, Paris), Gaspard Riche de Prony (1755-1839) (Paris), Denis-François Donnant (1769-18..) (Paris), Hendrik-Willem Tydeman (1778-1863) (Franeker), Simon L'Huillier (1750-1840) (Geneva), ...
In the *Almanach*, the long list of foreign or corresponding associates comprises names which are misspelt or enigmatic. Examples are Chapollion Figeni (Jacques-Joseph Champollion-Figeac), elder brother of the decipherer of the Rosetta Stone, and Stettio-Doria Proffalendi (Stylianos Dorias Prosalentis?) from Corfu Academy. (A Ionian Academy had been organised in Corfu also occupied by Napoléon.) Their expertise in Greek had presumably

motivated their nomination. Indeed, Philological studies had immensely developped in Göttingen under *le chevalier Heyne* (1729-1812), director of the famous university library.

4.3 Observatory

The first astronomical observatory in Göttingen had been installed in 1750 and the astronomer Tobias Mayer hired in 1751. The construction of a new observatory, decided by George III, was delayed by the French Revolutionary Wars. *Le chevalier* Gauss became its director in 1807. Actually, astronomers (and their active network) have a uncommon tradition for "Big Data" and mathematical models. Statistical modelling owes them at least L_1, L_2, minimax linear fits (Farebrother, 1999).

4.4 Miscellanea

A library, a museum, several institutes, seminars, ... were attached to the university. By 1800, the library counted more than 130,000 volumes while university libraries in Europe or America had typically no more than 30,000. The *Collection des modèles et machines* (Göttingen collection of Mathematical Models and Instruments) which is still maintained, was then headed by Professor Mayer.

5 A tentative sample of early statistical modellers

Below is a contestable list of figures related to Göttingen, to topics addressed during past IWSMs, and to my own statistical interests.

5.1 Gottfried Leibniz

Gottfried Wilhelm (von) Leibniz (1646-1716), the famous polymath, is usually recognized for his contribution to the infinitesimal calculus (wether independently of Isaac Newton or not, a debatable issue), to the field of mechanical calculators (arithmometer), and to the binary number system (see the logo of the eponymous university). Leibniz is also a fine historian. In 1676, Leibniz entered into the service of the Gelph family. (In particular he deconstructed the genealogy of the Gelph family.) Interested in actuarial matters, he also investigated the question of insurance and mortality (Rorhbasser, 2007). The mathematical formulations attempted below are tentative translations of verbal descriptions: The random life duration (T) could have cumulative distribution function $F_T(t) = F_T(1)+(1-F_T(1))\frac{t-1}{\theta-1}$ for $1 < t \leq \theta$, where θ is the climateric year 9×9, hazard function $\frac{1}{\theta-t}$ and conditional expectation $E[T|T > t_0] = \frac{\theta+t_0}{2}$ which Leibniz found in excess. This is certainly better than the tentative "model" for life duration discussed in Ephraim Chambers' *Cyclopædia* (1728) and reproduced by many (the *Encyclopédie de Diderot et de d'Alembert*: entry *politique arithmétique*; Denis-François Donnant's *Théorie élémentaire de la Statistique*, 1805) where the hazard function could translate as $\frac{1}{2t}$!

5.2 Tobias Mayer

Tobias Mayer (1723-1762) was a self-taught mathematician, cartographer, and astronomer. His work on the oscillation of the moon was published in 1750, *Kosmographische Nachrichten und Sammlungen auf das Jahr 1748* (p. 52-172). (`http://www.e-rara.ch/doi/10.3931/e-rara-2770verb`). He provided an elegant solution to the estimation of coefficients in linear regression ($E[Y] = X\beta$ and $Var(Y) = \sigma^2 I$), namely the particular construction of a matrix U such that $U'X$ is regular. (Note in passing that matrices did not exist then.) Hence $\hat{\beta}_M = (U'X)^{-1}U'Y$, $E[\hat{\beta}_M] = \beta$, and $Var(\hat{\beta}_M) = \sigma^2(U'X)^{-1}U'U(X'U)^{-1}$. Mayer's solution (1750) amounts to select an *ad hoc* partition of the observations in $|\beta|$ clusters, the indicator matrix of which is used to define X. In other words, by averaging the data in each cluster, β is the solution of a set of $|\beta|$ linear equations. (For some connections with clustering methods, see Falguerolles, 2009.) In 1788, also for computational simplification, Laplace restricted the coefficients in U to $+1$, -1, 0.

Appointed to a Chair of economy and mathematics at the *Georgia Augusta*, Tobias Mayer moved from Nuremberg to Göttingen in 1751. He became superintendent of the observatory in 1754. The Mayer mentioned in the *Almanach* is Tobias Mayer's son: alumni of the *Georgia Augusta*, professor in several German universities, Johann Tobias Mayer (1752-1830) obtained a position in Göttingen in 1799.

5.3 Ludwig Schlözer

August Ludwig von Schlözer (1735-1809) was the successor of Achenwall in Göttingen where he was promoted to a professorship in 1769. In the tradition of Cameralism, but also influenced by the ideas of Adam Smith (1723-1790), the Scottish moral philosopher and pioneer of political economy, his work was a solid attempt at conceptualization of statistics (Becker and Clark, 2001, and Garner, 2010). His numerous books were well received and, in particular, the *Theorie der Statistik nebst Ideen über das Studium der Politik überhaupt.* (See `http://ds.ub.uni-bielefeld.de/viewer/image/1493486/1/`.)

Two of the corresponding associates propagated Schlözer's treaty outside the German world. The French Denis-François Donnant rapidly translated the book into French. His translation which includes a preliminary discourse, some additions and remarks is titled *Introduction à la Science de la statistique, suivie ...* (Paris: Imprimerie Impériale, 1805). Donnant is more known for his translations (sometimes inventive) than for his own work. He had already translated and published William Playfair's *Statistical Breviary* (London: Wallis, 1801) with an addition which, in turn, Playfair translated and published (London: J. Whiting, 1805). The Dutch Hendrik-Willem Tydeman, then professor at Franeker university, also translated and published Schlözer's book: *Theorie der Statistiek of Staats-Kunde*, Groningen: Wijbe Wouters and Amsterdam: J.F. Nieman, 1807 (see Stamhuis, 2010).

5.4 Alexander von Humboldt

Alexander von Humboldt (1769-1859) was a Prussian geographer, naturalist and explorer. Educated in various German universities, he matriculated at the *Georgia Augusta* in 1789. Between 1799 and 1804, Humboldt and Aimé Bonpland (1773-1858), his French accomplice, travelled extensively in Latin America measuring and collecting almost everything. During this memorable expedition, Humboldt laid the foundation of physical geography, botanical geography, and meteorology. The impressive volume of Humboldt's publications hid some novel statistical graphics. The influence of the German August F. W. Crome (1753-1833), cameralist and statistician? Of the Scottish William Playfair (1759-1823)? In particular, Humboldt's *Essai sur la Géographie des Plantes ; acccompagnée d'un tableau physique des régions équinoxiales* (Paris: Levrault, 1805) is illustrated by the famous *tableau physique* showing the Chimborazo and the Cotopaxi volcanos in Ecuador (then a Spanish colony), and several physical phenomena linked to altitude. A useful website for seing Humboldt's *tableau* (and an amazing menagery of statistical graphics) is the Milestones project (Friendly and Denis, 2001). Humboldt's graphic was influential. For early followers, see Palsky (2010). A French example, *L'essai sur la statistique universelle du globe terrestre* (1815), is provided by Pierre-Bernard Barrau (1767-1843); born in Toulouse, Barrau was an unsuccessful pioneer in agriculture insurance and a naive statistical "believer".

5.5 Wilhelm Lexis

Wilhelm Lexis (1837-1914) was an economist, a statistician and a demographer. His academic mobility is quite exemplary: universities of Strasbourg (then Prussian), Dorpat (now Tartu), Freiburg im Breisgau, Breslau, and finally Göttingen in 1887. If Lexis has given his name to two tools reflecting questions still addressed by statistical modellers, it is often duly argued that he should not be credited for their invention but rather for their development. Stigler's law of eponymy again!

The Lexis ratio Q^2 aims at quantifying the stability of k replicated series of uncorrelated binary events (X_{ij}) with common length (n_0). The statistic Q^2 (or L^2) compares two variance estimates for the rate observed in a replication: $\frac{\overline{X}(1-\overline{X})}{n_0}$ and $\frac{1}{k-1}\sum_{i=1}^{k}(\overline{X_i}-\overline{X})^2$. A practical example taken from Richard von Mises's textbook (1957) is reanalysed by Gelman (2011) who elicits the link between Lexis ratio and the chi-square test statistics for overdispersion.

The Lexis diagram is a visualisation tool of individual lifetimes. Technically it is a planar representation of three dimensional data, hence the connection with stereograms. Keiding (2000) has thoroughly investigated the rise and the mathematical principles of theses graphics. The visualisation (Lexis pencils) of a moderate number of individual time changes in a number of categorical variables is considered in Francis and Fuller (1996); the analytical usefulness of this graphic is increased by zoomings and animations.

5.6 Ladislaus von Bortkiewicz

Ladislaus von Bortkiewicz (1868-1931) studied and taught at the universities of Strasbourg and Göttingen. In particular he defended his doctoral thesis at the *Georgia Augusta*. His famous monograph on the law of small numbers, (*Das Gesetz der Kleinen Zahlen* or 1898, Leipzig: Teubner), is dedicated to his advisor Lexis. (See `https://archive.org/details/dasgesetzderklei00bortrich`.) The book gives a talented presentation of the Poisson distribution (standard errors of standard errors are provided!) illustrated by detailed analyses of data sets (in particular, the famous two-way table of Prussian *Miltärpersonen* deaths by horse-kicks!). Undisputably his book is a forerunner to those of McCullagh and Nelder (1982, 1989) and of Aitkin and *al.* (1989) which very much influenced the creation of the IWSM. His examples and this workshop offers me a unique occasion to reminisce with nostalgia the GLIM package and its flexibility (Falguerolles and Francis, 1995).

Acknowledgments: I would like to thank François Bompaire, Brian Francis, Michael Friendly, Thomas Kneib, and Gilles Palsky for their encouragements and help in preparing this presentation. In the process, I had also the pleasure to trace "cousin" Johann Peter Falguerolles from Bremen and medical student in Erlangen (1785). All errors and approximations remain mine.

References

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989). *Statistical modelling in GLIM*. Oxford University Press.

Becker, P. and Clark, W. (2001). *Little Tools of Knowledge: Historical Essays on Academic and Bureaucratic Practices*. University of Michigan Press.

Block, M. (1878). *Traité théorique et pratique de la Statistique*. Paris: Guillaumin et Cie.

Bödeker, H. E. (2010). "Pour la vraie politique libre allez donc à Göttingen". Les théories de la politique à Göttingen autour de 1800. In: *Göttingen vers 1800*, H. E. Bödeker, Ph. Büttgen and M. Espagne (Eds.), Paris: Les Éditions du Cerf, pp. 401 – 456.

Décultot, E. (2008). La réception française du modèle universitaire allemand. L'exemple de l'université de Göttingen. In: *L'Allemagne face au modèle français*, F. Knopper and J. Mondot (Eds.), Toulouse: Presses Universitaires du Mirail, pp. 239 – 259.

Falguerolles A. de and Francis B. (1995). Fitting bilinear models in GLIM. *GLIM Newsletter*, **25**, 9 – 20.

Falguerolles, A. de (2009). La méthode d'ajustement de Mayer et ses liens avec les méthodes de classifications. *Mathématiques et Sciences humaines/Mathematics and Social Sciences*, **187**, 43 – 58.

Farebrother, R. W. (1999). *Fitting linear relationships, a history of the calculus of observations 1750-1900*. New York: Springer.

Francis, B. and Fuller, M. (1996). Visualization of Event Histories. *Journal of the Royal Statistical Society (Series A)*, **159**(2), 301 – 308.

Friendly, M. and Denis, D. J. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. Web document, `http://www.datavis.ca/milestones/`. (Accessed 29 April 2014.)

Garner, G. (2010). Économie politique et statistique à Göttingen autour de 1800. In: *Göttingen vers 1800*, H. E. Bödeker, Ph. Büttgen and M. Espagne (Eds.), Paris: Les Éditions du Cerf, pp. 457 – 479.

Gelman, A. (2011). Going beyond the book towards critical reading in statistics teaching. *Teaching statistics*, **34**(3), 82 – 86.

Keiding, N. (2000). Mortality measurements in the 1870s: diagrams, stereograms, and the basic differential equation. Talk given at the workshop "Lexis in context", Rostock. Web document, `http://www.demogr.mpg.de/papers/workshops/000828_paper02.pdf`. (Accessed 20 April 2004.)

Knopper, F. (2008). La Westphalie, un laboratoire des idées napoléoniennes. In: *L'Allemagne face au modèle français*, F. Knopper and J. Mondot (Eds.), Toulouse: Presses Universitaires du Mirail, pp. 181 – 196.

Mazliak, L. (2013). *Le voyage de Maurice Janet à Göttingen*. Paris: Éditions Matériologiques.

McCullagh, P. and Nelder J. A. (1989). *Generalized Linear Models (2nd Edition)*. London: Chapman and Hall.

Palsky, G. (2010). Le Tableau de la Hauteur des Montagnes. Un paysage de fantaisie entre art et géométrie. In: *Lenguajes y visiones del paisaje y del territorio*, N. Ortega Cantero, J. García Alvarez, M. Molla Ruiz-Gómez M. (Eds.), Madrid: UAM ediciones, pp. 297 – 307.

Pépin, D. (2005). Claude Bouchu et le mot "Statistique". *Journal de la Société française de Statistique*, **146**(3), 119 – 130.

Porter, Th. M. (1986). *The rise of statistical thinking 1820-1900*. Princeton University Press.

Rohrbasser, J.-M. (2007). Leibniz : assurance, risque et mortalité. *Asterion*, **5**, 197 – 218.

Salvioni, G. B. (1978). Prefazione. In: Georg Mayr, *La Statistica e la vita sociale, versione dal tedesco, approvata d'all Autore con introduzione storica, aggiunte note del Dott. Salvioni.* Torino e Roma: Ermanno Loescher.

Stamhuis, I. (2010). German *staatenkunde* or French 'Numbers and Equations'; Statistics and the Demise of the Dutch statistical Society. *Electronic Journ@l for History of Probability and Statistics*, 6 (2).

Bayesian nonparametric approaches for the analysis of compositional data based on Bernstein polynomials

Alejandro Jara[1], Andrés F. Barrientos[1]

[1] Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile. Work supported by Fondecyt grants 1141193 and 3130400.

E-mail for correspondence: atjara@uc.cl

Abstract: We discuss Bayesian nonparametric procedures for density estimation and fully nonparametric regression for compositional data, that is, data supported in a m–dimensional simplex Δ_m. The procedures are based on modified classes of multivariate Bernstein polynomials. We show that the modified classes retain the well known approximation properties of the classical versions defined on $[0,1]^m$ and Δ_m, $m \geq 1$. Based on these classes, we define prior distributions on the space of all probability measures defined on Δ_m, $\mathcal{P}(\Delta_m)$. We show that the processes are well defined, have large support and the frequentist asymptotic behaviour of the posterior distribution is appropriated. Finally, novel classes of probability models for sets of predictor-dependent probability distributions are proposed. Appealing theoretical properties such as support, continuity, marginal distribution, correlation structure, and consistency of the posterior distribution are studied.

Keywords: Simplex; Random Bernstein polynomials; Dependent Dirichlet processes

1 Introduction

Models for probability distributions based on convex combinations of densities from parametric families underly mainstream approaches to density estimation, including kernel techniques, nonparametric maximum likelihood and Bayesian nonparametric (BNP) approaches. From a BNP point of view, the mixture model provides a convenient set up for density estimation in that a prior distribution on densities is induced by placing a prior distribution on the mixing measure. On the real line, a mixture of normal densities induced by a Dirichlet process (DP) is often used to model smooth densities. Due to the flexibility and ease in computation,

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

these models are now routinely implemented in a wide variety of applications. While the normal kernel is a sensible choice on the real line, its usefulness is rather limited when considering densities on convex and compact subspaces, such as the closed unit interval or the m–dimensional simplex $\Delta_m = \{(y_1, \ldots, y_m) \in [0,1]^m : \sum_{i=1}^m y_i \leq 1\}$. Although methods based on the normal kernel could be used to deal with data supported on these spaces, by using transformations, the resulting model is susceptible to boundary effects.

Motivated by its uniform approximation properties, frequentist and Bayesian methods based on univariate Bernstein polynomials (BP) and more general discrete mixtures of beta distributions have been proposed for the estimation of probability distributions supported on bounded intervals (see, e.g., Petrone 1999). Extensions based on multivariate BP (MBP) defined on the unit hyper-cube have been considered in the statistical literature (see, e.g., Zheng et al., 2010). Multivariate extensions of Bernstein polynomials defined on Δ_m were considered by Tenbusch (1994), to propose and study a density estimator for the data supported on Δ_2. Although Tenbusch's estimator is consistent and optimal at the interior points of the simplex, it is not a valid density function for finite sample size.

We will discuss Bayesian nonparametric approaches for single density estimation and for the estimation of collections of conditional densities based on modified classes of MBP. The modified clases of MBP and its main properties are given in Section 2. The proposed models for probability measures defined on Δ_m are discussed in Section 3. Finally, proposed models for collections of probability measures defined on Δ_m are discussed in Section 4. Sections 2 and 3 summarise the works by Barrientos et al. (2014) and Barrientos and Jara (2014). Section 4 describes ongoing research on the subject.

2 Multivariate Bernstein polynomials on Δ_m

2.1 The original class

Tenbusch's estimator arises by taking G to be the restriction of the empirical CDF to Δ_2, and it is based on the following class of MBP. For a given bounded function $G : \Delta_m \longrightarrow \mathbb{R}$, the associated MBP of degree $k \in \mathbb{N}$ on Δ_m is defined by

$$B_{k,G}(\boldsymbol{y}) = \sum_{\boldsymbol{j} \in \mathcal{J}_m^k} G\left(\frac{j_1}{k}, \ldots, \frac{j_m}{k}\right) \frac{k!}{(\prod_{l=1}^m j_l!)\,(k - \sum_{l=1}^m j_l)!} \left(\prod_{l=1}^m y_l^{j_l}\right) \left(1 - \sum_{l=1}^m y_l\right)^{k - \sum_{l=1}^m j_l},$$

where $\boldsymbol{j} = (j_1, \ldots, j_m)$, and

$$\mathcal{J}_m^k = \left\{(j_1, \ldots, j_m) \in \{0, \ldots, k\}^m : \sum_{l=1}^m j_l \leq k\right\}.$$

It is not difficult to show, however, that if G is the restriction of the CDF of a probability measure on Δ_m, then $B_{k,G}(\cdot)$ is not the restriction of the CDF of a probability measure defined on Δ_m for a finite k; $B_{k,G}(\cdot)$ can be expressed as a linear combination of CDF's of probability measures defined on Δ_m, where the coefficients are nonnegative but do not add up to one.

2.2 The modified classes

To avoid the problem of Tenbusch's estimator, Barrientos et al. (2014) proposed a modified class of MBP, which is obtained by increasing the size of the set $\mathcal{J}_m^k$ and the domain of the function G. For a given function $G : \mathbb{R}^m \longrightarrow \mathbb{R}$, the associated MBP of degree $k \in \mathbb{N}$ on Δ_m is defined by

$$B_{1,k,G}(\boldsymbol{y}) = \sum_{\boldsymbol{j}\in\mathcal{H}_m^k} G\left(\frac{j_1}{k},\ldots,\frac{j_m}{k}\right) \frac{k!}{\left(\prod_{l=1}^m j_l!\right)\left(k-\sum_{l=1}^m j_l\right)!} \left(\prod_{l=1}^m y_l^{j_l}\right)\left(1-\sum_{l=1}^m y_l\right)^{k-\sum_{l=1}^m j_l}, \quad (1)$$

where $\mathcal{H}_m^k = \{(j_1,\ldots,j_m) \in \{0,\ldots,k\}^m : \sum_{l=1}^m j_l \leq k+m-1\}$.
Alternatively, Barrientos and Jara (2014) proposed another modified class of MBP, which is given next. For a given function $G : \mathbb{R}^m \longrightarrow \mathbb{R}$, the associated MBP of degree $k \in \mathbb{K} = \{l \in \mathbb{N} : l^{1/2} \text{ is an integer}\}$ on Δ_m is defined by

$$B_{2,k,G}(\boldsymbol{y}) = \sum_{\boldsymbol{j}\in\mathcal{H}_m^k} G\left(\frac{\mathcal{T}_k(j_1)}{\sqrt{k}},\ldots,\frac{\mathcal{T}_k(j_m)}{\sqrt{k}}\right) \frac{C(\boldsymbol{j})k!}{\left(\prod_{l=1}^m j_l!\right)\left(k-\sum_{l=1}^m j_l\right)!} \left(\prod_{l=1}^m y_l^{j_l}\right)\left(1-\sum_{l=1}^m y_l\right)^{k-\sum_{l=1}^m j_l}, \quad (2)$$

where

$$C(\boldsymbol{j}) = k^{-\tilde{m}(\boldsymbol{j})/2}\left(1 - I_{\mathcal{Q}^k_{\tilde{m}(\boldsymbol{j})}}(\boldsymbol{j})\right) + \frac{\tilde{m}(\boldsymbol{j})!(\sqrt{k}-1)!}{(\sqrt{k}+\tilde{m}(\boldsymbol{j})-1)!} I_{\mathcal{Q}^k_{\tilde{m}(\boldsymbol{j})}}(\boldsymbol{j}),$$

$I_A(\cdot)$ is the indicator function for set A, $\tilde{m}(\boldsymbol{j}) = \sum_{i=1}^m (1 - I_{\{0\}}(j_i))$,

$$\mathcal{Q}^k_{\tilde{m}(\boldsymbol{j})} = \left\{\boldsymbol{j} \in \mathcal{H}_m^k : \sum_{i=1}^m \mathcal{T}_k(j_i) = \sqrt{k} + \tilde{m}(\boldsymbol{j}) - 1\right\},$$

$\mathcal{T}_k(j) = \sum_{i=1}^{\sqrt{k}} i\, I_{A(k,i)}(j)$ and $A(k,i) = \{(i-1)\sqrt{k}+1,\ldots,i\sqrt{k}\}$.

2.3 Some properties of the modified classes

The modified classes $B_{1,k,G}$ and $B_{2,k,G}$ retain most of the appealing approximation properties of univariate BP and the standard class of MBP,

$B_{k,G}$. Specifically, if G is a real-valued function defined on $\mathbb{R}^m$ and $G|_{\Delta_m}$ its restriction on Δ_m, then the relations

$$\lim_{k\to\infty} B_{1,k,G}(\boldsymbol{y}) = G|_{\Delta_d}(\boldsymbol{y}),$$

and

$$\lim_{k\to\infty} B_{2,k,G}(\boldsymbol{y}) = G|_{\Delta_d}(\boldsymbol{y}),$$

hold at each point of continuity $\boldsymbol{y}$ of $G|_{\Delta_d}$. Furthermore, these relations hold uniformly on Δ_d if $G|_{\Delta_d}$ is a continuous function.
It is also possible to show that if G is the restriction of the CDF of a probability measure defined on Δ_m, then $B_{1,k,G}(\cdot)$ and $B_{2,k,G}(\cdot)$ are also restrictions of the CDF of probability measures defined on Δ_m. Furthermore, if G is the CDF of a probability measure defined on $\tilde{\Delta}_m = \{\boldsymbol{y} \in \Delta_m : y_j > 0, \quad j = 1, \ldots, m\}$, then $B_{1,k,G}(\cdot)$ and $B_{2,k,G}(\cdot)$ are restrictions of the CDF of probability measures with density functions given by the following mixtures of Dirichlet distributions,

$$b_{1,k,G}(\boldsymbol{y}) \quad = \sum_{\boldsymbol{j}\in\mathcal{H}^0_{k,d}} W_{1,k,\boldsymbol{j},G} \times \mathrm{d}(\boldsymbol{y} \mid \alpha(k,\boldsymbol{j})), \tag{3}$$

and

$$b_{2,k,G}(\boldsymbol{y}) \quad = \sum_{\boldsymbol{j}\in\mathcal{H}^0_{k,d}} W_{2,k,\boldsymbol{j},G} \times \mathrm{d}(\boldsymbol{y} \mid \alpha(k,\boldsymbol{j})), \tag{4}$$

respectively, where $\boldsymbol{j} = (j_1, \ldots, j_m)$,

$$\mathcal{H}^0_{k,m} = \left\{ (j_1, \ldots, j_m) \in \{1, \ldots, k\}^m : \sum_{l=1}^{m} j_l \leq k + m - 1 \right\},$$

$$W_{1,k,\boldsymbol{j},G} = G\left(\left(\frac{j_1 - 1}{k}, \frac{j_1}{k} \right] \times \ldots \times \left(\frac{j_m - 1}{k}, \frac{j_m}{k} \right] \right),$$

$$W_{2,k,\boldsymbol{j},G} = C(\boldsymbol{j}) G\left(\left(\frac{\mathcal{T}_k(j_1) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_1)}{\sqrt{k}} \right] \times \ldots \times \left(\frac{\mathcal{T}_k(j_m) - 1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_m)}{\sqrt{k}} \right] \right),$$

$\mathrm{d}(\cdot \mid (\alpha_1, \ldots, \alpha_m))$ denotes the density function of a m–dimensional Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_m)$, and

$$\alpha\,(k, \boldsymbol{j}) = \left(\boldsymbol{j},\ k + m - \sum_{l=1}^{m} j_l \right).$$

It is easy to see that $b_{1,k,G}(\cdot)$ and $b_{2,k,G}(\cdot)$ are polynomial functions of $\boldsymbol{y}$. These classes can approximate any element in the set of absolutely continuous probability measures defined on Δ_m and with α-Hölder continuous density function, $\alpha \in (0, 1]$. As a matter of fact, if G is an absolutely continuous probability measure defined on Δ_m, w.r.t. Lebesgue measure, with α–Hölder continuous density function g, then

$$\|\tilde{b}_{1,k,G} - g\|_\infty = \mathcal{O}\left(k^{-\alpha/2}\right),$$

and

$$\|\tilde{b}_{2,k,G} - g\|_\infty = \mathcal{O}\left(k^{-\alpha/2}\right).$$

3 Random MBP for density estimation

3.1 The probability models

We induce probability models for densities defined on Δ_m by considering a random k and G in expressions (3) and (4), respectively. Indeed, we define $\mathcal{P}(\Delta_m)$-valued stochastic processes by considering discrete stick-breaking process for G and appropriate priors for the polynomial degrees.
A random probability measure H_1 on $(\Delta_m, \mathbb{B}(\Delta_m))$ is said to be a stick-breaking MBP1 process with parameters

$$\left(\lambda_1, \{H_i^{v,1}\}_{i\geq 1}, \{H_i^{\theta,1}\}_{i\geq 1}\right),$$

written

$$H_1 \mid \lambda_1, \{H_i^{v,1}\}_{i\geq 1}, \{H_i^{\theta,1}\}_{i\geq 1} \sim \text{SBMBP1}\left(\lambda_1, \{H_i^{v,1}\}_{i\geq 1}, \{H_i^{\theta,1}\}_{i\geq 1}\right),$$

if there exists an appropriate probability space such that:

(1.i) $v_i \in [0,1]$, $i \geq 1$, are independent random variables with distribution $\mathcal{H}_i^{v,1}$, and such that

$$\sum_{i=1}^{\infty} \log\left[1 - E_{\mathcal{H}_i^{v,1}}(v_i)\right] = -\infty,$$

(1.ii) $\boldsymbol{\theta}_i \in \tilde{\Delta}_m$, $i \geq 1$, are independent random vectors with distribution $\mathcal{H}_i^{\theta,1}$.

(1.iii) $k \in \mathbb{N}$ is a discrete random variable with distribution indexed by a finite-dimensional parameter λ_1.

(1.iv) the density function of H_1, w.r.t. Lebesgue measure, is given by the following mixture of Dirichlet densities,

$$h_1(\boldsymbol{y}) = \sum_{\boldsymbol{j}\in\mathcal{H}_{k,m}^0} w_{1,k,\boldsymbol{j},G} \times \mathrm{d}(\boldsymbol{y} \mid \alpha(k,\boldsymbol{j})), \tag{5}$$

where

$$w_{1,k,\boldsymbol{j},G} = \left(\left\{\sum_{l_1=1}^{\infty} v_{l_1} \prod_{l_2<l_1} [1-v_{l_2}]\right\} I(\boldsymbol{\theta}_{l_1})_{A_{\boldsymbol{j},k}^1}\right),$$

with

$$A_{\boldsymbol{j},k}^1 = \left(\frac{j_1-1}{k}, \frac{j_1}{k}\right] \times \ldots \times \left(\frac{j_m-1}{k}, \frac{j_m}{k}\right].$$

In a similar way, a random probability measure H_2 on $(\Delta_m, \mathbb{B}(\Delta_m))$ is said to be a stick-breaking MBP2 process with parameters

$$\left(\lambda_2, \{H_i^{v,2}\}_{i\geq 1}, \{H_i^{\theta,2}\}_{i\geq 1}\right),$$

written

$$H_2 \mid \lambda_2, \{H_i^{v,2}\}_{i\geq 1}, \{H_i^{\theta,2}\}_{i\geq 1} \sim \text{SBMBP2}\left(\lambda^2, \{H_i^{v,2}\}_{i\geq 1}, \{H_i^{\theta,2}\}_{i\geq 1}\right),$$

if there exists an appropriate probability space such that:

(2.i) $v_i \in [0,1]$, $i \geq 1$, are independent random variables with distribution $\mathcal{H}_i^{v,2}$, such that

$$\sum_{i=1}^{\infty} \log\left[1 - E_{\mathcal{H}_i^v}(v_i)\right] = -\infty,$$

(2.ii) $\boldsymbol{\theta}_i \in \tilde{\Delta}_m$, $i \geq 1$, are independent random vectors with distribution $\mathcal{H}_i^{\theta,2}$.

(2.iii) $k \in \mathbb{K} = \{l \in \mathbb{N} : l^{1/2} \text{ is an integer}\}$ is a discrete random variable with distribution indexed by a finite-dimensional parameter λ_2.

(2.iv) the density function of H_2, w.r.t. Lebesgue measure, is given by the following mixture of Dirichlet densities,

$$h_2(\boldsymbol{y}) \quad = \quad \sum_{\boldsymbol{j}\in\mathcal{H}_{k,m}^0} w_{2,k,\boldsymbol{j},G} \times \mathrm{d}(\boldsymbol{y} \mid \alpha(k,\boldsymbol{j})), \tag{6}$$

where

$$w_{2,k,\boldsymbol{j},G} = C(\boldsymbol{j})\left(\left\{\sum_{l_1=1}^{\infty} v_{l_1} \prod_{l_2<l_1}\left[1-v_{l_2}\right]\right\} I(\boldsymbol{\theta}_{l_1})_{A_{\boldsymbol{j},k}^2}\right),$$

with

$$A_{\boldsymbol{j},k}^2 = \left(\frac{\mathcal{T}_k(j_1)-1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_1)}{\sqrt{k}}\right] \times \ldots \times \left(\frac{\mathcal{T}_k(j_m)-1}{\sqrt{k}}, \frac{\mathcal{T}_k(j_m)}{\sqrt{k}}\right].$$

Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the mappings induced by expression (5) and (6), respectively, which send the elements of the original probability space to their associated probability measures. Also, let $\mathcal{D}(\Delta_m) \subset \mathcal{P}(\Delta_m)$ be the space of all probability measures defined on Δ_m, absolutely continuous w.r.t. Lebesgue and with continuous density. It is possible to show that $\mathcal{T}_1$ and $\mathcal{T}_2$ are Borel measurable from the original probability space to $\mathcal{P}(\Delta_m)$ under the weak star topology and that is Borel measurable from the original probability space to $\mathcal{D}(\Delta_m)$ under the topology induced by the L_1-norm and the L_∞-norm.

3.2 The support and frequentist asymptotic behaviour properties

Large support is an important and basic property that any Bayesian nonparametric model should ideally possess. In fact, assigning positive mass to neighborhoods of any probability distribution is a minimum requirement (and almost a "necessary" property) for a model to be considered "nonparametric". This property is also important because it is typically a required condition for frequentist consistency of the posterior distribution. Even though the trajectories of a SBMBP1 or a SBMBP2 are only continuous distributions, its topological support, which is the smallest closed set of probability one, can be very large. As a matter of fact, it is possible to show that if the random polynomials degree k has full support, and that for every $i \geq 1$, $H_i^{v,1}$, $H_i^{\theta,1}$, $H_i^{v,2}$ and $H_i^{\theta,2}$ have positive density functions w.r.t. Lebesgue measure, then $\mathcal{P}(\Delta_m)$ is the support of H_1 and H_2 under weak star topology and $\mathcal{D}(\Delta_m)$ is the support of H_1 and H_1 under under the topology induced by the L_∞-norm. Under similar assumptions, it is also possible to show that every element of $\mathcal{D}(\Delta_m)$ is in the Kullback-Leibler support of the H_1 and H_2.
Now suppose that we observed a simple random sample of size n from a "true" probability distribution defined on $(\Delta_m, \mathbb{B}(\Delta_m))$, P, that is

$$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid P \overset{i.i.d.}{\sim} P.$$

A direct consequence of the support properties of the SBMBP1 and SBMBP2 is that the posterior measures of any weak neighborhood of P converges to one as the sample size goes to infinity, where posterior distributions arise by considering the i.i.d. sampling scheme and either SBMBP1 or SBMBP2 as a prior distribution for the sampling model.
Under more specific conditions on the definition of the SBMBPs, stronger consistency results can be shown. Specifically, if the induced prior distribution on the degree of the MBP has a particular tail behaviour, then for every $P \in \mathcal{D}(\Delta_m)$ the posterior measure of every L_1-norm neighborhood of it, converges to one as the sample size goes to infinity.
Finally, if stronger assumptions are made on the "true" probability model and the definition of the SBMBPs, it is possible to characterise the posterior concentration rate. As a matter of fact, it is possible to show that under certain prior specification, the concentration rate of the posterior distribution based on the SBMBP1 prior is slower than $n^{-\alpha/(2\alpha+m)}$, the optimal rate of convergence for multivariate α-smooth densities. However, it is also possible to show that if the "true" density belongs to a Holder class with α-regularity, of at most $\alpha = 1$, then convergence of the posterior distribution based on the SBMBP2 prior is at most $(\log n)^{4\alpha+m/(4\alpha+2m)}/n^{\alpha/(2\alpha+m)}$.

4 Random MBP for fully nonparametric regression

4.1 The inferential problem

Consider regression data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^p$ is a set of predictors, and $\boldsymbol{y}_i \in \Delta_m$ is the vector response variables. Rather than assuming an unknown functional form for the mean function or another functional, as is usually done in nonparametric regression, under the framework of fully nonparametric regression the problem is cast as inference for a family of conditional distributions

$$\{F_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^p\},$$

where $\boldsymbol{y}_i \mid \boldsymbol{x}_i \overset{ind.}{\sim} F_{\boldsymbol{x}_i}$. Therefore, from a Bayesian point of view, the definition of a fully nonparametric regression model requires of the definition of a probability model for the set of predictor-dependent continuous probability distributions $\mathcal{F} = \{F_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$, allowing the complete shape of the elements of $\mathcal{F}$ to change flexibly with the values of $\boldsymbol{x}$.
The problem of defining priors over related random probability distributions has received increasing attention over the past few years. To date, much effort has focused on constructions that generalize the widely used class of Dirichlet process priors for the analysis of data supported on the real line. Models based on MBP for data supported on Δ_m are discussed in the next section.

4.2 The models

To introduce dependence in the continuous random probability measures discussed in Section 3, we replace the stick-breaking mixing distribution in the definition of the processes by dependent stick-breaking process, which is defined by using transformed stochastic processes indexed by predictors $\boldsymbol{x} \in \mathcal{X}$. Set $\tilde{\mathcal{D}}(\Delta_m)^{\mathcal{X}} \subset \mathcal{D}(\Delta_m)^{\mathcal{X}}$, where

$$\tilde{\mathcal{D}}(\Delta_d)^{\mathcal{X}} = \left\{ \{P_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\} \in \mathcal{D}(\Delta_m)^{\mathcal{X}} : (\boldsymbol{y}, \boldsymbol{x}) \longrightarrow p_{\boldsymbol{x}}(\boldsymbol{y}) \text{ is continuous} \right\},$$

with $p_{\boldsymbol{x}}$ denoting the density of $P_{\boldsymbol{x}} \in \mathcal{D}(\Delta_m)$ w.r.t. Lebesgue measure. Let $\mathcal{V} = \{v_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$ be a set of known bijective continuous functions, such that for every $\boldsymbol{x} \in \mathcal{X}$, $v_{\boldsymbol{x}} : \mathbb{R} \longrightarrow [0,1]$, and such that for every $a \in \mathbb{R}$, $v_{\boldsymbol{x}}(a)$ is a continuous functions of $\boldsymbol{x}$.
Let $\mathcal{H}_1 = \{H_{1,\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$ be a $\mathcal{P}(\Delta_m)$-valued stochastic process such that:

(3.i) $\eta_1, \eta_2, \ldots,$ are independent and identically distributed real-valued stochastic processes of the form $\eta_i : \mathcal{X} \longrightarrow \mathbb{R}$, $i \geq 1$, with law indexed by a finite-dimensional parameter $\boldsymbol{\Psi}_1$.

(3.ii) $\theta_i \in \tilde{\Delta}_m$, $i \geq 1$, are independent random vectors with distribution $G_{1,0}$.

(3.iii) $k \in \mathbb{N}$ is a discrete random variable with distribution indexed by a finite-dimensional parameter λ_1.

(3.iv) For every $\boldsymbol{x} \in \mathcal{X}$, the density function of $H_{1,\boldsymbol{x}}$, w.r.t. Lebesgue measure, is given by a dependent mixture of Dirichlet densities,

$$h_{1,\boldsymbol{x}}(\boldsymbol{y}) \quad = \quad \sum_{\boldsymbol{j} \in \mathcal{H}^0_{k,m}} w_{1,k,\boldsymbol{j},G_{\boldsymbol{x}}} \times \mathrm{d}(\boldsymbol{y} \mid \alpha(k,\boldsymbol{j})), \tag{7}$$

where

$$w_{1,k,\boldsymbol{j},G_{\boldsymbol{x}}} = \left(\left\{ \sum_{l_1=1}^{\infty} v_{\boldsymbol{x}} \left\{ \eta_{l_1}(\boldsymbol{x}) \right\} \prod_{l_2<l_1} \left[1 - v_{\boldsymbol{x}} \left\{ \eta_{l_2}(\boldsymbol{x}) \right\} \right] \right\} I(\boldsymbol{\theta}_{l_1})_{A^1_{\boldsymbol{j},k}} \right).$$

The process $\mathcal{H}_1$ defined by (3.i)–(3.iv) is referred to as 'single-atoms' dependent MBP1 with parameters $(\lambda_1, \boldsymbol{\Psi}_1, \mathcal{V}, G_{1,0})$, and written

$$\mathcal{H}_1 \sim \theta\text{DMBP1}(\lambda_1, \boldsymbol{\Psi}_1, \mathcal{V}, G_{1,0}).$$

In a similar manner, let $\mathcal{H}_2 = \{H_{2,\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$ be a $\mathcal{P}(\Delta_m)$-valued stochastic process such that:

(4.i) $\eta_1, \eta_2, \ldots,$ are independent and identically distributed real-valued stochastic processes of the form $\eta_i : \mathcal{X} \longrightarrow \mathbb{R}$, $i \geq 1$, with law indexed by a finite-dimensional parameter $\boldsymbol{\Psi}_2$.

(4.ii) $\theta_i \in \tilde{\Delta}_m$, $i \geq 1$, are independent random vectors with distribution $G_{2,0}$.

(4.iii) $k \in \mathbb{K} = \{l \in \mathbb{N} : l^{1/2} \text{ is an integer}\}$ is a discrete random variable with distribution indexed by a finite-dimensional parameter λ_2.

(4.iv) For every $\boldsymbol{x} \in \mathcal{X}$, the density function of $H_{2,\boldsymbol{x}}$, w.r.t. Lebesgue measure, is given by a dependent mixture of Dirichlet densities,

$$h_{2,\boldsymbol{x}}(\boldsymbol{y}) \quad = \quad \sum_{\boldsymbol{j} \in \mathcal{H}^0_{k,m}} w_{2,k,\boldsymbol{j},G_{\boldsymbol{x}}} \times \mathrm{d}(\boldsymbol{y} \mid \alpha(k,\boldsymbol{j})), \tag{8}$$

where

$$w_{2,k,\boldsymbol{j},G_{\boldsymbol{x}}} = C(\boldsymbol{j}) \left(\left\{ \sum_{l_1=1}^{\infty} v_{\boldsymbol{x}} \left\{ \eta_{l_1}(\boldsymbol{x}) \right\} \prod_{l_2<l_1} \left[1 - v_{\boldsymbol{x}} \left\{ \eta_{l_2}(\boldsymbol{x}) \right\} \right] \right\} I(\boldsymbol{\theta}_{l_1})_{A^2_{\boldsymbol{j},k}} \right).$$

The process $\mathcal{H}_2$ defined by (4.i)–(4.iv) is referred to as 'single-atoms' dependent MBP2 with parameters $(\lambda_2, \boldsymbol{\Psi}_2, \mathcal{V}, G_{2,0})$, and written

$$\mathcal{H}_2 \sim \theta\text{DMBP2}(\lambda_2, \boldsymbol{\Psi}_2, \mathcal{V}, G_{2,0}).$$

Notice that the trajectories of both θDMBP1 and θDMBP2 are a.s. a density w.r.t. Lebesgue measure since, for every $\boldsymbol{x} \in \mathcal{X}$,

$$\sum_{i=1}^{\infty} \log \left[1 - E \left(v_{\boldsymbol{x}} \left\{ \eta_i(\boldsymbol{x}) \right\} \right) \right] = -\infty.$$

4.3 Continuity properties and the association structure

Assume that in the definition of θDMBP1 and θDMBP2, for every $j \in \mathbb{N}$, the stochastic process η_j is a.s. continuous. Then, for every $\{\boldsymbol{x}_j\}_1^\infty \subset \mathcal{X}$, such that $\lim_{j\to+\infty} \boldsymbol{x}_j \longrightarrow \boldsymbol{x}_0 \in \mathcal{X}$,

$$\lim_{j\to+\infty} \sup_{B\in\mathbb{B}(\Delta_m)} |H_{1,\boldsymbol{x}_j}(B) - H_{1,\boldsymbol{x}_0}(B)| = 0, \text{ a.s.},$$

and

$$\lim_{j\to+\infty} \sup_{B\in\mathbb{B}(\Delta_m)} |H_{2,\boldsymbol{x}_j}(B) - H_{2,\boldsymbol{x}_0}(B)| = 0, \text{ a.s.},$$

for every $\boldsymbol{x}_0 \in \mathcal{X}$, that is, $H_{1,\boldsymbol{x}_j}$ converges a.s. to $H_{1,\boldsymbol{x}_0}$ and $H_{2,\boldsymbol{x}_j}$ converges a.s. to $H_{2,\boldsymbol{x}_0}$, in total variation norm, as $\boldsymbol{x}_j \longrightarrow \boldsymbol{x}_0$.
If the θDMBP1 and θDMBP2 are defined such that, for every $\{\boldsymbol{x}_j\}_1^\infty \subset \mathcal{X}$, such that $\lim_{j\to+\infty} \boldsymbol{x}_j \longrightarrow \boldsymbol{x}_0 \in \mathcal{X}$, we have $\eta_i(\boldsymbol{x}_j) \xrightarrow{\mathcal{L}} \eta_i(\boldsymbol{x}_0)$, as $j \longrightarrow +\infty$, then, for all $\boldsymbol{y} \in \Delta_m$,

$$\lim_{j\to+\infty} \rho\left[H_{1,\boldsymbol{x}_j}(B_{\boldsymbol{y}}), H_{1,\boldsymbol{x}_0}(B_{\boldsymbol{y}})\right] = 1,$$

and

$$\lim_{j\to+\infty} \rho\left[H_{2,\boldsymbol{x}_j}(B_{\boldsymbol{y}}), H_{2,\boldsymbol{x}_0}(B_{\boldsymbol{y}})\right] = 1,$$

where $\rho(A, B)$ denotes the Pearson correlation between A and B, and $B_{\boldsymbol{y}} = [0, y_1] \times \ldots \times [0, y_m]$. Furthermore, the θDMBP1 and θDMBP2 can be specified such that

$$\lim_{j\to+\infty} Cov\left[H_{1,\boldsymbol{x}_{1j}}(B_{\boldsymbol{y}}), H_{1,\boldsymbol{x}_{2j}}(B_{\boldsymbol{y}})\right] = 0,$$

and

$$\lim_{j\to+\infty} Cov\left[H_{2,\boldsymbol{x}_{1j}}(B_{\boldsymbol{y}}), H_{2,\boldsymbol{x}_{2j}}(B_{\boldsymbol{y}})\right] = 0,$$

when $||\boldsymbol{x}_{1j} - \boldsymbol{x}_{2j}|| \longrightarrow \infty$.
Now, assume that for every $\{(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j})\}_1^\infty \subset \mathcal{X}^2$, such that $\lim_{j\to+\infty}(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}) = (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathcal{X}^2$, the θDMBP1 and θDMBP2 are defined such that

$$(\eta_i(\boldsymbol{x}_{1j}), \eta_i(\boldsymbol{x}_{2j})) \xrightarrow{\mathcal{L}} (\eta_i(\boldsymbol{x}_1), \eta_i(\boldsymbol{x}_2)),$$

as $j \longrightarrow +\infty$. Then, for every $\boldsymbol{y} \in \Delta_m$,

$$\lim_{j\to\infty} \rho\left[H_{1,\boldsymbol{x}_{1j}})(B_{\boldsymbol{y}}), H_{1,\boldsymbol{x}_{2j}}(B_{\boldsymbol{y}})\right] = \rho\left[H_{1,\boldsymbol{x}_1}(B_{\boldsymbol{y}}), H_{1,\boldsymbol{x}_2}(B_{\boldsymbol{y}})\right],$$

and

$$\lim_{j\to\infty} \rho\left[H_{2,\boldsymbol{x}_{1j}})(B_{\boldsymbol{y}}), H_{2,\boldsymbol{x}_{2j}}(B_{\boldsymbol{y}})\right] = \rho\left[H_{2,\boldsymbol{x}_1}(B_{\boldsymbol{y}}), H_{2,\boldsymbol{x}_2}(B_{\boldsymbol{y}})\right].$$

4.4 The support properties

If for every $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}^n$, $n \geq 1$, the joint distribution of

$$(\eta_i(\boldsymbol{x}_1), \ldots, \eta_i(\boldsymbol{x}_n))$$

has full support on $\mathbb{R}^n$, and, k and $G_{i,0}$ have full support, then $\mathcal{P}(\Delta_m)^{\mathcal{X}}$ and $\mathcal{D}(\Delta_m)^{\mathcal{X}}$ are the support of the θDMBP1 and θDMBP2 under the weak product topology and the L_∞ product topology, respectively. If, in addition, $\mathcal{X}$ is a compact set, and θDMBP1 and θDMBP2 are defined such that for any $\epsilon > 0$ and $[0,1]$–valued continuous function f defined on $\mathcal{X}$, we have that

$$P\left\{\sup_{\boldsymbol{x}\in\mathcal{X}} |v_{\boldsymbol{x}}(\eta_i(\boldsymbol{x})) - f(\boldsymbol{x})| < \epsilon\right\} > 0,$$

then $\tilde{\mathcal{D}}(\Delta_m)^{\mathcal{X}}$ is the support of θDMBP1 and θDMBP2 under the L_∞ topology, and

$$P\left\{\sup_{\boldsymbol{x}\in\mathcal{X}} \int_{\Delta_m} q_{\boldsymbol{x}}(\boldsymbol{y}) \log\left(\frac{q_{\boldsymbol{x}}(\boldsymbol{y})}{h_{1,\boldsymbol{x}}(\boldsymbol{y})}\right) d\boldsymbol{y} > \epsilon\right\} > 0,$$

and

$$P\left\{\sup_{\boldsymbol{x}\in\mathcal{X}} \int_{\Delta_m} q_{\boldsymbol{x}}(\boldsymbol{y}) \log\left(\frac{q_{\boldsymbol{x}}(\boldsymbol{y})}{h_{2,\boldsymbol{x}}(\boldsymbol{y})}\right) d\boldsymbol{y} > \epsilon\right\} > 0,$$

for every $\epsilon > 0$ and every $\{Q_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\} \in \tilde{\mathcal{D}}(\Delta_m)$, with density functions $\{q_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$.

4.5 The frequentist asymptotic behaviour

It is possible to show that the posterior distribution associated with the random joint distribution induced by θDMBP1 and θDMBP2, $m_1(\boldsymbol{y}, \boldsymbol{x}) = q(\boldsymbol{x})h_{1,\boldsymbol{x}}(\boldsymbol{y})$ and $m_2(\boldsymbol{y}, \boldsymbol{x}) = q(\boldsymbol{x})h_{2,\boldsymbol{x}}(\boldsymbol{y})$, respectively, where q is the density generating the predictors, is weakly consistent at any joint distribution of the form $m_0(\boldsymbol{y}, \boldsymbol{x}) = q(\boldsymbol{x})q_0(\boldsymbol{y} \mid \boldsymbol{x})$, where $\{q_0(\cdot \mid \boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\} \in \widetilde{\mathcal{D}}(\Delta_m)^{\mathcal{X}}$.
Now, if $\mathcal{X} = [0,1]^p$, and the θDMBP1 and θDMBP2 are defined such that:

(5.i) $\theta_1, \theta_2, \ldots$, have full support on $\tilde{\Delta}_m$.

(5.ii) $\eta_j(\boldsymbol{x}) = \eta_0(A_j^{1/2}\boldsymbol{x})$ for each $\boldsymbol{x} \in \mathcal{X}$, where η_0 is a base Gaussian process with covariance kernel $c_0(\boldsymbol{x}, \boldsymbol{x}') = \tau^2 \exp\{-\|\boldsymbol{x} - \boldsymbol{x}'\|^2\}$, and that there exits $\kappa, \kappa_0 > 0$ and a sequence $\delta_n = O((\log n)^2/n^{5/2})$ such that $P\{A_j < \delta_n\} \leq \exp\{-n^{-\kappa_0} j^{(\kappa_0+2)/\kappa} \log j\}$ for each $j \geq 1$. In addition, assume that there exists a sequence $r_n \uparrow \infty$ such that $r_n^p n^\kappa (\log n)^{p+1} = o(n)$ and $P\{A_n > r_n\} \leq \exp\{-n\}$.

(5.iii) for every $v_{\boldsymbol{x}} \in \mathcal{V}$, $v_{\boldsymbol{x}} \equiv \Phi$, where Φ is the CDF of a standard normal distribution.

(5.iv) k has full support.

(5.v) there exists a sequence $k_n \in \mathbb{N}$ such that $\log((k_n + m - 1)!(k_n - 1)!^{-1}) \preceq O(n)$ and $P\{k(\omega) > k_n\} \preceq O(\exp\{-n\})$, where $\preceq$ stands for inequality up to a constant multiple or if the constant multiple is irrelevant to the given situation.

Then the posterior distribution associated with the random joint distribution induced by the θDMBP1 and θDMBP2 models $m_1(\boldsymbol{y}, \boldsymbol{x})$ and $m_2(\boldsymbol{y}, \boldsymbol{x})$, respectively, is L_1-consistent at any joint distribution of the form $m_0(\boldsymbol{y}, \boldsymbol{x}) = q(\boldsymbol{x})q_0(\boldsymbol{y} \mid \boldsymbol{x})$, where $\{q_0(\cdot \mid \boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\} \in \widetilde{\mathcal{D}}(\Delta_m)^{\mathcal{X}}$.

5 Concluding remarks

We have proposed novel classes of probability models for single probability distributions and sets of predictor-dependent probability distributions for data supported on Δ_m. The proposal corresponds to extensions of the Dirichlet-Bernstein and dependent stick-breaking-Bernstein priors.
The proposed classes have appealing theoretical properties such as full support, continuity, known marginal distribution, well behaved correlation function, and its posterior distribution is consistent. The proposed models can be fit using standard MCMC algorithms for Dirichlet process based models.

Acknowledgments: The first author was supported by Fondecyt 1141193 grant. The second author was supported by Fondecyt 3130400 grant.

References

Barrientos, A. F. and Jara, A. (2014). Posterior convergence rate of a class of Dirichlet process mixture models for compositional data. *Technical report, Department of Statistics, Pontificia Universidad Católica de Chile.*

Barrientos, A. F., Jara, A., and Quintana, F. (2014). Bayesian density estimation for compositional data using random Bernstein polynomials. *Technical report, Department of Statistics, Pontificia Universidad Católica de Chile.*

Petrone, S. (1999). Random Bernstein polynomials. *Scandinavian Journal of Statistics*, **26**, 373 – 393.

Tenbusch, A. (1994). Two-dimensional Bernstein polynomial density estimators. *Metrika*, **41**, 233 – 253.

Zheng, Y., Zhu, J. and Roy, A. (2010). Nonparametric Bayesian inference for the spectral density function of a random field. *Biometrika*, **97**, 238 – 245.

Consistent estimation of mixed models by ignoring non-ignorable missingness

Sophia Rabe-Hesketh[1], Anders Skrondal [1]

[1] Graduate School of Education, University of California, Berkeley, USA

E-mail for correspondence: sophiarh@berkeley.edu

Abstract: In longitudinal data, maximum likelihood estimators of mixed-effects model parameters are consistent if missingness depends only on the covariates. Missingness can also depend on observed outcomes, under correct specification of the covariance structure, but this result is useful only under monotone missingness. When missingness depends on unobserved outcomes or on the random effects, it is said to be not missing at random (NMAR). For such NMAR missingness, joint modeling of the outcomes and missingness has been advocated, but these approaches are known to rely on unverifiable assumptions. In this talk, I will consider methods that ignore NMAR missingness but are consistent for (some of) the parameters of interest. An example of such "protective" estimators are conditional maximum likelihood estimators, also known as fixed-effects estimators. For binary data, fixed-effects approaches can be used to obtain consistent estimators of regression coefficients under a wide range of NMAR mechanisms (Skrondal and Rabe-Hesketh, 2014, Biometrika 101, 175-188). Other protective estimators for binary and continuous outcomes will also be discussed in this talk.

Keywords: mixed models; missing observations

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

Regularization and Sparsity in Discrete Structures

Gerhard Tutz[1]

[1] Ludwig-Maximilians-Universität München, Germany

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

Abstract: Modeling categorical data with many categories in the predictor or response is a challenge because many parameters are needed to specify the link between predictors and responses. An attractive way to reduce the complexity of the estimation problem is regularization by structured penalties. Penalization has been well investigated for metric predictors but categorical data call for penalty terms that are tailored to the categorical nature of the involved variables. In particular one should distinguish between ordered and un-ordered categorical predictors and allow for appropriate clustering of categories. In addition to tailored penalty terms for cross sectional data we consider regularized estimators for repeated measurements. The considered fixed effects models allow to model the heterogeneity of the population and represent an alternative to the widely used random effects models. As an alternative to penalization tree-based estimators are considered to obtain clusters of categories in high dimensional problems.

Keywords: Regularization; Categorical data; Fixed Effects models; Fusion.

1 Introduction

With the introduction of the lasso (Tibshirani, 1996) regularization methods for regression and classification have become a topic of intensive research. Regularization methods aim at a sparse representation of the link between predictors and responses. Only those components should be included in the model that are really needed to model the effect of explanatory variables on an outcome variable. In particular categorical variables are a challenge to sparsity because typically one needs at least one parameter for each category. As a consequence, if the number of categories in a predictor is large maximum likelihood estimates tend to fail, they are not unique or deteriorate. If the outcome variable is categorical the problems increase because for each predictor variable one needs a parameter linked to the categories of the outcome variable.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

Another feature of categorical data is that other structures than for metric data are of interest. While regularization for metric predictors often means variable selection and therefore identification of parameters that should be set to zero, for a categorical predictor one also wants to know which categories have to be distinguished when modelling the effect on an outcome variable. Therefore one wants to identify clusters of categories that share the same effect. Clustering is not restricted to predictor variables, it also is a challenge in categorical outcomes and in the modelling of subject-specific effects.

In the following we will consider methods of feature extraction for discrete structures. One method is regularization or constrained estimation by use of penalty terms in the tradition of the lasso. We will also consider tree-based methods, which have advantages in the case of very large numbers of categories. Boosting, which is usually an efficient tool to extract information in regression problems (Friedman et al, 2000), will be neglected because it is less appropriate for discrete structures.

2 Penalized Regression

In generalized linear models (GLMs) the conditional response $\mu = E(y|\boldsymbol{x})$ is specified by

$$\mu = h(\eta) \quad \text{or} \quad g(\mu) = \eta$$

where $h(.)$ denotes the response function and $g(.) = h(.)^{-1}$ the link function. The linear predictor is determined by the predictors $\boldsymbol{x}$ in the form $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$. In addition, $y|\boldsymbol{x}$ follows a simple exponential distribution (McCullagh & Nelder, 1989).

Regularization methods that use penalty terms are obtained by maximizing the *penalized log-likelihood*

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \lambda J(\boldsymbol{\beta}),$$

where $l(\boldsymbol{\beta})$ is the usual log-likelihood of the GLM, λ is a tuning parameter, and $J(\boldsymbol{\beta})$ is a functional that penalizes the size and structure of the parameters. A classical penalty is the ridge penalty $\sum_j \beta_j^2$, which goes back to Hoerl and Kennard (1970). It shrinks estimates toward zero and is able to stabilize estimates but unable to detect structures in the predictor. In the following alternative penalty terms are considered that enforce the detection of interesting structures in discrete data.

3 Selection and Clustering for Categorical Predictors

Let categorical predictors $C_j, j = 1, \ldots, p$, have values $C_j \in \{0, \ldots, k_j\}$. They can be included into the model by using dummy variables defined by $x_{jr} = 1$ if $C_j = r$ and $x_{jr} = 0$ otherwise, yielding the linear predictor

$$\eta = \sum_{j=1}^{p} \sum_{r=1}^{k_j} x_{jr}\beta_{jr} + \boldsymbol{z}^T\boldsymbol{\gamma} = \sum_{j=1}^{p} \boldsymbol{x}_j^T\boldsymbol{\beta}_j + \boldsymbol{z}^T\boldsymbol{\gamma},$$

where $\boldsymbol{z}$ is a vector of additional variables with weight vector $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_j^T = (\beta_{j1}, \ldots, \beta_{jk_j})$ collects all parameters linked to variable C_j. The predictor C_j adds k_j parameters, the total number of parameters contributed by the categorical predictors is $k_1 + \cdots + k_p$, which can be very large, in particular, if several categorical predictors are included.
In selection problems for categorical predictors it should be distinguished between two problems:

- Which categorical predictors should be included in the model?
- Which categories within one categorical predictor should be distinguished?

For the first problem, namely variable selection, Yuan & Lin (2006) proposed the *group lasso*, which uses the penalty

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sqrt{k_j} ||\boldsymbol{\beta}_j||_2, \tag{1}$$

where $||\boldsymbol{\beta}_j||_2 = (\beta_{j1}^2 + \cdots + \beta_{jk_j}^2)^{1/2}$ is the L_2-norm of the parameters of the jth group. The penalty encourages sparsity in the sense that either $\hat{\boldsymbol{\beta}}_j = \mathbf{0}$ or $\beta_{jr} \neq 0$ for $r = 1, \ldots, k_j$. Thus, it aims at variable selection in contrast to parameter selection. Meier et al. (2008) showed that under sparsity the resulting estimates are consistent even when the number of predictors is larger than the sample size. The penalty selects predictors but typically, if a predictor is in the model, all the parameter estimates differ and no clustering is obtained.
A penalty that enforces the building of clusters of categories that share the same effect is

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{r<s} w_{rs}^{(j)} |\beta_{jr} - \beta_{js}|, \tag{2}$$

where the sum is over all categories $r, s \geq 0$ and implicitly the reference category zero has been chosen by setting $\beta_{j0} = 0$. The $w_{rs}^{(j)}$ are additional, appropriately chosen weights. By using the L_1-penalized differences between all pairs of parameters that are linked to one categorical predictor the penalty tends to form clusters of categories that have the same effect. Since the parameter for the reference category ($\beta_{j0} = 0$) is included in the sum the penalty also enforces variable selection. In the extreme case, for $\lambda \to \infty$, all parameter estimates become zero and the categorical predictors are excluded.
The penalties (1) and (2) can be recommended for *nominal predictors* only. For *ordinal categorical predictors* they ignore the information contained in the ordering of categories. With the focus on selection the group lasso penalty should be replaced by

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sqrt{k_j} ||\boldsymbol{D}_j \boldsymbol{\beta}_j||_2,$$

where $\boldsymbol{D}_j$ is a matrix that generates differences of fixed order from the parameters linked to the jth predictor. In the simplest case of order one differences one obtains $\boldsymbol{D}_j\boldsymbol{\beta}_j = \sum_r |\beta_{jr} - \beta_{j,r-1}|$. The penalty enforces selection of whole groups of parameters and simultaneously smoothes over the ordered categories.

If the objective is the identification of clusters of categories it is natural to assume for ordered predictors that clusters of categories refer to adjacent categories. Thus the penalty should enforce the fusion of adjacent categories, which is obtained by using

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{r=1}^{k_j} w_r^{(j)} |\beta_{jr} - \beta_{j,r-1}|$$

with corresponding weights $w_r^{(j)}$. The effect of the penalty is that one obtains step functions for the ordered predictor, categories that have the same effect are fused. For nominal categorical predictors the fusion type penalties were considered by Bondell & Reich (2009). Gertheiss & Tutz (2010) and Tutz & Gertheiss (2014) considered clustering of categories for nominal and ordinal predictors.

As an illustrative example we consider the Munich rent data. The data set consists of 2053 housholds with the response variable being monthly rent per square meter in Euro. Available predictors are the urban district (nominal factor), the year of construction, the number of rooms, the quality of residential area (ordinal factors), floor space (metric) and five additional binary variables, hot water supply available, central heating available, tiled bathroom (yes/no), for details, see Gertheiss & Tutz (2010). For illustration we show the coefficient paths for the two predictors urban district and year of construction.

It should be noted that the given penalty terms can be seen as basic components to obtain sparsity in terms of variables and clusters. In applications they can apply to main effects but also to interaction terms. Moreover, it is often useful to combine several penalties including simple smoothing penalties as the ridge or extended ridge penalties. An exemplary complex modeling problem that calls for combinations of penalties is discrete survival. With discrete time $T \in \{1, \ldots, K\}$ the hazard is defined by the conditional probability $\lambda(r|\boldsymbol{x}) = P(T = r|T \geq r)$ and models have the form $g(\lambda(r|\boldsymbol{x})) = \beta_{0r} + \sum_j x_j\beta_{jr}$, where β_{jr} are time-varying coefficients. Estimation of discrete survival can be obtained by considering the conditional survival of a given time, that is, the event $T > r|T \geq r$ as a binary event. Then discrete time is an ordered categorical predictor and the time-varying effects can be seen as an interaction between the x-predictors and time. A useful penalty for this modelling problem is

$$J(\boldsymbol{\beta}) = \sum_r (\beta_{0r} - \beta_{0,r-1})^2 + \sum_{j=1}^{p} \sum_r |\beta_{jr} - \beta_{j,r-1}|.$$

The first term is a generalized ridge penalty that smoothes the baseline

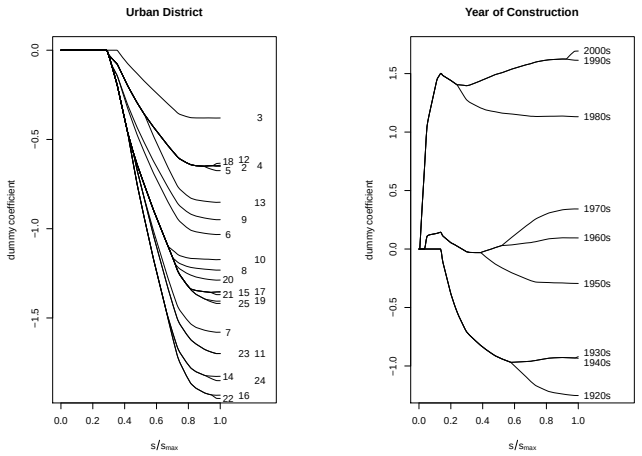


FIGURE 1. Paths of dummy coefficients for urban district and year of construction (rent data).

hazard, the second term reduces the number of parameters to be estimated by assuming that the time-varying effect of covariates are constant over adjacent categories.

4 Categorical Responses

For categorical responses variable regularization to obtain sparsity has to be adapted to the multivariate nature of the response. The classical model for response categories $Y \in \{1, \ldots, K\}$ is the multinomial model, which can be seen as a multivariate GLM. In its generic form it specifies

$$\pi_r = P(Y = r|\boldsymbol{x}) = \frac{exp(\beta_{r0} + \boldsymbol{x}^T\boldsymbol{\beta}_r)}{\sum_{s=1}^{k} exp(\beta_{s0} + \boldsymbol{x}^T\boldsymbol{\beta}_s)} = \frac{\exp(\eta_r)}{\sum_{s=1}^{k} \exp(\eta_s)}, \qquad (3)$$

where $\boldsymbol{\beta}_r^T = (\beta_{r1}, \ldots, \beta_{rp})$. Since parameters $\beta_{10}, \ldots, \beta_{K0}$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are not identifiable, additional constraints are needed. Typically, one of the response categories is chosen as reference category, for example, by setting $\beta_{K0} = 0$, $\boldsymbol{\beta}_K = \mathbf{0}$.
In the multinomial logit model the effect of covariates is specified by the linear predictors $\eta_r,\ r = 1, \ldots, K-1,$, which correspond to the log odds

between category r and the reference category k. We will consider a more general version of the model that allows for category-specific variables. For example, when the response is the choice of the transportation mode, the potential attributes are price and duration, which vary across the alternatives and therefore are category-specific. Then, in addition to the global predictors $\boldsymbol{x}$, a set of category-specific predictors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ is available, where $\boldsymbol{w}_r$ contains the attributes of category r. The set of linear predictors generalizes to

$$\eta_{ir} = \beta_{r0} + \boldsymbol{x}^T\boldsymbol{\beta}_r + (\boldsymbol{w}_{ir} - \boldsymbol{w}_{iK})^T\boldsymbol{\alpha}, \qquad r = 1, \ldots, K-1. \tag{4}$$

The second term specifies the effect of the global variables, and the third term specifies the effect of the difference $\boldsymbol{w}_{ir} - \boldsymbol{w}_{iK}$ on the choice between category r and the reference category. In the choice of the transport mode it can be the difference in price that has an effect on the choice.

If one uses the simple lasso, which penalizes all parameters by a sum over $|\beta_{rj}|$ (Friedman et al, 2010) one does not obtain variable selection but parameter selection because if one of the parameters β_{rj} is not deleted the whole variable x_j is still in the model. To select variables one has to group all the parameters that correspond to one variable and penalize them simultaneously. This is obtained by the categorically structured (CATS) penalty

$$J(\boldsymbol{\theta}) = \psi \sum_{j=1}^{p} \phi_j \, ||\boldsymbol{\beta}_{.j}|| \; + \; (1-\psi) \sum_{l=1}^{L} \varphi_l \, |\alpha_l|, \tag{5}$$

where $\boldsymbol{\beta}_{.j}^T = (\beta_{1j}, \ldots, \beta_{K-1,j})$ collects all parameters linked to predictor x_j, ψ is an additional tuning parameter that balances the penalty on the global and the category-specific variables. The parameters ϕ_j and φ_l are weights that assign different amounts of penalization to different parameter groups. Typically they are chosen by $\phi_j = \sqrt{K-1}$ and $\varphi_l = 1$.

The penalty enforces variable selection, that is, all the parameters in $\boldsymbol{\beta}_{.j}$ are simultaneously shrunk toward zero. It is strongly related to the classical group lasso (Yuan & Lin, 2006). However, in the group lasso the grouping refers to the parameters that are linked to a categorical predictor within a univariate regression model whereas in the present model grouping arises from the multivariate response structure. Preliminary versions of the penalty have been considered by Tutz (2012), Tutz et al. (2012) and Simon et al. (2013).

5 Subject-Specific Models

In this section models for repeated measurements are considered. For repeated measurements penalty methods provide an alternative to random effects models with good performance in terms of estimation accurracy. Repeated measurements can be represented by $(y_{ij}, \boldsymbol{x}_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, where y_{ij} denotes the response of unit i at measurement occasion j, and $\boldsymbol{x}_{ij}$ is a vector of covariates that potentially varies across measurements. A common approach to model heterogeneity across units is by

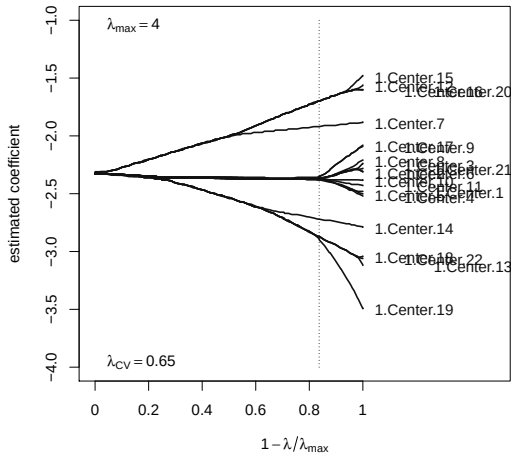


FIGURE 2. Coefficient paths for the beta blocker data. The very right end of the figure relates to ML estimates; that is, to $\lambda = 0$. The left end relates to the minimal value of λ giving maximal penalization; in this case $\lambda = 4$.

random effects models. In *generalized linear mixed effects model* (GLMM), the structural assumption specifies that the conditional means, $\mu_{ij} = E(y_{ij}|b_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$, have the form

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i \tag{6}$$

where g is a monotonic and continuously differentiable link function, $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ is a linear parametric term with parameter vector $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$ that includes an intercept, and $\boldsymbol{z}_{ij}$ is a covariate vector associated with random effects. The second term contains the random effects that model the heterogeneity of the units. For the random effects, one assumes a distributional form, typically a normal distribution, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{Q})$.

The focus of the random effects models is on the fixed effects; the distribution of the random effects is mainly used to account for the heterogeneity of the units. Although it is the most popular model that accounts for heterogeneity, it has several drawbacks. The assumption of a specific distribution for the random effects may affect the inference. In particular, if the distributional assumption is far from the data generating distribution, inference can be strongly biased. Moreover, assuming a continuous distribution prevents that the effects of units can be the same. Therefore, by assumption, no clustering of units is available. One further aspect is that it is assumed that the random effects and the covariates observed per second level unit are independent; a criticism that has a long tradition, in particular in the econometric literature. For an overview on the choice between fixed and random effects models, see, Townsend et al (2013).

As an alternative we consider the *fixed effect* or *subject-specific model*

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{\beta}_i \tag{7}$$

The model specifies that each unit has its own coefficient $\boldsymbol{\beta}_i, i = 1, \ldots, n$. The problem with these models is that the large number of parameters can render the estimates unstable and encourage overfitting. Typically, there is not enough information available to distinguish among all the units; but under the assumption that observations form clusters with respect to their effect on the response, the number of parameters can be reduced and estimates are available. The tool to obtain sparsity of subject-specific parameters and clusters is the use of the penalty

$$J(\boldsymbol{\beta}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n) = \sum_{r>m} ||\boldsymbol{\beta}_r - \boldsymbol{\beta}_m||. \tag{8}$$

If $\lambda = 0$, one obtains the unpenalized estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ and each unit has its own parameter. If $\lambda \to \infty$, the penalty enforces that the estimates of all subject-specific parameters are the same. It has been demonstrated in Tutz & Oelker (2014) that the method outperforms the random effects model, in particular if correlation between the random intercept and the random effect, the so-called level 2 endogeneity, is present. The model is also compared to alternative approaches as the discrete mixture model (Aitkin, 1999).

As an example we consider the modeling of the effect of beta blockers on the mortality after myocardial infarction, see also Aitkin (1999). In a 22-center clinical trial, for each center, the number of deceased/successfully treated patients in control/test groups was observed. The binary response (1 = deceased/0 = not deceased) suggests a mixed logit model of the form

$$\text{logit } P(y_{ij} = 1) = \beta_0 + \beta_{i0} + \beta_T \cdot \text{Treatment}_{ij}, \quad i = 1, \ldots, 22 \text{ Centers}, \tag{9}$$

where Treatment_{ij} codes the treatment in hospital i for patient j. If β_{i0} is replaced by a random effect b_i with normal distribution, implicitly the hospitals are considered as a random sample and all the effects of hospitals are assumed to differ. In contrast, the fixed effect model with regularization assumes that some of the hospitals have the same effect. Figure 2 shows the coefficient built-ups against regularization, where the vertical line refers to the cross-validated choice of the tuning parameter. It is seen that there seem to be essentially five clusters of hospitals which share the same strength.

6 Tree-Based Approaches

Penalization is a useful tool to identify relevant categorical predictors and clusters of categories but becomes computationally demanding if the number of categories is very large. In this case approximations or alternative procedures have to be used. One alternative that is considered in the following is based on trees. The big advantage of classical trees or recursive partitioning procedures as CART (Breiman et al, 1984) and C4.5 (Quinlan, 1993) is that they automatically find interactions. The concept of interactions is at the core of recursive partitioning. But the focus on interactions

can also turn into a disadvantage because common trees do not allow for a linear or smooth component in the predictor. Below the root node most nodes represent interactions with the effect that main effects and potentially linear or additive effects of covariates are neglected. That means for categorical predictors that typically interactions are fitted but no clustering of main effect is detected. A version of trees that is able to detect clusters are structured trees, which fit the predictor

$$\eta = tr(C) + \boldsymbol{z}^T\boldsymbol{\gamma},$$

where $tr(C)$ is the tree component of the predictor and $\boldsymbol{z}^T\boldsymbol{\gamma}$ is the familiar linear term containing further variables. For simplicity we start with one categorical predictor $C \in \{1, \ldots, k\}$ in the tree component. When a tree is built, successively a node A, that is, a subset of the predictor space, is split into subsets with the split determined by only one variable. For a *nominal* categorical variable $C \in \{1, \ldots, k\}$, the partition has the form $A \cap S, \quad A \cap \bar{S}$, where S is a non-empty subset $S \subset \{1, \ldots, k\}$ and $\bar{S} = \{1, \ldots, k\} \setminus S$ is the complement. Thus, after several splits the predictor $tr(C)$ represents a clustering of the categories $\{1, \ldots, k\}$, and the tree term can be represented by

$$tr(C) = \alpha_1 1_{S_1}(C) + \cdots + \alpha_m 1_{S_m}(C),$$

where $S_1, \ldots, S_m$ is a partition of $\{1, \ldots, k\}$, and $1_S(.)$ denotes the indicator function, $1_S(C) = 1$ if $C \in S$, $1_S(C) = 0$, otherwise. For an *ordinal* categorical variable $C \in \{1, \ldots, k\}$ the partition into two subsets has the form $A \cap \{C \leq c\}, \quad A \cap \{C > c\}$, based on the threshold c on variable C. Thus during the building of a tree clusters of adjacent categories are formed.

In the case of more than one categorical predictor trees form clusters that combine the predictors. Thus the subsets do not refer to a single variable. A *structured tree*, which is proposed here forces the tree to form clusters only for one variable. Then, with p predictors $C_1, \ldots, C_p$, $C_j \in \{1, \ldots, k_j\}$, the tree component has the form

$$tr(C_1, \ldots, C_p) = tr(C_1) + \cdots + tr(C_p),$$

where $tr(C_r)$ is the tree for the rth variable, that means it represents clusters of the rth variable with the cluster form determined by the scale level of the corresponding variable. A traditional tree hardly finds clusters for single components. It typically produces clusters that combine several variables, in particular, mixing nominal and ordinal predictors.

For an ordinal categorical predictor clusters of adjacent categories can be found by fitting a model with predictor

$$\eta = \alpha_l I(C \leq c) + \alpha_r I(C > c) + \boldsymbol{z}^T\boldsymbol{\gamma},$$

where $I(.)$ denotes the indicator function. By use of the split-point c the model splits the predictor space into two regions, $C \leq c$ and $C > c$ yielding

two clusters of adjacent categories. An equivalent representation of the predictor, which is more familiar from the fitting of linear terms, is

$$\eta = \beta_0 + \alpha I(z > c) + \boldsymbol{z}^T\boldsymbol{\gamma}.$$

with the transformation of parameters given by $\beta_0 = \alpha_l$ and $\alpha = \alpha_r - \alpha_l$. For a nominal predictor $C \in \{1, \ldots, k\}$ splitting is much harder because one has to consider all possible partitions that contain two subsets. That means one has $2^{k-1} - 1$ candidates for splitting. But it has been shown that for regular trees it is not necessary to consider all possible partitions. One simply orders the predictor categories by increasing mean of the outcome and then splits the predictor as if it were an ordered predictor. It has been shown that this gives the optimal split in terms of various split measures, see Breiman et al (1984), Ripley (1996).
The fitting of a structured tree is a forward strategy that includes estimation and selection steps. The basic form is the following.

Structured Tree

Step 1 (Initialization)

(a) Estimation: Fit the candidate GLMs with predictors

$$\eta = \beta_0 + \alpha_{ij} I(C_i > c_{ij}) + \boldsymbol{z}^T\boldsymbol{\gamma}, \quad i = 1, \ldots, p, j = 1, \ldots, k_i$$

(b) Selection
Select the model that has the best fit. Let $c^*_{i_1,j_1}$ denote the best split, which is found for variable C_{i_1}. That means that $c^*_{i_1,j_1}$ is from the set of possible splits for C_{i_1}.

Step 2 (Iteration)

For $l = 1, 2, \ldots,$

(a) Estimation: Fit the candidate models with predictors

$$\eta = \beta_0 + \sum_{s=1}^{l} \alpha_{i_s,j_s} I(C_{i_s} > c^*_{i_s,j_s}) + \alpha_{ij} I(C_i > c_{ij}) + \boldsymbol{z}^T\boldsymbol{\gamma},$$

for all i and all values $c_{ij} \in C_i$ that have not been selected in previous steps.

(b) Selection
Select the model that has the best fit yielding the new cut point $c^*_{i_{l+1},j_{l+1}}$ that is found for variable $C_{i_{l+1}}$.

In the sequence of selected cut-points $c^*_{i_1,j_1}, c^*_{i_2,j_2}, \ldots$ and corresponding estimates $\hat{\alpha}_{i_1,j_1}, \hat{\alpha}_{i_2,j_2}, \ldots$ the first index refers to the variable and the

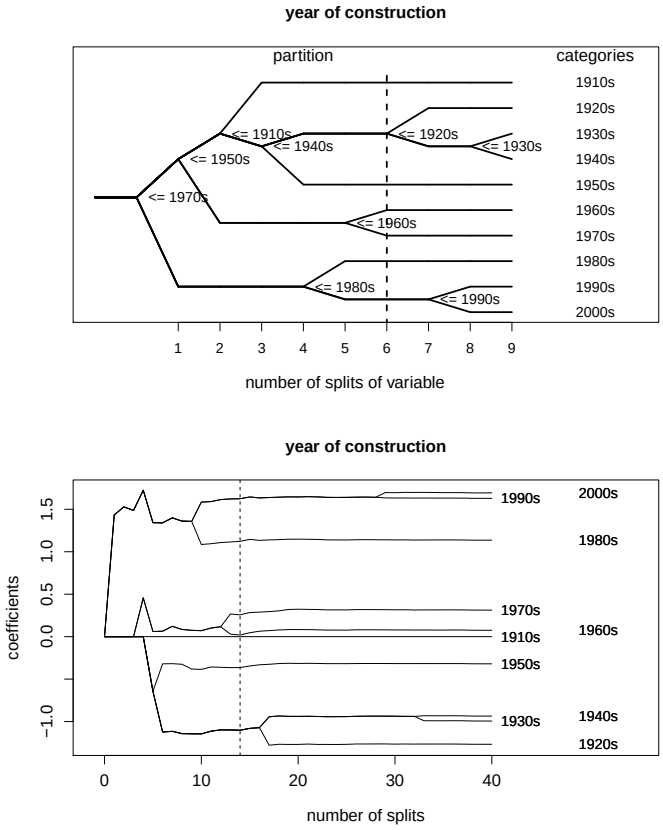


FIGURE 3. Results for the ordinal predictor year of construction for the analysis of the Munich rent standard data. Upper panel: resulting tree for year of construction, lower panel: paths of coefficients against all splits.

second to the split for this variable. As always in forward procedures one has to specify a stopping criterion, which can be cross-validation or in our case of fitting trees test statistics that evaluate if a further split is warranted.

For illustration we consider again the Munich rent data, where one has one nominal predictor (urban district), three ordinal predictors (year of construction in decades, number of rooms, quality of residential area), one metric variable (floor space) and five binary variables. In the additive part of the structured tree we model the effect of the metric predictor by cubic regression splines and include the binary variables in a linear form. The fusion of categories obtained by the tree is illustrated for the predictor year of construction. Figure 3 shows the resulting tree and the coefficient paths over the splits for the predictor decade of construction. The upper panel shows the successive splits against the number of splits in this predictor.

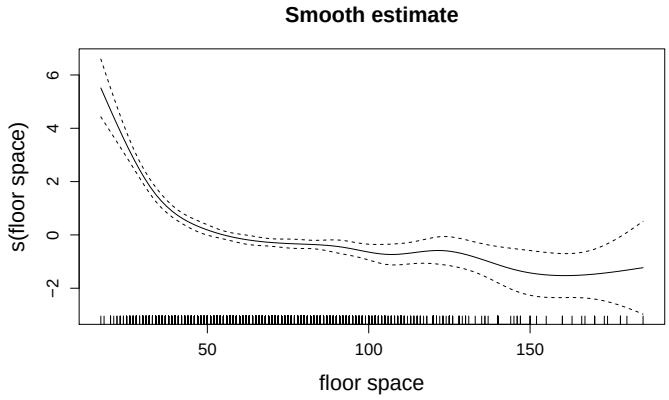


FIGURE 4. Resulting function of the smooth estimation of predictor floor space of the Munich rent data in the additive part of the structured tree.

The lower panel shows the coefficients plotted against the splits in all of the predictors. It is seen, in particular from the first steps, that estimates can change when other variables are included. But after about 14 splits the estimates are very stable. Model selection by p-values with significance level 0.05 yields seven clusters marked by the dashed lines in both panels. Similar pictures are obtained for the other categorical predictors. For the metric predictor floor space one obtains a decreasing function, pictured in Figure 4, which means that the net rent per square meter decreases with growing floor space.

It should be noted that the structured tree is not a tree in the sense of traditional recursive partitioning, where models are fitted recursively to sub samples defined by nodes. In structured trees one obtains for each of the categorical predictors that are used in the tree component a separate tree. The obtained trees show which categories have to be distinguished given the other predictors are included in the model. Only because of this feature they are competitors to regularization approaches that are able to handle a large number of categories and predictors. Related approaches, but not for categorical predictors, are the so-called partially linear trees, see Chen et al (2007), Dusseldorp et al (2010).

References

Aitkin, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. *Biometrics*, 55, 117 – 128.

Bondell, H. D. and Reich, B. J. (2009). Simultaneous Factor Selection and Collapsing Levels in ANOVA. *Biometrics*, 65, 169 – 177.

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

CA

Townsend, Z., Buckley, J., , Harada, M., and Scott, M. (2013). The Choice Between Fixed and Random Effects. In: M. Scott, J. Simonoff, B. Marx *The SAGE Handbook of Multilevel Modeling*, SAGE.

Chen, J., and Yu, K., and Hsing, A., and Therneau, T. (2007). A partially linear tree-based regression model for assessing complex joint gene–gene and gene–environment effects. *Genetic epidemiology*, 31, 238–251.

Dusseldorp, E., Conversano, C. and Van Os, B. (2010). Additive logistic regression: A statistical view of boosting. *Journal of Computational and Graphical Statistics*, 19, 514 – 530.

Friedman, J. H. and Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28, 337 – 407.

Friedman, J. H. and Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1), 1 – 22.

Gertheiss, J. and Tutz, G. (2010). Sparse Modeling of Categorial Explanatory Variables. *Annals of Applied Statistics*, 4, 2150 – 2180.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Bias Estimation for Nonorthogonal Problems. *Technometrics*, 12, 55 – 67.

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models (Second Edition). *Chapman & Hall*, second edition

Meier, L. and van de Geer, S. and Bühlmann, P. (2008). The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society, Series B*, 70, 53 – 71.

Quinlan, J. R. (1993). Programs for Machine Learning. *Computational Statistics, Data Analysis*, Morgan Kaufmann PublisherInc.

Ripley, B. D. (1996). Pattern Recognition and Neural Networks. *Cambridge University Press*, Cambridge

Simon, N. and Friedman, J. and Hastie, T. and Tibshirani, R. (2013). A Sparse-group lasso. *Journal of Computational and Graphical Statistics*, 17(3)

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

Tutz, G. and Gertheiss, J. (2014). Rating Scales as Predictors – the Old Question of Scale Level and some Answers. *Psychometrika, to appear*

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

Tutz, G. and Oelker, M. (2014). Modeling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures. *Technical Report*, 156, Department of Statistics LMU Munich.

Tutz, G., Pössnecker, W., Uhlmann, L. (2012). Variable Selection in General Multinomial Logit Models. *Technical Report*, 126, Department of Statistics LMU Munich.

Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society*, B 68, 49 – 67.

General smooth additive modelling

Simon N. Wood[1]

[1] University of Bath, UK.

E-mail for correspondence: simon.wood@r-project.org

Abstract: Regression models built using random effects and penalized reduced rank spline smoothers are popular, with the link between smoothers and random effects providing a reliable computational and inferential framework for their practical use. Indeed for exponential family responses it is possible to produce computational methods approaching the routine reliability of methods for generalized linear models. This talk will discuss the extent to which a similar framework can be produced for more general models built in terms of smooth functions, when these result in non-exponential family likelihoods.

The generic framework explored is that of quadratically penalized likelihood maximization, with smoothing parameters and other variance components estimated by Laplace Approximate Marginal Likelihood (LA-ML) optimization. The idea is that models can be built in terms of mixtures of any reduced rank quadratically penalized basis expansions (including simple Gaussian random effects, Gaussian Markov random fields, one dimensional splines, thin plate splines, tensor product splines, etc.). Reliable and efficient computational methods are developed for this setting, based on recognition of the need to deal stably with some difficult log determinant calculations and the need for stable computation as smoothing parameters tend to infinity. The basic computational approach is to optimize the LA-ML w.r.t. smoothing parameters by Newton's method, with each Newton step requiring a penalized likelihood maximization and implicit differentiation step to find the model coefficients and their derivatives w.r.t. the smoothing parameters. Interval estimates can be based on a standard Bayesian view of the smoothing process, while AIC based model selection is also possible based on suitably corrected versions of AIC.

The framework is implemented in R packages mgcv 1.8-0, and includes Tweedie, negative binomial, beta, scaled t and ordered categorical models as special cases, as well as additive Cox proportional hazard models, GAMLSS models such as zero inflated Poisson and Gaussian location scale models, and multivariate Gaussian Additive models. Examples such as adaptive signal regression for ordered categorical responses will be presented.

Keywords: Additive models; penalized smoothing; smoothness estimation

---

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

1 Backgound

Generalized additive models (GAMs) based on reduced rank spline smooths are popular in part because of the rich variety of model terms that can be represented this way (e.g. figure 1) and in part because of the reliable statistical and computational framework available for their use.

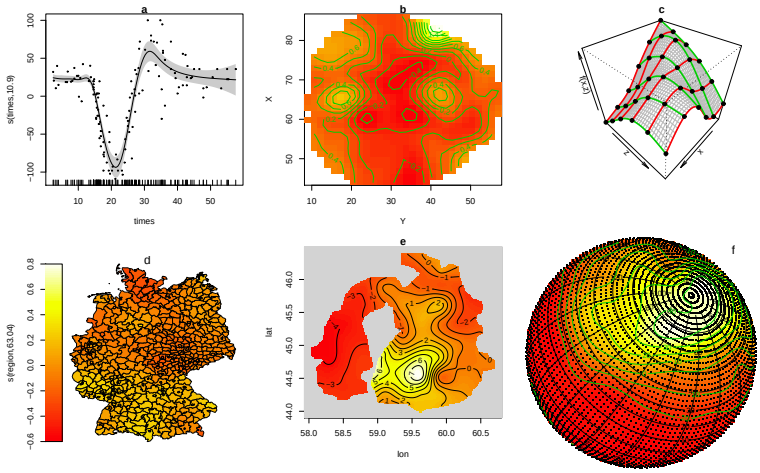


FIGURE 1. Some of the rich variety of smooth model components that can be represented using relatively low rank basis expansions with quadratic penalties. From top left: Adaptive P-spline, thin plate regression spline, tensor product interaction smooth, Gaussian Markov random field, soap film smoother and spline on the sphere.

GAMs were originally defined for exponential family responses, and this remains the setting for which the most reliable and general smoothing parameter estimation methods are available. However many further generalizations have been proposed beyond exponential family, and it would be convenient to have generally applicable and efficient smoothing parameter selection methods which would cover these too.
Generically we would like methods for dealing with the situation in which a likelihood depends on covariates via smooth functions of the covariates, represented using quadratically penalized basis expansions, so that estimation is via the penalized likelihood maximization problem of maximising

$$\mathcal{L}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\sum_j \lambda_j \boldsymbol{\beta}^T \mathbf{S}^j \boldsymbol{\beta}.$$

w.r.t. $\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the vector of basis coefficients and other model parameters, while the λ_j are smoothing parameters, and the $\mathbf{S}^j$ are (fixed known)

penalty coefficient matrices.

2 Model estimation

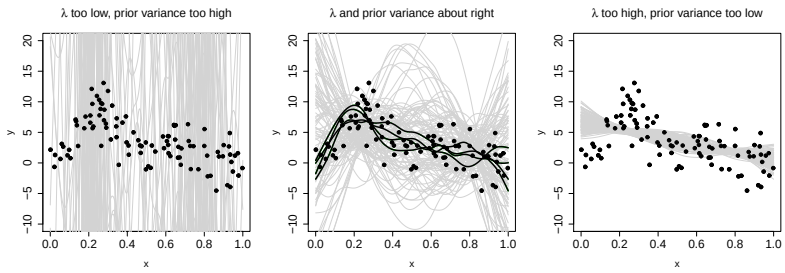


FIGURE 2. Illustration of marginal likelihood based smoothing parameter estimation. Marginal likelihood selects smoothing parameters to maximize the average likelihood of random draws from the prior on the function space. In all panels the grey curves are random draws from the prior, black dots are data to fit, and black curves are those with high likelihood, according to a threshold, all curves are centred on the best fit line through the data. Left: the smoothing parameter is too low, corresponding to high prior variance, the curves from the prior are so variable that all have very low likelihood. Middle the smoothing parameter is about right, with this degree of variability some draws (e.g. the black curves) are close enough to the truth to have high likelihood. Right: the smoothing parameter is too high, so that the curves are tightly bunched with none being close enough to the data to have high likelihood.

One way to estimate the smoothing parameters is to view the smoothness penalties as being induced by improper Gaussian priors on the model coefficients, $\boldsymbol{\beta}$ and then integrating $\boldsymbol{\beta}$ out of the joint density of the data and $\boldsymbol{\beta}$ to obtain a marginal likelihood for the smoothing parameters, which can be maximized to find them. Figure 2 illustrates that this is not as artificial as it might at first appear.

In practice the required integral is intractable, but a Laplace approximation is possible. If $\hat{\boldsymbol{\beta}}$ is the maximizer of $\mathcal{L}$ and $\boldsymbol{\mathcal{H}}$ is the negative Hessian of $\mathcal{L}$ w.r.t. the model coefficients then a Laplace approximate Marginal Likelihood (REML) score is

$$\mathcal{V}(\boldsymbol{\lambda}) = \mathcal{L}(\hat{\boldsymbol{\beta}}) + \frac{1}{2}\log|\mathbf{S}^{\lambda}|_{+} - \frac{1}{2}\log|\boldsymbol{\mathcal{H}}| + \frac{M_p}{2}\log(2\pi).$$

where $\mathbf{S}^{\lambda} = \sum_j \lambda_j \mathbf{S}^j$ (and $|\cdot|_+$ denotes a generalized determinant - the product of the non-zero eigenvalues of a matrix).

Model fitting is then most reliably undertaken by Newton optimization of $\mathcal{V}$, with each trial $\log\lambda$ Newton proposal requiring an inner Newton iteration to find $\hat{\boldsymbol{\beta}}$ and evaluate $\mathcal{V}$. The outer Newton iteration requires first and second derivatives of $\mathcal{V}$ w.r.t. $\log\boldsymbol{\lambda}$ to be computed, and these in turn require implicit differentiation to be applied in order to obtain

derivatives of the $\hat{\boldsymbol{\beta}}$ w.r.t. $\log \boldsymbol{\lambda}$. With careful structuring this optimization can be made quite efficient, with the whole process being accomplished in of the order of 10 outer iterations each of leading order cost $O(Mnp^2)$, where M is the number of smoothing parameters, n the number of data and p the number of model coefficients. However the process has one major potential source of instability.

Log determinant terms such as $\log |\mathbf{S}^\lambda|_+$ can not be reliably computed numerically for general $\mathbf{S}^\lambda$, in particular when λ_j become widely disparate in magnitude. In such cases naive determinant computations based on Choleski, QR or symmetric eigen decompositions can fail to bear any relation to the correct log determinant value as a result of finite precision arithmetic issues. The solution is to split $\mathbf{S}^\lambda$ by diagonal block where possible, and to employ similarity transform methods on overlapping blocks to ensure that large elements in one block do not incorrectly contaminate computations in another. In practice this means some initial reparameterization of smooths is necessary before fitting, with adaptive orthogonal re-parameterization considered at each step of the outer Newton method, to ensure stable computation of $\mathcal{V}$ and its derivatives.

3 Further Inference

Once fitting is accomplished, the Bayesian model underpinning the Marginal Likelihood can be re-used to obtain an approximate posterior for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}|\mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, \mathcal{H}^{-1}).$$

The resulting confidence intervals have good frequentist properties (Nychka, 1988), but note that the approach is really Bayesian here, rather than being a frequentist random effect analysis: the modeller almost never expects the smooth functions of a model to be re-sampled from the prior with each re-sampling of the data. Approximate p-values for testing smooths or random effects for equality to zero can also be computed using the methods of Wood (2013). AIC model comparison is also possible, fixing the problems identified in Greven and Kneib (2010) via a first order correction for smoothing parameter uncertainty in the computation of the AIC penalty.

4 Software and a simple example

The new methods have been implemented in R package `mgcv` version 1.8-0. Particular examples implemented are beta, ordered categorical, Tweedie, negative binomial simple zero inflated Poisson and scaled t regressions. The Cox Proportional Hazards model, multivariate Gaussian additive models, Gaussian location scale models and 2 stage zero inflated Poisson models have also been implemented. All can be used with function `gam` in `mgcv`, with the last 3 models mentioned requiring multiple formulae specifying the multiple linear predictors used.

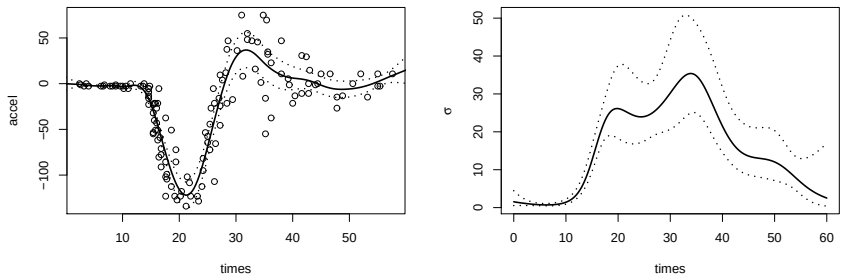


FIGURE 3. Scale location model fit to the motorcycle crash data as a simple example of the new method operation.

For example the following code fits a Gaussian scale-location model to the classic motorcycle crash test data, using an adaptive P-spline to smooth the mean and a thin plate regression spline for the standard deviation.

```
library(MASS)
library(mgcv)
m <- gam(list(accel~s(times,k=30,bs="ad"),~s(times)),
         data=mcycle,family=gaulss)
```

The resulting model fit is shown in figure 3

Acknowledgments: Part of the work reported here is collaborative with Benjamin Säfken and Natalya Pya. The work is funded by the UK EPSRC.

References

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, **83**, 1134 – 1143.

Greven, S. and Kneib, T. (2010) On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, **97**, 773 – 789

Wood, S.N. (2013). A simple test for random effects in regression models. *Biometrika*, **100**, 1005 – 1010.

Wood, S.N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, **100**, 221 – 228.

Part II - Contributed Papers

She'll be coming 'round the mountain: Simple models of complex spatial behaviour

Haakon Bakka[1]

[1] Norwegian University of Science and Technology (NTNU), Norway

E-mail for correspondence: haakon.bakka@math.ntnu.no

Abstract: Classical models in spatial statistics assume that the correlation between two points depends only on the distance between them. In practice, however, the shortest distance may not be an appropriate measure of the separation of two points. Real life is not stationary! For example, when modelling fish near the shore, correlation should not take the shortest path going across land, but should travel along the shoreline. Similar problems occur in ecology, where animal movement depends on the terrain or the existence of animal corridors.
In this talk, I will show that this type of expert knowledge can be combined with stochastic partial differential equation (SPDE) models to generate easy-to-use, computationally efficient non-stationary models. These models can be easily constructed and interpreted by non-experts. The title makes more sense after looking at figure 1.

Keywords: GMRF; Non-stationary; SPDE; Spline; Ecology.

1 Introduction

The motivation for this work came from wanting to model complex spatial domains, but the approach turned out to be more general.
Let us first think about smoothing and inference on complex domains. As a general example, we consider modelling a species that can only thrive in an aquatic environment, like fish lice. In this case, it would be natural to remove the part of our domain that was land, by cutting a hole, because we do not want our inference to be correlated across land. Or, in the case of a spline, we do not want to smooth across land. This turns out to be a hard problem to solve. For an enlightening summary of difficulties involved in smoothing problems on complex domains, we refer you to section 2 in Ramsay (2002).
Returning to the inference view, we will consider the Stochastic Partial Differential Equation (SPDE) approach introduced by Lindgren *et al.* (2011).

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

The SPDE approach considers the spatial Gaussian field (GF) as a solution to an SPDE, and approximates it by a Gaussian Markov random field (GMRF) which has good sparseness properties. One of the main benefits of this formulation is that we can leverage the knowledge and intuition in the fields of physics, numerics, and PDE analysis, which is how we came up with the ideas presented in this talk.
One feature of the SPDE approach is that you have to provide boundary conditions to the SPDE. In the case where the boundary condition is known (e.g. by some physical phenomena like a no-slip condition), the SPDE approach is already able to incorporate this easily. We will consider the harder problem, where the boundary conditions are unspecified.
No-one knows how to deal with unspecified boundary conditions in a good way, so we fudge it. Typically, by increasing the domain to contain points we don't care about, in such a way that the new boundary is far away from our original domain. Then we require the normal derivative to be zero at the new boundary. This becomes problematic when you want to cut a hole in your domain. Then it may not be possible to introduce these artificial points, and the approximation becomes dependent on the boundary assumptions. A comparison on how to handle boundary conditions can be found in section 5.2 in Wood et. al. (2008), where they comment on the work of Ramsay (2002).
The method presented here does not assign boundary conditions to the boundary of the hole that you cut out of the domain. Instead, we solve a different SPDE inside the subregion representing the hole. This approach differs substantially from the solutions presented by Ramsay and by Wood.

2 The purpose of this model

The purpose of our work on the Difficult Terrain models is to

- Model a GF with Matérn correlation with a varying correlation range, making the model non-stationary,
- Give computationally feasible algorithms for a large number of nodes / spatial points (in 2014 we estimate this to be 10^6 nodes).

We expect this to provide ways to

- Handle complex domains in both smoothing and spatial inference, without parametrizing the boundary,
- Cut holes out of the domain by letting the local range there be essentially zero,
- Generalize the current smoothing techniques by being able to infer from the data how the holes in the domain are to be handled (e.g. removed, ignored, shrunk, or a mix of these),
- Pose scientific questions to the data, on the form "is there any difference between the spatial behaviour in these two areas?"

For a proof of concept, we refer the reader to the figures. See figure 1 for a demonstration of how the correlation curves behave when you use this method to include the non-stationarity caused by an island at sea. See figure 2 for an example of how this is used to get non-stationary behaviour close to a non-trivial shoreline. See figure 3 for an example of how this can be used to model several narrow passes.

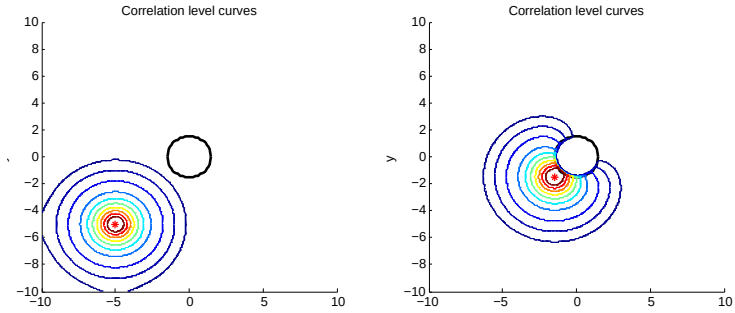


FIGURE 1. Level curves for correlation (80% to 2.5%), near an isolated island (or, if you will, a mountain that you have to "go around"). The left plot shows stationary behaviour away from the island. The right plot shows the non-stationary correction; the red dot is not correlated with the point on the other side of the island. The Difficult Terrain weights are $h = 1$ as default, and $h = 20$ inside the circle.

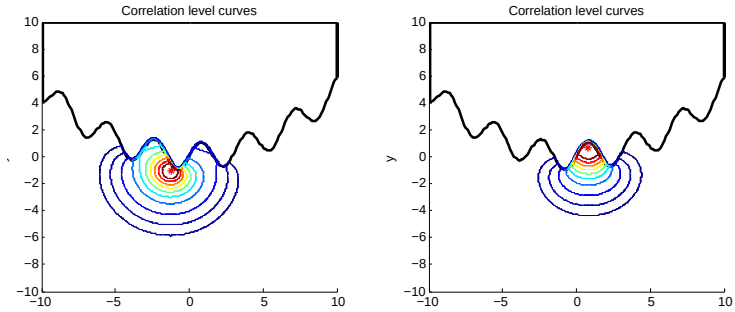


FIGURE 2. Level curves for correlation (80% to 2.5%), near a shoreline. Weights are $h = 1$ below the boundary, and $h = 20$ above the boundary.

3 The Difficult Terrain models

First, we will give some background on the SPDE approach, then we will present the Difficult Terrain approach, and lastly, we will discuss its interpretations.

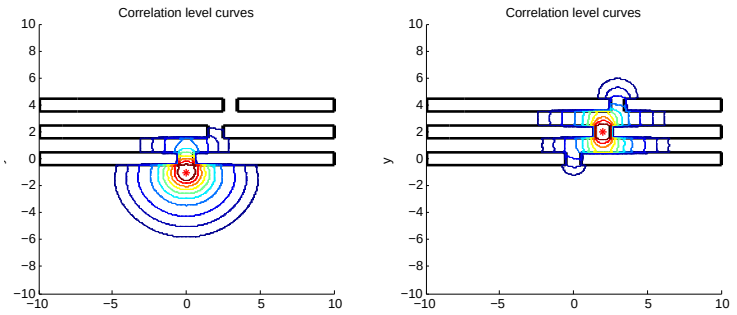


FIGURE 3. Level curves for correlation (80% to 2.5%), as you pass through three corridors. Weights are $h = 1$ as default, and $h = 20$ inside the six rectangles.

3.1 Stationary GF as a solution to an SPDE

A stationary (and isotropic) GF with Matérn correlation function can be expressed as the solution to

$$u(s) - \nabla h^{-2} \nabla u(s) = h^{-1} \tau \mathcal{W}(s), \tag{1}$$

where $u(s), s \in \Omega \subseteq \mathbb{R}^2$ is the GF, h and τ are constants. Also, $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$, and $\mathcal{W}(s)$ denotes white noise. For proofs, and to see how this can be approximated with a GMRF, see Lindgren *et al.* (2011), and note that we have taken their equation (2), but re-parametrized it and fixed $\alpha = 2$. With our parametrization, the correlation range and the variance of the field are

$$\rho \propto h^{-1},$$
$$\sigma^2 \propto \tau^2.$$

3.2 The Difficult Terrain non-stationarity

Introduce non-stationarity by letting $h = h(s)$, so that

$$u(s) - \nabla h(s)^{-2} \nabla u(s) = h(s)^{-1} \tau \mathcal{W}(s). \tag{2}$$

Let us immediately restrict ourselves to locally constant h;

$$h(s) = h_i \text{ on } \Omega_i,$$

where our domain Ω is the disjoint union of these Ω_i. Now, we note that inside the subdomain Ω_i we are solving equation (1), for a GF with range proportional to h_i^{-1}, and constant variance.

3.3 Difficult Terrain interpretation

How to interpret this locally constant correlation range? Let us consider a scaling of distances, as in figure 4. E.g. a distance in Ω_3 where $h_3(s) = 2$ is twice as demanding/difficult/resource intensive as the similar distance in the subdomain where $h(s) = 1$. And so, the correlation range will be halved in Ω_3.

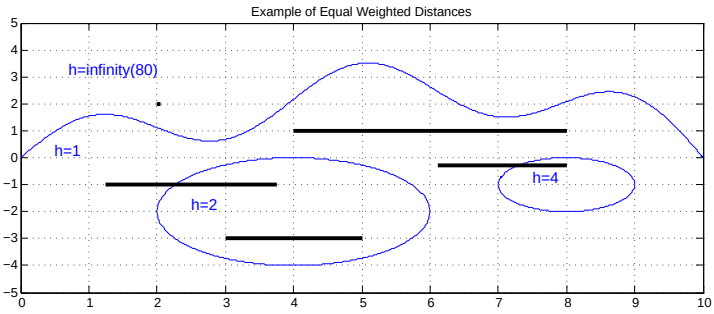


FIGURE 4. This figure shows five yardsticks of equal length, placed in a domain where distances are scaled by $1/h$. The yardstick at $(2, 2)$ is hard to spot as its weighted length is 0.05.

3.4 Generalization

There is one obvious generalization worth mentioning, namely letting $h(s)$ and $\tau = \tau(s)$ vary throughout the domain. This only requires minimal changes to the FEM algorithm, if any at all. So why not? The problem, ladies and gentlemen, is how to parametrize these functions in a simple and intuitive way. In a way that can be understood by non-experts; in a way that you feel comfortable putting priors on the hyper-parameters.

4 Computationally feasible inference

Now we will outline how to implement inference in a fast way. We want to consider problems with many nodes, so speed is an issue!
The SPDE approach, enables us to use GMRFs as approximations to GFs, by solving the SPDE with a FEM algorithm. GMRFs give huge computational benefits though sparse matrix computations, by representing the field as

$$\mathbf{u} \sim \mathcal{N}(0, Q^{-1}),$$

where Q is a sparse (symmetric positive definite) matrix.
The precision matrix Q needs to be computed for each value of the hyper-parameters that are to be explored during inference. To compute the GMRFs in the figures we have used precision matrices of size 160 000 times

160 000. Performing inference with these, even though they are very sparse, would be hard with Markov Chain Monte Carlo (MCMC) algorithms. Instead, we propose to use the Integrated Nested Laplace Approximation (INLA) algorithm.
The INLA algorithm calculates the posterior for the hyper-parameters without sampling. It is therefore significantly faster than MCMC algorithms, and has different convergence/approximation properties. For more details see Rue *et al.* (2009).

5 Future work

There are several parts that remain to be done. We plan to integrate this into the R-INLA package (see www.r-inla.org), to facilitate fast computations and widespread use. We need to find relevant data sets to illustrate how this method can be useful. We would also like to compare results with the mentioned spline smoothing techniques.

Acknowledgments: Special thanks to my supervisors Daniel Simpson and Håvard Rue.

References

Lindgren, F., Rue, H. and Lindstrom, J. (2011). *An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach.* J. R. Statist. Soc. B (2011) 73, Part 4, pp. 423-498.

Ramsay, T. (2002). *Spline smoothing over difficult regions.* J. R. Statist. Soc. B (2002) 64, Part 2, pp. 307-319.

Rue, H., Martino, S. and Chopin, N. (2009). *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.* J. R. Statist. Soc.B (2009) 71, Part 2, pp. 319-392

Wood, S.N, Bravington, M.V. and Hedley, S.L. (2008). *Soap film smoothing.* J. R. Statist. Soc. B (2008) 70, Part 5, pp. 931-955.

Modeling Exposure-lag-response associations in Survival Models with application to nutrition of critically ill patients

Andreas Bender[1], Fabian Scheipl[1], Wolfgang Hartl[1], Helmut Küchenhoff[1]

[1] Ludwig-Maximilians-Universität München, Germany

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

Abstract: In survival analysis estimation of time-varying effects has become a standard procedure, for example by extending the standard Cox PH model. Modeling so-called exposure-lag-response associations is less common and advanced methods have been introduced just recently in the context of distributed-lag models known from time-series analysis. We propose a new approach, using piece-wise exponential models to estimate a wide variety of effects, including potentially smooth, potentially time-varying effects as well as cumulative effects with leads and lags, taking advantage of the advanced inference methods known for generalized additive models. Our research has been motivated by a multi-center study that included over ten thousand patients with the goal of analyzing the effect of parenteral nutrition on short term survival of critically ill patients in intensive care units.

Keywords: exposure-lag-response associations; survival analysis; piece-wise exponential models; time-varying effects; cumulative effects; linear functionals.

1 Exposure-lag-response associations

The effect of prescribed calories and proteins on survival of critically ill patients has been a topic of discussion with controversial results (Heyland et al., 2011). We analyze data from a multi-center study of critically ill patients in intensive care units (ICU). For each patient a twelve day protocol has been conducted, recording their caloric intake alongside their goal calories. Let cal_{it_ν} denote the amount of calories received at study day $t_\nu, t_\nu = 1, \ldots 12$ and $pcal_i$ the amount of prescribed calories for patient $i, i = 1, \ldots n$. Then the caloric adequacy at day t_ν for patient i is denoted as $\nu_{it_\nu} = cal_{it_\nu}/pcal_i$. One essential question is the association between caloric adequacy and short term survival after ICU admission. The effect is

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

possibly time-varying and lagged such that nutrition on study-day t_ν is only effective in a predetermined time-window $[t_\nu + t_{lag}, t_\nu + t_{lead}]$, where t_{lag} and t_{lead} are tuning parameters (see Figure 1 for three possible definitions of leads and lags). In addition, we want the effect to be possibly cumulative, in the sense that the effect on the hazard in interval j(t) is affected by the sum of multiple past days of nutrition (depending on the definition of leads and lags). This defines the class of so called *exposure-lag-response* associations (Thomas, 1988), which have been revised and extended recently in the context of distributed-lag models (Gasparrini, 2013).

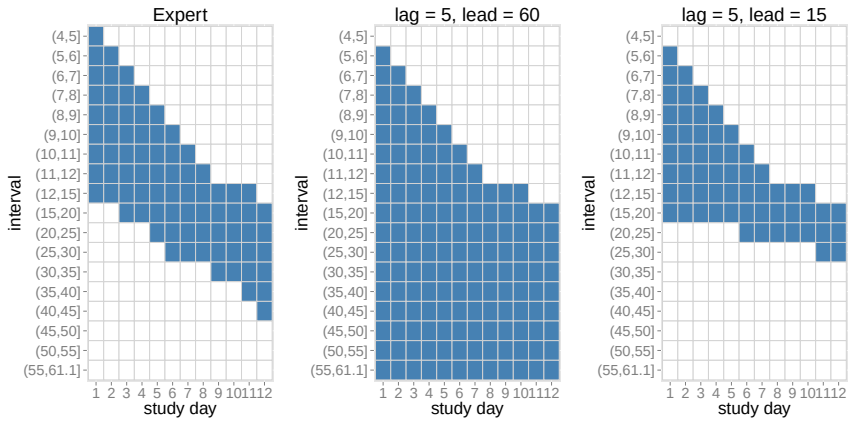


FIGURE 1. Three different lag and lead schemes for the possible effect of nutrition at time t_ν on survival in interval j. Row-wise: Which of the study days 1-12 are in effect in interval j. Column-wise: In which intervals is nutrition of study-day t_ν possibly effective.

2 Model

We use the framework of piece-wise exponential models (PEM) to model the hazard rate $\lambda(t, \mathbf{x}, \boldsymbol{\nu})$ for death at time $t > 4$ given confounder vector $\mathbf{x}$, nutrition variable vector $\boldsymbol{\nu}$ and given survival to $t > 4$. For fixed time intervals defined by cut-points $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \ldots, \kappa_I = t_{\max})$, where $t_{\max}$ is the maximal follow-up time, the hazard rate for patient i at time t, $\kappa_{j-1} < t \leq \kappa_j$ in the j-th interval is given by

$$\lambda(t, \mathbf{x}_i, \boldsymbol{\nu}_i) = \exp\left(g_0(j(t)) + \sum_{p=1}^{P} f(x_{ip}, t) + \sum_{q=1}^{Q} \sum_{t_\nu \in \mathcal{T}(j(t))} g(\nu_{iq}(t_\nu), j(t), t_\nu) \right)$$

where $g_0(j(t))$ represents the baseline hazard rate in interval $j(t)$, $f(\mathbf{x}_{.p}, t)$, $p = 1, \ldots, P$, are (potentially time-varying, potentially smooth) effects of confounders $\mathbf{x}_{.p}$ and $g(\nu_{iq}(t_\nu), j, t_\nu)$ is the partial effect of nutrition variable $\boldsymbol{\nu}_{.q}$ at time t_ν on the hazard in interval $j(t)$.

All time-dependent effects are assumed to be piece-wise constant in the intervals such that e.g. $g_0(t) = g_0(j(t))$ for all $t \in (\kappa_{j-1}, \kappa_j]$. We consider time in days. The interval borders $\boldsymbol{\kappa} = (4, 5, \ldots, 12, 15, 20, \ldots, 55, \infty)$ were chosen based on the shape of a nonparametric estimate of the marginal hazard rate. The likelihood for a PEM is proportional to that of a Poisson model with (1) one observation for each interval for each subject, yielding around 10^5 pseudo-observations in total, (2) offsets $o_{ij} = \max(0, \min(\kappa_j - \kappa_{j-1}, t_i - \kappa_{j-1}))$, where t_i is the observed time under risk for subject i and (3) responses y_{ij} equal to the event indicators δ_{ij}, with $\delta_{ij} = 0$ if subject i survived interval j and $\delta_{ij} = 1$ if not.
This allows us to use well established smoothing methods to estimate various kinds of smooth and smoothly time-varying effects by means of generalized additive models. To estimate the nutrition effect we use P-Spline tensor product smooths with B-Spline bases spanned over the dimensions of time $j(t)$ and nutrition-time t_ν. For combinations of $j(t)$ and t_ν outside the valid Lag/Lead window (see Figure 1) the respective entries in the model matrix are set to zero.

3 Effect of hypocaloric nutrition in critical care

We apply our approach to a multi-center study on the 60 day survival of patients from 352 Intensive Care Units (ICU) across 33 countries. The main interest was in the effect of nutrition as described in section 1. As confounders we used *Year* of ICU admission, *Admission Category* and initial *Diagnosis*, *Age*, *Apache II Score* and a set of variables describing the patients status during the first 4 days (since we started analysis at day 5). To account for patients who partially or fully switched to oral intake as compared to enteral or parenteral nutrition, we discretized the caloric adequacy ν_{it_ν} into intervals $0 - 30\%$, $30 - 70\%$, $> 70\%$ and assigned patients with additional oral intake to the next higher category. With the category $0 - 30\%$ as reference category, we included two linear functional terms in our model for the remaining categories, such that the estimated contribution of the nutrition variable $q \in \{30 - 70\%, > 70\%\}$ to the log-hazard of patient i in interval $j(t)$ can be rewritten as

$$\sum_{t_\nu \in \mathcal{T}(j(t))} \hat{g}(\nu_{i,q}(t_\nu), j(t), t_\nu) = \sum_{\ell=1}^{q_\nu} f_w(\mathbf{L}_i, \ell, j(t)) \cdot \hat{f}_q(\ell, j(t))$$

where $q_\nu = d_1 \cdot d_2$, d_1,d_2 number of marginal B-Spline bases on which $\hat{f}_q$ had been estimated. $\hat{f}_q$ is the estimated effect of nutrition variable $q \in \{30 - 70\%, > 70\%\}$. $\mathbf{L}_i$ are known weights for each combination of $j(t)$ and t_ν defined by the discretized version of the nutrition ν_{it_ν} and the Lag/Lead specification. Finally f_w is a known function that maps weights $\mathbf{L}_i$ of interval $j(t)$ and nutrition at time t_ν to the according effect of $\hat{f}_q$.
The tensor product smooth $\hat{f}_q$ is difficult to interpret on its own because we also need to incorporate weights $\mathbf{L}_i$ and sum up over all effects for

nutrition days effective in interval $j(t)$. Therefore we present the estimated effect of nutritional adequacy as cumulative effects at each interval $j(t)$ for three fictitious patients (see Figure 2). The results have to be interpreted as compared to a patient with hypocaloric nutrition throughout the 12 day nutrition protocol. It appears that in the first couple of days nutrition of $> 70\%$ is especially advantageous compared to hypocaloric nutrition.

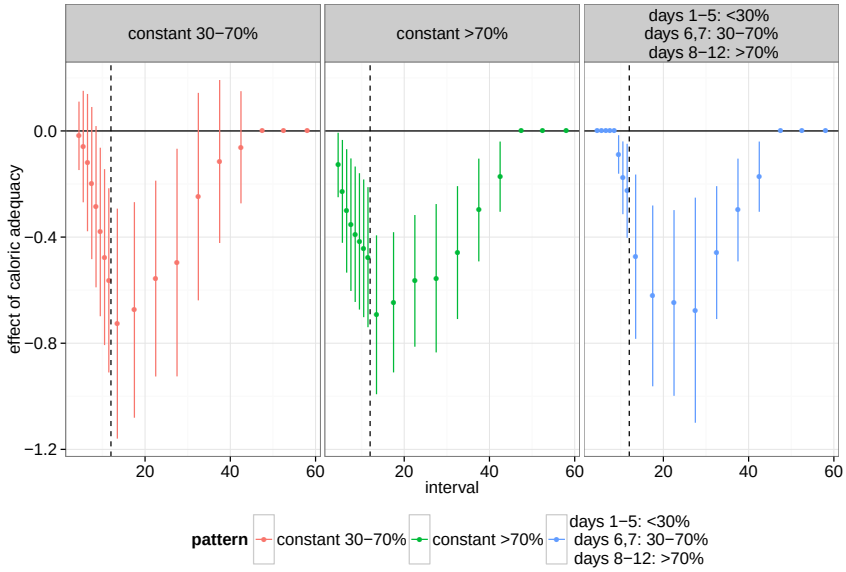


FIGURE 2. The cumulative effect of caloric adequacy on log hazard-rate for three patients with caloric adequacy of $30-70\%$ or $> 70\%$ throughout the nutrition period and a typical patient, that received $0-30\%$ in the first five days, $30-70\%$ on days 6-7 and $> 70\%$ afterwards. Effects have to be interpreted in comparison to a fictitious patient who received $< 30\%$ of prescribed calories throughout.

Summarizing, our approach makes it possible to analyze complex exposure-lag-response structures and is a valuable contribution to the discussion on the effect of nutrition on survival of critically ill patients.

Acknowledgments: We thank Prof. Dr. Daren Heyland for providing the data set and helpful cooperation.

References

Berhane, K., Hauptmann, M., Langholz, B. (2008). Using tensor product splines in modeling exposure-time-response relationships: Application to the Colorado Plateu Uranium Miners cohort. *Statistics in Medicine* **27**, 5484 — 5496.

Gasparrini, A. (2013). Modeling Exposure-lag-response Associations with Distributed Lag Non-linear Models. *Statistics in Medicine*.

Heyland, D. K., Cahill, N. , and Day, A. G. (2011). Optimal amount of calories for critically ill patients. Depends on how you slice the cake!. *Critical care medicine* **39**, 2619 — 2626.

Thomas, D. C. (1988). Models for exposure-time-response relationships with application to cancer epidemiology. *Annual Review of Public Health*, **9**, 451 — 482.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B*, **73 (1)**, 3 — 36.

The Functional Linear Array Model and an Application to Viscosity Curves

Sarah Brockhaus[1], Fabian Scheipl[1], Torsten Hothorn[2], Sonja Greven[1]

[1] Institut für Statistik, Ludwig-Maximilians-Universität München, Germany
[2] Institut für Epidemiologie, Biostatistik und Prävention, Universität Zürich, Switzerland

E-mail for correspondence: sarah.brockhaus@stat.uni-muenchen.de

Abstract: We propose the Functional Linear Array Model (FLAM) which is a model class containing scalar-on-function, function-on-scalar and function-on-function regression models. Mean, median, quantile as well as generalized additive regression models for functional and scalar responses are contained as special cases in this general framework. Estimation is conducted using a boosting algorithm, which allows for numerous covariates and automatic, data-driven model selection. Our motivating application is an experiment on viscosity of resin measured over time for different experimental settings. We fit a function-on-scalar regression model in order to determine the factors that affect the hardening process. An implementation of our methods is provided in the R add-on package FDboost.

Keywords: boosting; functional data analysis; structured additive regression.

1 Introduction

In an experiment the viscosity of resin was measured over time under different experimental settings to determine the factors affecting the curing process (Wolfgang Raffelt, Technical University of Munich, Institute for Carbon Composites). The ideal viscosity-curve should have low values in the beginning and then increase quickly. That corresponds to high fluidity during filling of the mold and a rapid hardening. As experimental factors temperature of resin, temperature of curing agent, temperature of tools, rotational speed and mass flow were investigated. For the five binary factors a fractional factorial design with 16 factor combinations was conducted with 4 replications per experimental setting, resulting in 64 observed curves. Due to technical reasons the measuring method has to be changed in a certain

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

range of viscosity. The first measuring method gives observations every two seconds, the second every ten, inducing missing values in the curves with the earlier change time. After the change of method some curves show large amounts of measurement error (Figure 1).
For the modeling, all main effects and interactions of first order are of interest, resulting in 15 effects. All in all it is necessary to estimate a robust (median) regression model for functional response, incorporating model selection and missing values.
We introduce the Functional Linear Array Model (FLAM), a very general framework for functional regression models with functional or scalar response. Among the most general competing frameworks for functional regression models are a Bayesian wavelet based approach (Meyer et al. 2013) and an approach based on additive mixed models (Scheipl et al. 2014). Neither framework discusses a unified model class for scalar and functional responses and both are limited to model the expectation of the conditional distribution. Our unified approach is the first to cover functional regression models for both scalar and functional response in one framework and to go beyond modeling the conditional mean.
In the following we will present the FLAM and its estimation by boosting, followed by the analysis of the viscosity data.

2 The functional linear array model

We consider data (Y, X), where the response Y is from the space of square integrable functions over a real interval $\mathcal{T}$, which consists of a single point for the special case of a scalar response. The covariates $X \in \mathcal{X}$ are from a product space of suitable spaces. The spaces for the single variables in X are assumed to be the real numbers or the space of square integrable functions. As generic model we set up the following additive regression model:

$$\boldsymbol{\xi}(Y|X = x) = h(x) = \sum\nolimits_j h_j(x), \tag{1}$$

where $\boldsymbol{\xi}$ is some transformation function, for instance the expectation, the median or some quantile and $h(x)$ is the linear predictor which is the sum of partial effects $h_j(x)$. Each effect $h_j(x)$ is a real valued function over $\mathcal{T}$ and is represented using a tensor product basis

$$h_j(x)(t) = (\mathbf{b}_j(x)^{\mathrm{T}} \otimes \mathbf{b}_Y(t)^{\mathrm{T}})\, \boldsymbol{\theta}_j, \tag{2}$$

where $\otimes$ is the Kronecker product, $\mathbf{b}_j : \mathcal{X} \to \mathbb{R}^{K_j}$ is a vector of basis functions depending on one or several covariates, $\mathbf{b}_Y : \mathcal{T} \to \mathbb{R}^{K_Y}$ is a vector of basis functions over the domain of the response and $\boldsymbol{\theta}_j \in \mathbb{R}^{K_j K_Y}$ is the vector of coefficients. In the case of scalar-on-function regression, $\mathbf{b}_Y(t) \equiv 1$ with $K_Y = 1$. Regularization of the effects is achieved by penalization. A suitable penalty matrix can be constructed as $\mathbf{P}_{jY} = \lambda_j(\mathbf{P}_j \otimes \mathbf{I}_{K_Y}) + \lambda_Y(\mathbf{I}_{K_j} \otimes \mathbf{P}_Y)$, where $\mathbf{P}_j$ is an appropriate penalty matrix for $\mathbf{b}_j(x)$, $\mathbf{P}_Y$ is an appropriate penalty matrix for $\mathbf{b}_Y(t)$ and λ_j,

$\lambda_Y \geq 0$ are the corresponding smoothing parameters. The penalty term has a quadratic form, resulting in a Ridge-type penalty. As we represent all effects as Kronecker product of two bases and use a Ridge-type penalty, the model is a special case of a generalized linear array model (Currie et al. 2006). Thus we denote model (1) as Functional Linear Array Model (FLAM). The array model framework allows us to estimate the model very efficiently by taking advantage of the special Kronecker structure in the design matrix. Some choices for the bases and their penalty matrices are given later in Section 5 for the case of function-on-scalar regression. The model class, however, is much more flexible allowing, e.g., for effects of functional covariates, interactions of scalar and functional covariates and group specific effects.

3 Estimation

The basic idea for the estimation of a FLAM (1) is the use of an adequate loss function that represents the estimation problem. The choice of the loss function depends on the transformation function $\boldsymbol{\xi}$ and on the conditional distribution of the response. For continuous response a typical choice is the squared error loss, yielding mean regression for a normally distributed response. The more robust absolute error loss yields median regression and more generally the check function can be used to obtain quantile regression. If the conditional distribution of the response is assumed to be of the exponential family, the loss function is the corresponding negative log-likelihood. To measure the loss of a functional response, a loss function is needed that is defined for a whole trajectory. We obtain such a loss function by integration of the loss over the domain of the response:

$$\ell\left((Y,X),h\right) = \int_{\mathcal{T}} \rho((Y,X),h)(t)dt, \tag{3}$$

where ρ is a loss function, e.g. the L_2-loss $\rho_{L_2}((Y,X),h) = \frac{1}{2}(Y - h(X))^2$.

4 Boosting

We use a component-wise boosting algorithm to fit the FLAM (1) based on the loss function (3). Boosting is an ensemble method that pursues a divide-and-conquer strategy for optimizing an expected loss criterion. The estimator is updated step-by-step to minimize the loss criterion along the steepest gradient descent. The model is represented as the sum of simple (penalized) regression models, the so called base-learners, that fit the negative gradient in each step. The base-learners specify the type of covariate effects. The loss criterion determines which characteristic of the response variable's conditional distribution is the goal of optimization (Bühlmann and Hothorn 2007). The aim of boosting is to find the solution of the optimization problem

$$h^* = \underset{h}{\operatorname{argmin}}\ \mathrm{E}\ \ell\left((Y,X),h\right). \tag{4}$$

In practical problems the integral is approximated by the weighted sum over the observed points and the expectation has to be replaced by the observed mean, yielding optimization of the empirical risk.
We adapt and extend the boosting algorithm developed by Hothorn et al. (2013) for the estimation of models with tensor product bases as base-learners.

5 Application to viscosity data

As discussed in the introduction we need a robust (median) regression model with model selection that can deal with missing values. Median regression is obtained by using the absolute loss, variable selection is inherent to boosting as it selects base-learners step-by-step and missing values are treated by setting the corresponding weights to zero.
In order to fit the model we need base-learners for an effect of the form $x\beta(t)$, where x is dummy-coded and $\beta(t)$ is the smooth effect over time. This is obtained by setting $\mathbf{b}_j(x)^{\mathrm{T}} = (1\ x)$ and $\mathbf{b}_Y(t)$ to the vector of cubic B-Splines basis functions evaluated at t. The smooth intercept is represented by setting $\mathbf{b}_1(x)^{\mathrm{T}} = (1)$. After fitting the model, the intercept-part is subtracted from each coefficient function and added to the global intercept. The penalty matrix for the linear term $\mathbf{P}_j$ is set to $\mathbf{0}$ and as the penalty matrix $\mathbf{P}_Y$ a squared second order difference matrix is used.

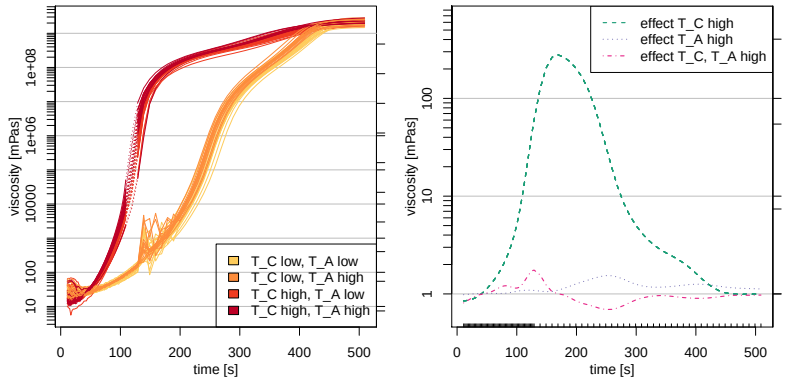


FIGURE 1. Viscosity measures over time with temperature of tools (T_C) and temperature of resin (T_A) coded in colors and missing values represented as dotted lines (left panel). The estimated coefficients for the model with T_C, T_A and their interaction (right panel).

The optimal stopping iteration is determined by 10-fold bootstrapping over curves. In the resulting model all main effects and some of the interaction effects are selected. But most base-learners contribute very small effects to the prediction of the viscosity. To obtain a parsimonious model we conduct stability selection (Shah and Samworth 2013) with a per-family-error-rate

of 2 and an expected number of terms in the model of 5. Using 100 subsamples, the effects for temperate of tools (T_C), temperature of resin (T_A) and their interaction are selected into the model, yielding

$$\text{med}\left(\log(\text{vis}_i(t))\right) = \beta_0(t) + \text{T\_A}_i\beta_A(t) + \text{T\_C}_i\beta_C(t) + \text{T\_AC}_i\beta_{AC}(t),$$

where T_A and T_C are coded as -1 for the lower and 1 for the higher level and the interaction T_AC is 1 if both temperatures are in the higher level and -1 otherwise.
The estimated coefficients can be seen in Figure 1 in the right panel. Temperature of tools has a very strong influence. From about 40 seconds on the resin in the setting with the higher tool-temperature is curing faster. For temperature of resin the effect is similar but much smaller. If temperature of tools is in its higher level the viscosity curves have the desired shape (low in the beginning, increasing rapidly); the effect is more pronounced if temperature of resin is in the higher level as well. That means for practical purposes that it is most important to control the temperature of tools. Temperature of resin has some impact as well, but the other factors do not have to be controlled precisely. All analyses are fully reproducible as data and code are part of the R add-on package FDboost (Brockhaus, 2014).

6 Conclusions

We provide a model class for both functional and scalar response, containing mean, median and quantile regression as special cases. The FLAM has a modular structure: the transformation function allows to choose which feature of the conditional distribution of the response to model and the additive predictor allows the specification of a variety of covariate effects. We take advantage of the Kronecker product structure of the design matrix to achieve computational efficiency using linear array models. The estimation is done by boosting as it is well suited to the modular structure of the model class.

Acknowledgments: Thanks to Benjamin Hofner for his advice on stability selection and to Wolfgang Raffelt for providing us with the viscosity data. The work of Sarah Brockhaus, Fabian Scheipl and Sonja Greven was funded by Emmy Noether grant GR 3793/1-1 from the German Research Foundation.

References

Brockhaus, S. (2014) (2014). FDboost: Boosting functional regression models. R package version 0.0-5, available at http://CRAN.R-project.org/package=FDboost.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477 – 505, with discussion.

Currie, I. D., Durban, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 259 – 280.

Hothorn, T., Kneib, T. and Bühlmann, P. (2013). Conditional transformation models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **76**, 3 – 27.

Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P. and Morris, J. S. (2013). Bayesian Function-on-Function Regression for Multi-Level Functional Data. Tech. rep., The Selected Works of Jeffrey S. Morris, available at http://works.bepress.com/jeffrey_s_morris/52.

Scheipl, F., Staicu, A. and Greven, S. (2014). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, in press, DOI 10.1080/10618600.2014.901914

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **75**, 55 – 80.

Reconstructing Mortality Series by Cause of Death: Two alternative approaches

Carlo G. Camarda[1]

[1] Institut National d'Études Démographiques, Paris, France

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

Abstract: Regular revisions of the classification of diseases and the consequent disruptions of mortality series are known issues in long-term cause of death analyses. Given basic assumptions and medical knowledge on eventual exchanges among causes of death, reconstruction of coherent mortality series by cause of death can be viewed as a constrained optimization problem. A combination of a penalized likelihood approach with either a quadratic programming solver or an asymmetric penalty allows to estimate exchanges among causes of death that smoothly vary over age-groups. The approach is illustrated using Russian data on digestive diseases.

Keywords: Quadratic programming; asymmetric penalty; smoothing; classification of causes of death; mortality series; inequality constraints.

1 Introduction

Mortality data are commonly collected by year, age, sex and cause of death (COD). Whereas years, ages and sex are well defined, classification of CODs changes regularly over time to reflect progress in medical knowledge. Because of these periodic revisions, mortality time series by COD shows disruptions which are not due to variability in mortality trends. We seek a method that is able to redistribute counts in an earlier period among causes from the newer classification allowing the construction of coherent long-term series.
The conventional method uses medical knowledge to identify closed groups and eventual exchanges among old and new CODs within these groups. Moreover it assumes that proportions among new COD for the whole age range are equal in most cases in the years before and after the revision. Aiming for reasonable trends for each COD, this approach redistribute death counts in an heuristic way and it sometimes requires subjective adjustments.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

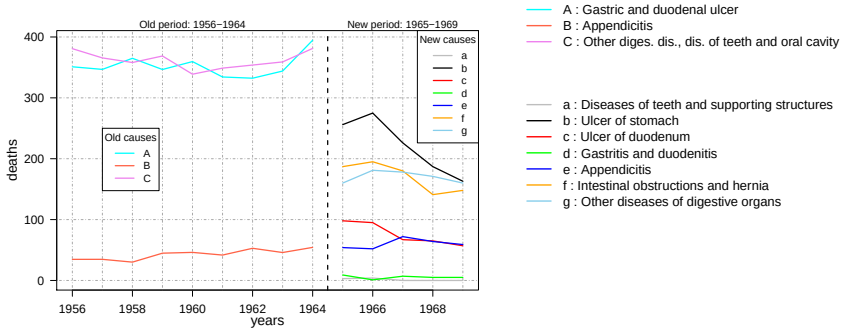


FIGURE 1. Deaths by digestive diseases. Russia, ages 50-54, 1956-1965.

In this paper we embed the whole system in a least-squares problem with asymmetric constraints, retaining the assumption on proportions and the knowledge of possible COD exchanges. Moreover we generalize the method allowing exchanges between old and new CODs which smoothly vary by age. The proposed model can be thus tackled either with a smooth quadratic programming or an asymmetric penalization approach.

2 Example

Figure 1 illustrates mortality series for digestive diseases for a specific age-group (50-54) for Russia from 1956 to 1965. Disruptions in 1965 due to a change in COD classification are evident. The complete dataset includes all age-groups from 20-24 to 85+ years old (Meslé et al. 2003).
An additional input is the Boolean matrix $\boldsymbol{T}$ which identifies possible exchanges between old and new CODs. In our example:

$$\boldsymbol{T} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccccccc} a & b & c & d & e & f & g \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}.$$

3 Model

Data are two sets of three-dimensional arrays (age, year, CODs) of death counts, one for each classification period. Estimation will be only based on data from the last (first) year of the old (new) classification: $\boldsymbol{D} = (d_{ik})$ and $\boldsymbol{Y} = (y_{il})$, where $i = 1, \ldots, m$ indexes the age-groups and CODs are labeled by k or l according to the classification period. Redistribution of counts during the whole old period will be automatically done fixing the estimated exchanges between CODs.

For ease of presentation we will first present the model for one age groups. We assume equal proportions by new CODs in the period of change (e.g. 1964 and 1965) and compute the expected deaths by new CODs:

$$\tilde{\boldsymbol{d}} = (\tilde{d}_l) = \frac{y_l}{\sum y_l} \sum d_k$$

Given the potential exchanges described in $\boldsymbol{T}$, we can express the expected deaths as

$$\tilde{\boldsymbol{d}} = \breve{\boldsymbol{X}}\, \breve{\boldsymbol{p}}\,,$$

where $\breve{\boldsymbol{p}} = (\breve{p}_{kl})$ are the proportions of deaths belonging to COD k that get redistributed into COD l and $\breve{\boldsymbol{X}}$ is the associated design matrix.
We aim to redistribute all death counts and a set of equality constraints can describe this hypothesis. Nevertheless we can incorporate into the system these constraints reducing the number of unknowns. In formula:

$$\begin{bmatrix} \tilde{d}_a \\ \tilde{d}_b \\ \tilde{d}_c \\ \tilde{d}_d \\ \tilde{d}_e - d_B \\ \tilde{d}_f \\ \tilde{d}_g - d_A - d_C \end{bmatrix} = \boldsymbol{r} = \boldsymbol{X}\,\boldsymbol{p} = \begin{bmatrix} d_C & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & d_A & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & d_A & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_A & d_C & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & d_C & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & d_C \\ -d_C & -d_A & -d_A & -d_A & -d_C & -d_C & -d_C \end{bmatrix} \begin{bmatrix} p_{Ca} \\ p_{Ab} \\ p_{Ac} \\ p_{Ad} \\ p_{Cd} \\ p_{Ce} \\ p_{Cf} \end{bmatrix}.$$

4 A quadratic programming approach

Estimating proportions the vector $\boldsymbol{p}$ is bounded between 0 and 1. We can thus view the whole system as a quadratic programming problem:

$$\underset{\boldsymbol{p}}{\text{minimize}} \quad f(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})\boldsymbol{p} + \boldsymbol{r}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{p} \qquad \text{s.t.} \quad \boldsymbol{B}\,\boldsymbol{p} \geq \boldsymbol{b} \qquad (1)$$

where

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{I}_{|\boldsymbol{p}|} \\ -\boldsymbol{I}_{|\boldsymbol{p}|} \end{bmatrix}, \quad \boldsymbol{b}^{\mathrm{T}} = \mathtt{vec}(\mathbf{1}_{|\boldsymbol{p}|}\,[0, -1])\,.$$

$\boldsymbol{I}_{|\boldsymbol{p}|}$ is the identity matrix of dimension equal the length of $\boldsymbol{p}$.
Noteworthy that both design matrix and response vector are uniquely defined by $\boldsymbol{T}$ and they are presented here only for the Russian data. More complex and larger exchanges among CODs are possible.
The dual method of Goldfarb and Idnani (1983) implemented in the `R`-routine `solve.QP` was used to solve the quadratic programming problem iterating Cholesky and QR factorizations and procedures.

4.1 Smooth proportions over age

We generalize the approach above allowing series of p_{kl} to smoothly varies over age-groups. We thus consider the complete matrix $\boldsymbol{D}$ and compute the matrix of expected deaths $\tilde{\boldsymbol{D}}$, always assuming equal proportions by new CODs in the two years of transition for each age-group.

Two-dimensional response, model matrix and vector of unknown proportions are augmented versions over i of the uni-dimensional structures presented above. In particular, each element of the previous $\boldsymbol{X}$ is replaced by a diagonal matrix over age-groups of the associated vector of deaths, e.g. $\mathtt{diag}(d_{1k}, d_{2k}, \ldots)$ instead of the simple d_k.
We enforce smoothness of the series of coefficients by penalizing the minimization problem in (1):

$$\underset{\boldsymbol{p}}{\text{minimize}} \quad f(\boldsymbol{p}) = \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{P})\boldsymbol{p} + \boldsymbol{r}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{p} \qquad \text{s.t.} \quad \boldsymbol{B}\,\boldsymbol{p} \geq \boldsymbol{b} \quad (2)$$

where $\boldsymbol{p} \in \mathbb{R}^{m\,|p|}$ and inequality matrix and vector are consequently expanded. The penalty $\boldsymbol{P}$ measures roughness of neighboring coefficients and is given by:

$$\boldsymbol{P} = \boldsymbol{I}_{|\boldsymbol{p}|} \otimes \lambda\,\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Delta}\,.$$

Symbol $\otimes$ denotes the Kronecker product of two matrices and $\boldsymbol{\Delta} \in \mathbb{R}^{(m-2)\times m}$ computes the second order of the coefficients $\boldsymbol{p}_{kl}$ (Eilers and Marx, 1996). We assume isotropic penalization of the series of proportions, i.e. a single λ for all $\boldsymbol{p}_{kl}$ over i. This smoothing parameter was chosen based on visual inspection.

5 An alternative: asymmetric penalty

A way for bypassing the implementation of quadratic programming systems is to introduce an asymmetric penalty that shrinks the coefficients within certain bounds.
Following Eilers (2005), we iteratively solve the following system:

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{P} + \boldsymbol{P}_b)\tilde{\boldsymbol{p}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}\,.$$

The additional penalty term is given by

$$\boldsymbol{P}_b = \kappa\,\mathtt{diag}(\boldsymbol{v}^0) \;+\; \kappa\,\mathtt{diag}(\boldsymbol{v}^1)$$

where

$$v_j^0 = \begin{cases} 1 & \text{if} \quad \tilde{p}_j < 0 \\ 0 & \text{if} \quad \tilde{p}_j \geqslant 0 \end{cases} \qquad \text{and} \qquad v_j^1 = \begin{cases} 1 & \text{if} \quad \tilde{p}_j > 1 \\ 0 & \text{if} \quad \tilde{p}_j \leqslant 1 \end{cases}.$$

This means that the second penalty only has the effect where proportions lie outsize the $[0, 1]$ interval. The parameter κ is rather large (here 10^8) meaning that where the penalty is in effect, it really has a very strong influence.
Differences between smooth quadratic programming approach and asymmetric penalization are indistinguishable, therefore we will only present the outcomes from the former approach.

6 Application

Figure 2 presents the estimated smooth proportions of deaths redistributed from old to new CODs in the years of classification revision. We also plot the estimations when equal proportions are assumed by age as well as an independent fit for each age.

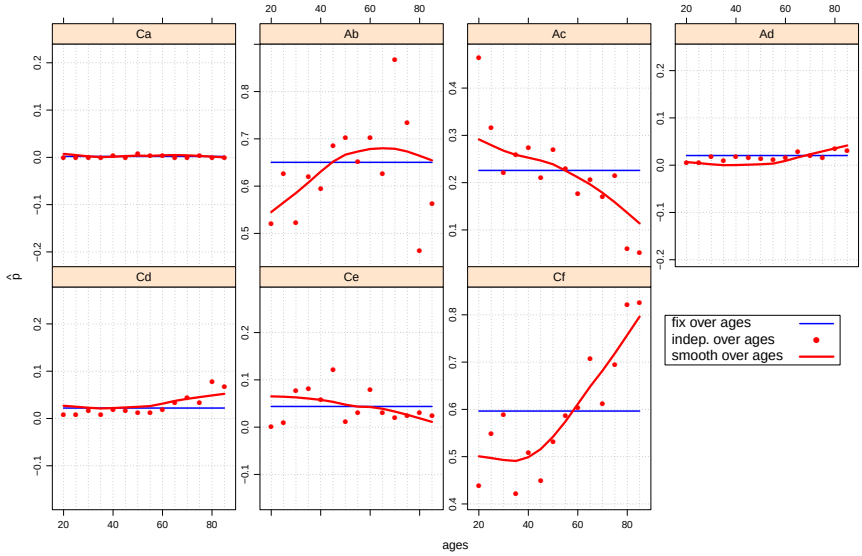


FIGURE 2. Estimated $\boldsymbol{p} = (p_{ikl})$ over ages: proportions of deaths belonging to COD k that get redistributed into COD l. Digestive diseases in Russia 1956-1965.

The estimated smooth proportions are then used to redistribute deaths in the old period by new COD and reconstruct mortality series over both periods. Figure 3 (left panel) shows that these series no longer present disruptions and typical age-patterns are well described, too (see a specific COD on the right panel of Figure 3).

7 Concluding remarks

In this paper we present two interchangeable approaches for reconstructing coherent mortality series by cause of death. Starting from basic assumptions on eventual exchanges among causes of death between two revisions, we describe the problem as a least-squares system with inequality constrains. Either smooth quadratic programming or asymmetric penalization allows us to estimate the proportions assuming smoothness over age-groups.
Alternative approaches are also available for coping with the presented issue. We plan to model the logit of $\boldsymbol{p}$ adding a small ridge penalty.
Often assuming equal proportions of counts by COD in the years of change may well be too strong. We plan to extend the model assuming a smooth

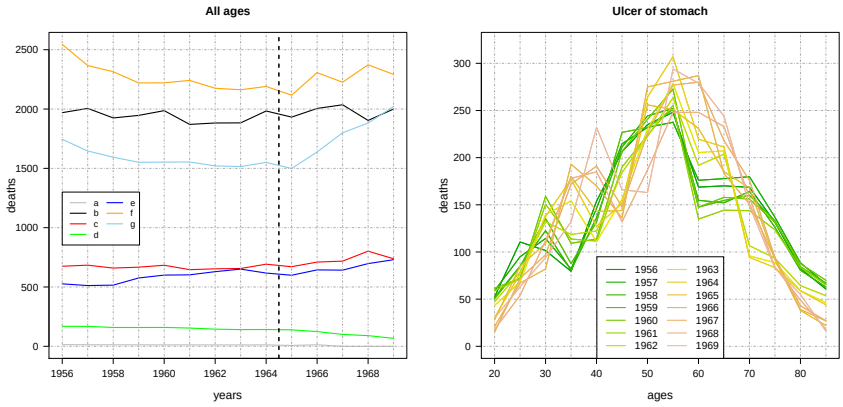


FIGURE 3. Left: Deaths counts for all ages. Digestive diseases in Russia, ages 20-85, 1956-1965. Right: age-patterns for all years of a specific COD.

change of the proportions over the new period and back-forecasting these proportions in the last year of the old classification. Furthermore, the proposed methods are expendable in other contexts in which time-series are disrupted for recognized reasons, e.g. series with historic border changes.

References

Eilers, P.H.C. (2005). Unimodal smoothing. *Journal of Chemometrics*, **19**, 317 – 328.

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

Meslé, F., Vallin, J., Hertrich, V., Andreev, E., and Shkolnikov, V. (2003). Causes of death in Russia: assessing trends since the 1950s. In: *Population of Central and Eastern Europe: Challenges and Opportunities*, Warsaw, European Population Conference, pp. 389 – 414.

Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, **27**, 1 – 33.

Functional linear mixed model for irregularly spaced phonetics data

Jona Cederbaum[1], Sonja Greven[1], Marianne Pouplier[2], Phil Hoole[2]

[1] Department of Statistics, LMU Munich, Germany
[2] Institute of Phonetics and Speech Processing, LMU Munich, Germany

E-mail for correspondence: Jona.Cederbaum@stat.uni-muenchen.de

Abstract: The investigation of sibilant assimilation is an integral part of speech production research. Apart from their functional character, data in this field often have a crossed correlation structure (speaker by target word) and measurements are commonly observed irregularly or even sparsely over time. We extend the linear mixed model to correlated functional data which are observed irregularly or even sparsely. Estimation is based on dimension reduction via functional principal component analysis, whereby the random effects are expanded in truncated bases of eigenfunctions of the estimated respective auto-covariances. We propose two ways of estimating the weights of the basis expansions. The first is straightforward and computationally efficient. The second allows for statistical inference such as the construction of point-wise confident bands for the mean function.

Keywords: Functional Linear Mixed Model; Functional Principal Component Analysis; Sparse Functional Data; Correlated Functional Data.

1 Introduction

Advancements in technology allow today's scientists to collect an increasing amount of data consisting of functional observations rather than single data points. Most methods in functional data analysis assume that observations are a) independent and/or b) observed at a large number of the same equidistant measurement points.
Speech production research is only one of numerous fields in which the data often do not meet these strong requirements. The sibilant assimilation data we present in this work have a crossed correlation structure due to repeated observations for speakers and for target words and measurement points differ between the observed curves.
We propose a model that extends conventional regression approaches by

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

accounting for both a) correlation between functional data (fd) and for b) irregular spacing of – possible few – measurement points. The model is a functional analogue to the linear mixed model (lmm), frequently used to analyze scalar correlated data. Random effects are replaced by functional random effects.
Dimension reduction becomes indispensable when dealing with fd. We use functional principal component analysis (fpca) to extract the main modes of variation in the data. Functional random intercepts are expanded in bases of eigenfunctions of the estimated respective auto-covariance functions. For the estimation of the basis weights, we propose to either compute them directly as best linear unbiased predictors (BLUPs) of the resulting lmm or to embed our model in a more general framework of Scheipl et al. (2014). The first approach is straightforward and computationally more efficient, it does not require additional estimation steps. The latter has the advantage that all model components are estimated in one framework and it allows for statistical inference.
Our approach extends two existing fpca procedures. It is a generalization of the PACE algorithm of Yao et al. (2005) who allow for irregular spacing and sparseness of measurement points but assume that the functions are uncorrelated. And it extends the longitudinal functional principal component analysis of Greven et al. (2010) to irregularly or sparsely observed measurements and to more complex correlation structures.

2 The Sibilant Assimilation Data

In order to investigate German sibilant assimilation, Pouplier and Hoole recorded acoustic data of 9 speakers. For each speaker, 5 repetitions of the same 16 target words were recorded. Each repetition was summarized in a functional index over time, varying between +1 and -1. We focus on the assimilation of the two sibilants *s* (index value 1) and *sch* (or $\int$, index value -1). Half of the target words contain the sequence /s#$\int$/, such as *Kuerbi**sSch**ale*, the other half contain the sequence /$\int$#s/, e.g. *Gula**schS**ymbol*. The resulting acoustic data are depicted in Figure 1. In the left, one can see the data for order /s#$\int$/ and in the right, data for order /$\int$#s/ are shown.
A special focus lies on the asymmetry arising from the order of the sibilants. We investigate under which conditions (order, stress of syllables, vowel context) sibilants assimilate and if assimilation is symmetric with respect to the orders /s#$\int$/ and /$\int$#s/. Moreover, we study the effect of stress, vowel context and possible interactions with sibilant order.
Due to differing reading length, the time scale is standardized to a [0,1] interval, resulting in irregular spacing of the measurements. Correlation in the data arises from repeated measurements of speakers and of target words. In a preprocessing step, we mirror the curves of order /$\int$#s/ along the time axis such that all curves have an index dynamic ranging from +1 to -1. This allows us to incorporate the index values of the two sibilant

orders in one single statistical model.
Data of this kind have frequently been analyzed by taking the index values at the 25%, 50%, and 75% points of the time interval followed by the application of multivariate methods. However, it is obvious that a lot of information is lost by this approach. Moreover, the lack of ideal reference curves makes the interpretation more difficult.

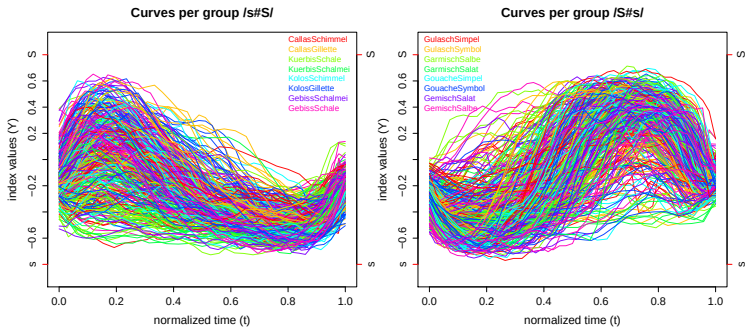


FIGURE 1. Sibilant assimilation data. Shown are the index curves over time colored by word. Curves belonging to one word are the same color. Left: Curves of order /s#∫/. Right: Curves of order /∫#s/.

3 The Functional Linear Mixed Model for our Data

We account for correlation between measurements of each speaker and of each target word, by applying a functional linear mixed model (flmm) with a crossed design of the form

$$Y_{ijh}(t) = \mu(t, x_{ij}) + B_i(t) + C_j(t) + E_{ijh}(t) + \varepsilon_{ijh}(t), \qquad (1)$$

with $Y_{ijh}(t)$ denoting the index value of the hth repetition of speaker i and target word j at time $t \in [0, 1]$. $\mu(t, x_{ij})$ is a smooth mean function depending on known covariates x_{ij} (stress and vowel context). It can include interaction terms. Correlation between measurements of the same speaker is captured by the speaker-specific functional random intercept $B_i(t)$. Analogously, the functional random intercept $C_j(t)$ accounts for target word-specific deviations of the mean. $E_{ijh}(t)$ comprises interactions between speakers and target words as well as smooth random curve-specific deviations and $\varepsilon_{ijh}(t)$ is white noise measurement error with variance $\sigma^2(t)$. We assume that $B_i(t), C_j(t), E_{ijh}(t)$ and $\varepsilon_{ijh}(t)$ are independent for all curves. Note that the irregular spacing of the data comes into play in the estimation.

4 Estimation of the FLMM

We estimate the proposed model based on dimension reduction via fpca. In a first step, the mean function $\mu(t, x_{ij})$ is estimated under the working inde-

pendence assumption, $Y_{ijh}(t) = \mu(t, x_{ij}) + \varepsilon_{ijh}(t)$, using penalized splines. Subsequently, index values are centered, $\tilde{Y}_{ijh}(t) = Y_{ijh}(t) - \hat{\mu}(t, x_{ij})$, and their cross products are used for the estimation of the auto-covariances of the functional random intercepts based on the variance decomposition

$$Cov\{\tilde{Y}_{ijh}(t), \tilde{Y}_{i'j'h'}(t')\} = Cov\{B_i(t), B_{i'}(t')\}\delta_{ii'} + Cov\{C_j(t), C_{j'}(t')\}\delta_{jj'} + \left[Cov\{E_{ijh}(t), E_{i'j'h'}(t')\} + \sigma^2(t)\delta_{tt'}\right]\delta_{ii'}\delta_{jj'}\delta_{hh'},$$

with $\delta_{xx'}$ denoting the Kronecker delta. Assuming smoothness of the auto-covariances, we use bivariate penalized splines implemented in the `R-package mgcv` (Wood, 2006) to estimate the auto-covariances from the above additive model. Strength is borrowed across curves. We expand the functional random intercepts in truncated bases of eigenfunctions of the auto-covariances (evaluated on a regular grid). We propose two alternatives for the estimation of the basis weights:

a) The weights are directly obtained as BLUPs of the resulting lmm

$$Y_{ijh}(t) = \hat{\mu}_{ijh}(t, x_{ij}) + \sum_{k=1}^{N_B} \xi_{ik}^B \hat{\phi}_k^B(t) + \sum_{k=1}^{N_C} \xi_{jk}^C \hat{\phi}_k^C(t) + \sum_{k=1}^{N_E} \xi_{ijhk}^E \hat{\phi}_k^E(t),$$

where ξ_i^B, ξ_j^C, and ξ_{ijh}^E are uncorrelated random weights and $\phi_k^B(t)$, $\phi_k^C(t)$, and $\phi_k^E(t)$ denote the eigenfunctions. Truncation levels N_B, N_C, and N_E are chosen by a pre-specified proportion of explained variance.

b) We incorporate the estimated eigenfunctions and eigenvalues in a framework for additive regression models for correlated functional responses of Scheipl et al. (2014) who use tensor product representation for each model term. Estimation is conducted via mixed model representation using again the `mgcv` package. Although an additional estimation step is necessary, it has the advantage that the mean function $\mu(t, x_{ij})$ is re-computed in the same framework allowing for statistical inference and the construction of point-wise confident bands.

5 The General Functional Linear Mixed Model

For notational simplicity, we have so far focused on the model we apply to our data (1). This model can be generalized as follows

$$Y_i(t) = \mu(t, x_i) + z_i^\top U(t) + E_i(t) + \varepsilon_i(t), \tag{2}$$

were $Y_i(t)$ is the functional response observed at arguments t in some set $\mathcal{T}$, a closed interval in $\mathbb{R}$. Time intervals are a special case of $\mathcal{T}$, but the functions can also vary subject to distances or other arguments.
$\mu(t, x_i)$ is a fixed main effect surface dependent on a vector of known covariates x_i. The random intercepts in model (1) are replaced by a vector-valued

random process $U(t)$. z_i is a covariate vector. $E_i(t)$ is a curve-specific deviation in form of a vector-valued smooth residual curve, and $\varepsilon_i(t)$ is white noise measurement error with variance $\sigma^2(t)$ which captures random uncorrelated variation within each curve. Estimation can be directly generalized for model (2).

6 Results

We analyzed the sibilant assimilation data with the proposed flmm with a crossed design structure (1). Estimation of the basis weights in the framework of additive regression models yields estimates and point-wise confident bands for the global mean function and for covariate effects. Our results suggest that sibilant assimilation of s and $\int$ is not symmetric. The effect of the sibilant order (dummy coded: 0:/s#$\int$/, 1:/$\int$#s/) with point-wise confident bands is shown on the left in Figure 2.
We also find that the covariate stress interacts with order whereas the interaction between vowel context and order is not significant as assessed by the point-wise confident bands. Moreover, our model allows a variance decomposition into speaker-, target word-, and curve-specific variability and gives an estimate for the measurement error.
We find that the covariates describe the target words so well that no word-specific functional random intercept is needed. The estimated eigenfunctions give us the main directions of variability. In Figure 2 on the right, we exemplarily display the estimated mean and the effect of adding and subtracting a suitable multiple of the first estimated principal component (PC) for the speaker-specific functional random intercept. This principal component shows how well speakers differentiate between the two sibilants. Speakers with negative weights distinguish better between s and $\int$ than speakers with positive weights. The basis weights can be used for additional analyses. Finally, predictions of the functional random intercept curves show us how speaker- and curve-specific deviations look like.

7 Conclusion

We extend the linear mixed model to correlated and irregularly or sparsely sampled functional data. Functional random effects are expanded in parsimonious bases of eigenfunctions which are estimated from the data. We propose two ways of estimating the basis weights. We demonstrate the practical relevance by applying our approach to data from speech production research in order to answer questions concerning the asymmetry of sibilant assimilation.

Acknowledgments: Special thanks to Fabian Scheipl and Simon Wood. Sonja Greven and Jona Cederbaum were funded by Emmy Noether grant GR 3793/1-1 from the German Research Foundation.

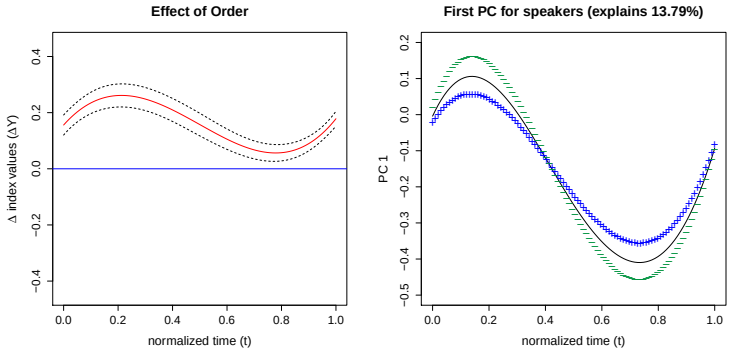


FIGURE 2. Left: Effect of sibilant order with point-wise confident bands. Right: Mean function and the effect of adding (+) and subtracting (-) a suitable multiple of the first estimated principal component for the speaker-specific random intercept.

References

Greven, S., Crainiceanu, C., Caffo, B., and Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, **4**, 1022 – 1054.

Scheipl, F., Staicu, A.-M., and Greven, S. (2014). Functional Additive Mixed Models. *Journal of Computational and Graphical Statistics*, to appear.

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

Yao, F., Müller, H., and Wang, J. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, **100(470)**, 577 – 590.

A multivariate random component model for simultaneously interval censored and right truncated data.

Sammy Chebon[1], Helena Geys[1,2], Ann De Smedt[2], Christel Faes[1]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[2] Janssen Pharmaceutica NV., Turnhoutseweg 30, Beerse, Belgium

E-mail for correspondence: sammy.chebon@uhasselt.be

Abstract: This paper deals with the analysis of simultaneously right truncated and interval censored time-to-event data with possible overdispersion and cluster induced correlation. The analysis is extended to the multivariate case through correlated shared random components (frailties).

Keywords: Truncation; Interval censoring; Shared frailty; Combined Frailty.

1 Introduction

In evaluation of compound effects during drug development, various endpoints are often simultaneously evaluated as there is no one specific standard endpoint that genuinely points to the compound effect. Subsequently, the resulting dataset is of multivariate form and analysis of such data should account for the possible correlation between different end-points. The modified HET-CAM^{VT} experiment (Van Goethem, 2006; De Smedt, 2007) provides a prime example of such a dataset. The experiment examines the viability of using a chicken egg to evaluate the injection-site-reaction properties of a given compound. A typical feature of these data is simultaneous right truncation and interval censoring. A two part model is employed: a logistic regression model for the probability of an event being observed during the experiment and a parametric Weibull model for the exact event time conditional on event occurrence. A combined frailty model (Molenberghs, 2007) is considered for the dependence in the univariate case while a correlated shared normal frailty is used to extend the analysis to the multivariate case.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

2 Data and Methods

2.1 The Data

The original dataset consisted of 14 endpoints classified under three major outcomes: Hemorrhage (H1-H4), Lysis (L5 - L8) and Coagulation(C9-C14). Each of the 14 endpoints measures time to occurrence of an irritation indicator on a fertilized egg after treatment with a compound formulation. A total of 390 eggs in batches of 3 were considered. The set-up is such that the eggs are checked for the event of interest 5 minutes after treatment, then after 15, 30 45 and 60 minutes. After this the experiment is terminated and it is unknown whether the event will eventually occur or not. Due to very small proportions of events, endpoints L6 and L8 were not considered further in this paper. Two covariates, compound concentration (0 to 10 mg/ml) and vehicle (Captisol© - 1 or Dextrose - 0) are also recorded. The objective is to identify the compound formulation's potential to cause injection site reaction and determine whether the vehicle used in compound administration has an influence.

2.2 Methodology

Note that for each egg ID i, the event of interest is either observed within the observation interval, $[0, \tau_i]$ with probability $P(\delta_i = 1) = \pi_i$ or fails to be observed with probability $P(\delta_i = 0) = 1 - \pi_i$. Furthermore, the exact time to observed events is only known to lie in the interval (t_{i1}, t_{i2}) while the unobserved event-times are truncated at time τ_i. Therefore, the data is respectively interval censored and right-truncated. The contributions to the likelihood are therefore as follows:

1. If the event occurs, it occurs in the interval $[t_{i1}, t_{i2}]$ and the likelihood is given by

$$L_i = P(T_i \in [t_{i1}, t_{i2}] | \delta_i = 1, X_i, \theta) = \pi_i \frac{S(t_{i1}|X_i, \theta_i) - S(t_{i2}|X_i, \theta_i)}{1 - S(\tau|X_i, \theta_i)} \tag{1}$$

2. If the event does not occur before the end of the study, $t_i = \tau$, the likelihood contribution is

$$L_i = P(\delta_i = 0) = 1 - \pi_i. \tag{2}$$

The general likelihood contribution for any given egg is therefore given as

$$L_i(\theta_i, \pi | X, \delta_i, t_{i1}, t_{i2}) = (1 - \pi_i)^{1-\delta_i} (\pi_i)^{\delta_i} \left(\frac{S(t_{i1}|X_i, \theta_i) - S(t_{i2}|X_i, \theta_i)}{1 - S(\tau_i|X, \theta_i)} \right)^{\delta_i} \tag{3}$$

where θ_i is the vector of parameters in the survival function; Both π_i and θ_i can be modelled as functions of covariates X , i.e. $\pi_i = \pi_i(X)$ and

$\theta_i = \theta_i(X)$, although the two sets of covariates need not be the same. The Survival function $S(.)$, can further be modeled with univariate, shared, or combined frailty (Molenberghs, 2007) terms. That is,

$$S(t_i k|X_{ik}, \theta_i, Z_{ik}) = e^{-Z_{ik}H_0(t_{ik},\theta)e^{\beta' X_{ik}+u_k}} \tag{4}$$

where Z_{ik} and u_k represent the univariate and clustering/shared frailty terms respectively. Here, $i = 1, 2, 3$ while $k = 1, ..., 130$. Omission of Z_{ik} or u_k in eq. 4 respectively results in a shared or univariate frailty model. Furthermore, Z_{ik} is assumed to follow a non-negative distribution e.g. $gamma(\alpha, \alpha)$ or log-normal while $u_k \sim N(0, \sigma^2)$ is assumed to follow a normal distribution. The Weibull function is used to model the baseline hazard $H_0(t_{ik})$ although other parametric shapes can be considered.
For multivariate extension, we consider that each event time, $T_{ir} : r = 1, 2, ..., m$ can separately be modeled using equation 4. However, we restrict ourselves to the shared frailty models. Assuming that conditional on the random effects, the event times are independent, the joint likelihood of the m outcomes is then the product of individual likelihoods in eq (3). That is

$$\begin{aligned} &L_i(\theta_{i1}, \ldots, \theta_{im}; \pi_{i1}, \ldots, \pi_{im}|X_i; \delta_{i1}, \ldots, \delta_{im}; t_{1i1}, \ldots t_{i2}; u_1, \ldots, u_m) \\ &= (1-\pi_{i1})^{1-\delta_{i1}}(\pi_{i1})^{\delta_{i1}} \left(\frac{S(t_{1i1}|X_i,\theta_{1i},u_1)-S(t_{1i2}|X_i,\theta_{1i},u_1)}{1-S(\tau_{1i}|X_i,\theta_{1i},u_1)}\right)^{\delta_{1i}} \cdots \\ &\cdots \times (1-\pi_{mi})^{1-\delta_{mi}}(\pi_{mi})^{\delta_{mi}} \left(\frac{S(t_{mi1}|X_i,\theta_{mi})-S(t_{mi2}|X_i,\theta_{mi},u_m)}{1-S(\tau_{mi}|X,\theta_{mi},u_m)}\right)^{\delta_{mi}} \end{aligned} \tag{5}$$

where the random effects are assumed to follow a multivariate normal distribution,

$$\begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix} \sim MVN \left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \cdots & \rho_{1,m}\sigma_1\sigma_m \\ \vdots & \ddots & \vdots \\ \rho_{1,m}\sigma_m\sigma_1 & \cdots & \sigma_m^2 \end{pmatrix} \right] \tag{6}$$

Extending this approach to the combined model is straiht forward. The dependence between any two event-time components is however assumed to be only through the correlated normal random effects. Model fitting is done using the pairwise pseudo-likelihood method of Fieuws and Verbeke (2006).

3 Results

The outcomes were first analyzed univariately with the best model selected based on the AIC. The univariate (heterogeneity) gamma frailty model was found to fit best for outcome H1. For outcomes C11, C12, C13 and C14, the combined model was found to fit best while for the rest of the outcomes, the shared frailty model had the lowest AIC. Based on plots of observed (red continuous line) and predicted (blue dotted line) cumulative incidence

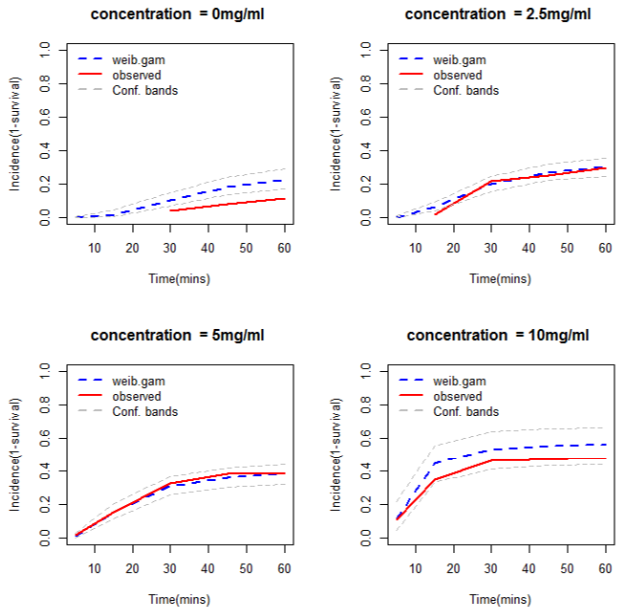


FIGURE 1. Cumulative incident rate for outcome H1

curves, the model showed a good predictive ability as illustrated in Figure 1 below for outcome H1 (first instance of diffuse bleeding).

To demonstrate our results for the multivariate extension, four randomly chosen outcomes: H2 (first punctual bleeding), L5 (first lysis effect in small blood vessels), C9 (first effect in small blood vessels-decreased blood flow) and C14 (general decrease to complete blockage in blood flow) were considered. Tables 1 and 2 below compare the results obtained on the univariate outcomes model and the multivariate respectively. The upper panel of each table represents the logistic/occurrence model part while the lower represents the event-time model. From both tables, the odds of an egg experiencing an event within the observation window increases with compound concentration. The Vehicle Captisol© tends to have low initial event rate compared to Dextrose although not significant for most outcomes.

TABLE 1. Parameter (Std. errors) for univariate shared frailty model

	H2	L5	C9	C14
Intercept	-2.320 (0.453) †	-1.433 (0.483) †	-0.774 (0.509)	-1.643 (0.336) †
Conc.	0.228 (0.037) †	1.226 (0.333) †	1.777 (0.731) †	0.732 (0.079) †
Vehicle	0.151 (0.467)	-0.608 (0.501)	0.951 (0.548)	-1.178 (0.376) †
λ	0.061 (0.078)	0.180 (0.202)	0.207 (0.141)	0.002 (0.002)
γ	2.561 (0.563) †	2.899 (0.337) †	3.284 (0.364) †	2.883 (0.283) †
Conc.	0.073 (0.211)	0.275 (0.148)	1.163 (0.341) †	0.548 (0.143) †
Vehicle	-3.134 (1.291) †	-2.205 (1.236)	-0.413 (0.889)	-1.567 (1.195)
$Vehicle \times Conc.$	0.433 (0.231)	0.031 (0.157)	-0.674 (0.357)	-0.128 (0.170)
σ^2	1.860	2.826	3.264	1.131

TABLE 2. Parameter estimates(Std. errors) for multivariate shared frailty model

	H2	L5	C9	C14
Intercept	-2.320 (0.367) †	-1.433 (0.344)†	-0.774 (0.357)	-1.643 (0.388) †
Conc.	0.228 (0.035) †	1.226 (0.129)†	1.777 (0.316) †	0.732 (0.125) †
Vehicle	0.151 (0.368) †	-0.608 (0.386)†	0.951 (0.391)†	-1.178 (0.438) †
λ	0.077 (0.094)	0.191 (0.159)	0.182 (0.160)	0.002 (0.0.001)
γ	2.561 (0.507) †	2.903 (0.260)†	2.487 (0.225)†	3.314 (0.364)†
Conc.	0.032 (0.208)	0.266 (0.152)	1.202 (0.344)†	0.598 (0.123)†
Vehicle	-3.270 (1.393) †	-2.169 (0.935)	-0.260 (0.916)	-1.213(1.001)
Vehilce × Conc.	0.458 (0.236)	0.027 (0.165)	-0.720 (0.339)	-0.176 (0.145)
σ^2	1.975	2.811	3.173	1.728

Table 3 presents the resulting correlation matrix based on the pairwise fitting approach. The outcomes exhibit moderate correlation although correlation across most coagulation outcomes tended to be higher (*results omitted*).

TABLE 3. Correlation Matrix

	H2	L5	C9	C13
H2	1			
L5	0.56	1		
C9	0.60	0.54	1	
C13	0.39	0.52	0.55	1

4 Conclusions

Although fitted under a pseudo-likelihood approach multivariate parameter estimates are very close to the univariate estimates. The model predictive performance is sound as indicated by plots. There is a significant increase in the probability of event occurrence with compound concentration based on the logistic model. The same is reflected in the event-time model though not significant in most cases. The initial low event rates in Captisol© could be attributed to a likely complex formed between the vehicle and the compound as noted by Hermans(2009).

References

De Smedt A, Goethem FV, Sysmans M, Vermeiren F, Broeckaert F , Gompel JV. (2007) *Evaluation of the concentration-and time-dependent irritation effects in a modified het-cam assay* Poster presented on World Congress of Alternatives (Tokyo, 2007)

Fieuws S., Verbeke G. (2006) *Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles* Biometrics, **62**, pp. 424 – 431.

Hermans A. (2009) *A modified Het-Cam assay: A new approach to predict in vivo vascular toxicities* Unpublished master Thesis 2009.

Molenbergs G, Verbeke G, Demetrio G. (2007) *An extended random-effects approach to modelling repeated overdispersed count data.*

Life Data Analysis, **13**, pp. 13– 513.

Van Goethem F, De Smedt A, Sysmans M, Vermeiren F, Broeckaert F, Van Gompel J, Lampo A, Vanparys P. *A modified het-cam assay: a new approach to predict vascular in vivo irritation in inflammation.*(2006) Poster presented on INVITOX congress (O0stend, 2006).

Smooth mixed models for balanced longitudinal data

Iain D. Currie[1]

[1] Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, UK

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

Abstract: Centring is a familiar device much used in the analysis of fixed effects models, but rarely seen in the context of mixed models, since the distribution of the random effects usually brings about the required centring. However there are mixed models, notably in the analysis of longitudinal data, where the distribution of the random effects does not bring about the necessary centring. In such models we use a conditional argument to centre the distribution of the random effects directly: the resulting estimates are of necessity unbiased. In general, we define a new class of mixed models, *centred mixed models*. We give some examples of models in this class and discuss efficient estimation of the fixed and random effects, and the variance parameters. We illustrate our method with the analysis of some Canadian weather data.

Keywords: Bias; Centring; Longitudinal data; Mixed models; Smoothing.

1 Introduction

We consider balanced longitudinal data $\boldsymbol{Y} = [\boldsymbol{y}_1 : \ldots : \boldsymbol{y}_n] = (y_{i,j})$, $m \times n$, with m observations on each of n subjects at times $\boldsymbol{t} = (t_1, \ldots, t_m)^{\mathrm{T}}$. One such data set is illustrated in the left panel of Figure 1, which shows the daily average temperature (taken over the period 1960-1994) in 35 Canadian cities; the mean temperature over the 35 cities is also shown. These data are available in `CanadianWeather` in the library `fda` in `R` (R Core Team, 2013). These data are balanced, and simple data analysis shows that both population mean and subject (city) effects are curved. Ruppert *et al.* (2003, ch 9) described these curves with truncated lines as follows:

$$y_{i,j} = \beta_0 + \beta_1 t_i + \sum_{k=1}^{K} u_k (t_i - \tau_k)_+ + b_{0,j} + b_{1,j} t_i + \sum_{k=1}^{\breve{K}} v_{k,j} (t_i - \breve{\tau}_k)_+ + \epsilon_{i,j}. \quad (1)$$

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

Here $x_+ = \max\{0, x\}$, and $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_K\}$ and $\breve{\boldsymbol{\tau}} = \{\breve{\tau}_1, \ldots, \breve{\tau}_{\breve{K}}\}$ are sets of K and $\breve{K}$ equally spaced internal knots in $\boldsymbol{t}$; usually, we take $\breve{K} < K$. Ruppert *et al.* (2003, p192) expressed the model as a mixed model, but did not "explore the use of standard software for fitting". Durban *et al.* (2005) showed that the model could be fitted in a straightforward fashion with the `lme` function in `R`, thus making the model widely available. However, Djeundje and Currie (2010) and Heckman *et al.* (2013) showed, by considering balanced data, that (1) could lead to seriously biased estimates of both the population and subject curves. In this paper, we propose a modification to the distribution of the random effects of the original mixed model of Ruppert *et al.* (2003) which corrects for the bias.

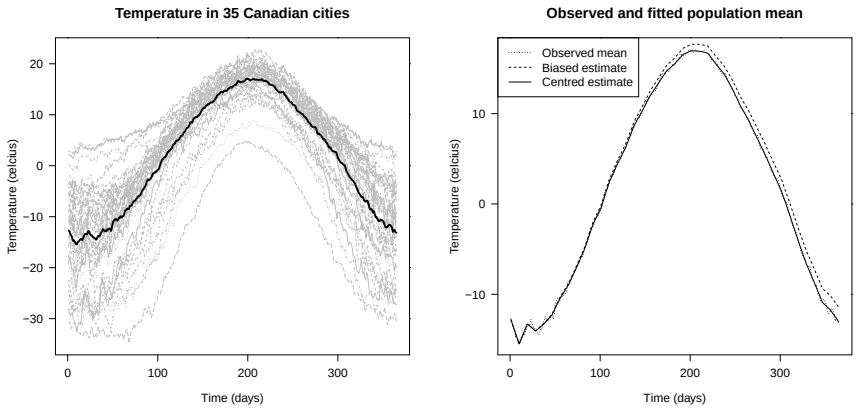


FIGURE 1. Left: temperature in 35 Canadian cities (grey), mean (black); right: observed, biased and centred estimates (forward bases) of population mean with $K = 39$ and $\tilde{K} = 19$.

The plan of the paper is as follows: in section 2 we describe our main idea, centred random effects, with reference to three examples, and present the results of the analysis of the Canadian weather data; in section 3 we define a new class of mixed models, *centred mixed models*, and describe efficient estimation in this class; in section 4 we draw some conclusions and indicate the direction of future work.

2 Examples

2.1 One-way layout

The usual model for the balanced one-way layout with m observations on each of n subjects is

$$y_{i,j} = \mu + u_j + \epsilon_{i,j},\ i = 1, \ldots, m,\ j = 1, \ldots, n,\ \epsilon_{i,j} \sim \mathcal{N}(0, \sigma_e^2). \quad (2)$$

The parameters in (2) are not identifiable so some condition is required to ensure that μ estimates the population mean. In the fixed effects model the

constraint $\sum u_j = 0$ ensures that $\hat{\mu}$ does indeed estimate the population mean. In the mixed model we suppose the $u_j \sim \mathcal{N}(0, \sigma_u^2)$ and, as is well known, this implies that the BLUP of the u_j satisfies $\sum \tilde{u}_j = 0$; in other words, correct identification of population mean and subject effects is implicit in the distribution of the random effects, and no explicit constraints are required.

We can write model (2) in matrix/vector form as

$$\boldsymbol{y} \mid \boldsymbol{u} = \mu \mathbf{1}_N + (\boldsymbol{I}_n \otimes \mathbf{1}_m)\boldsymbol{u} + \boldsymbol{\epsilon}, \;\; \boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \boldsymbol{I}_n), \;\; \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N) \quad (3)$$

where $\boldsymbol{I}_s$ is an identity matrix of size s, $\mathbf{1}_s$ is a vector of 1's of length s, $N = mn$ and $\otimes$ denotes the Kronecker product. We ask the following question: what is the appropriate way to place explicit constraints on the random effects $\boldsymbol{u}$ to ensure that $\sum \tilde{u}_j = 0$? We observe that

$$\boldsymbol{u} \mid (\textstyle\sum u_j = 0) \sim \mathcal{N}\left(\mathbf{0}, \sigma_u^2 (\boldsymbol{I}_n - \tfrac{1}{n}\mathbf{1}_n \mathbf{1}_n^{\mathrm{T}})\right) = \mathcal{N}\left(\mathbf{0}, \sigma_u^2 \boldsymbol{H}_n\right), \quad (4)$$

say, and propose that the distribution of $\boldsymbol{u}$ in (3) be replaced by the conditional distribution in (4). The matrix $\boldsymbol{H}_n$ is a familiar device and is known as the *centring matrix*, since $\boldsymbol{H}_n \boldsymbol{w}$ returns $(w_1 - \bar{w}, \ldots, w_n - \bar{w})^{\mathrm{T}}$ for any vector $\boldsymbol{w} = (w_1, \ldots, w_n)^{\mathrm{T}}$.

The two models, (3) and its centred version, give exactly the same estimates of the fixed effect μ and the random effects $\boldsymbol{u}$, and the same residual likelihood and hence the same estimates of the variance components σ_e^2 and σ_u^2; ie, in this example where the constraints on $\boldsymbol{u}$ are implicit in the original mixed model (3), explicit centring of the random effects is redundant as far as parameter estimates are concerned.

2.2 Linear population and subject effects

We consider the mixed model where population mean and subject effects are linear:

$$y_{i,j} = \beta_0 + \beta_1 t_i + u_{0,j} + u_{1,j}\, t_i + \epsilon_{i,j},\; i = 1, \ldots, m,\; j = 1, \ldots, n, \quad (5)$$

where $\boldsymbol{u}_j = (u_{0,j}, u_{1,j})^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$, and $\boldsymbol{\Sigma}_0$, 2×2, is a positive definite symmetric matrix. Let $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ be the vector of fixed effects, $\boldsymbol{u} = (\boldsymbol{u}_1^{\mathrm{T}}, \ldots, \boldsymbol{u}_n^{\mathrm{T}})^{\mathrm{T}}$ be the vector of random effects, and $\boldsymbol{T}_0 = [\mathbf{1}_m : \boldsymbol{t}]$. Corresponding to (3) we have

$$\boldsymbol{y} \mid \boldsymbol{u} = [\mathbf{1}_n \otimes \boldsymbol{T}_0]\boldsymbol{\beta} + [\boldsymbol{I}_n \otimes \boldsymbol{T}_0]\boldsymbol{u} + \boldsymbol{\epsilon}, \;\; \boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_0), \;\; \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N). \quad (6)$$

The model is readily fitted with `lme` and the BLUP of the random effects satisfies $\sum \tilde{u}_{0,j} = \sum \tilde{u}_{1,j} = 0$. As in the one-way layout, the distribution of $\boldsymbol{u}$ ensures that the estimates are correctly centred. The centred mixed model corresponding to (6) replaces $\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_0)$ with

$$\boldsymbol{u} \mid (\textstyle\sum u_{0,j} = \sum u_{1,j} = 0) \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{H}_n \otimes \boldsymbol{\Sigma}_0\right). \quad (7)$$

Again, we find that model (6) and its corresponding centred version give the same estimates of $\boldsymbol{\beta}$, $\boldsymbol{u}$, σ_e^2 and $\boldsymbol{\Sigma}_0$.

2.3 Curved population and subject effects

We can write model (1) compactly as

$$\boldsymbol{y} \mid (\boldsymbol{u}, \boldsymbol{b}, \boldsymbol{v}) = [\mathbf{1}_n \otimes \boldsymbol{T}_0]\boldsymbol{\beta} + [\mathbf{1}_n \otimes \boldsymbol{T}]\boldsymbol{u} + [\boldsymbol{I}_n \otimes \boldsymbol{T}_0]\boldsymbol{b} + [\boldsymbol{I}_n \otimes \breve{\boldsymbol{T}}]\boldsymbol{v} + \boldsymbol{\epsilon}, \quad (8)$$

where $\boldsymbol{T}$, $m \times K$, and $\breve{\boldsymbol{T}}$, $m \times \breve{K}$, are regression matrices of truncated lines. The distribution of the random effects is defined as follows:

$$\boldsymbol{u} = (u_1, \ldots, u_K)^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \boldsymbol{I}_K), \quad (9)$$

$$\boldsymbol{b} = (\boldsymbol{b}_1^{\mathrm{T}}, \ldots, \boldsymbol{b}_n^{\mathrm{T}})^{\mathrm{T}},\ \boldsymbol{b}_j = (b_{0,j}, b_{1,j})^{\mathrm{T}},\ \boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}_0), \quad (10)$$

$$\boldsymbol{v} = (\boldsymbol{v}_1^{\mathrm{T}}, \ldots, \boldsymbol{v}_n^{\mathrm{T}})^{\mathrm{T}},\ \boldsymbol{v}_j = (v_{1,j}, \ldots, v_{\breve{K},j})^{\mathrm{T}},\ \boldsymbol{v} \sim \mathcal{N}\left(\mathbf{0}, \sigma_v^2 \boldsymbol{I}_{n\breve{K}}\right). \quad (11)$$

Finally, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N)$. This is the model proposed by Ruppert *et al.* (2003) and fitted with the `lme` function in `R` by Durban *et al.* (2005). We fit the model with $K = 39$ and $\breve{K} = 19$. If we want $\boldsymbol{T}_0\hat{\boldsymbol{\beta}} + \boldsymbol{T}\tilde{\boldsymbol{u}}$ to estimate the population mean then we must check that the subject random effects are appropriately centred. We find that $\sum \tilde{b}_{0,j} = \sum \tilde{b}_{1,j} = 0$, so the linear component is correctly centred but $\sum \tilde{v}_{k,j} \neq 0$ for each $k = 1, \ldots, \breve{K}$. The right panel of Figure 1 shows that the estimate of the population mean is indeed biased. We centre the model by replacing the distribution $\boldsymbol{v} \sim \mathcal{N}\left(\mathbf{0}, \sigma_v^2 \boldsymbol{I}_{n\breve{K}}\right)$ with its centred distribution

$$\boldsymbol{v} \mid \left(\textstyle\sum_j v_{k,j} = 0,\ k = 1, \ldots, \tilde{K}\right) \sim \mathcal{N}\left(\mathbf{0}, \sigma_v^2 \boldsymbol{H}_n \otimes \boldsymbol{I}_{\breve{K}}\right). \quad (12)$$

The right panel of Figure 1 shows that the bias of the estimate of the population mean has been removed: the centred estimate now tracks the observed mean very well.

The problem with the original formulation is that the subject effects are not centred. Let $S_j(t)$ denote the j th subject effect and $\tilde{S}_j(t)$ be the BLUP of $S_j(t)$ in the centred model. Then

$$\sum_j \tilde{S}_j(t) = \sum_j \left(\tilde{b}_{0,j} + \tilde{b}_{1,j} t + \sum_k \tilde{v}_{k,j} (t - \breve{\tau}_k)_+ \right) = 0,\ \forall\, t. \quad (13)$$

In other words, centring the random effects centres the entire curve; we call this *perfect centring*. Table 1 shows values of the mean bias per city, $\sum_j \tilde{S}_j(t)/n$, for selected t. Truncated lines come in two versions: the *forward basis* with truncated lines of slope $+1$ and the *backward basis* with truncated lines of slope -1. Fixed effects models with forward and backward bases are parameterizations of each other, but in a mixed model the two models are different. The forward basis is more usual but results for both bases are shown in Table 1. The bias with the forward basis is substantial.

3 Centred mixed models

Equations (4), (7) and (12) show a common pattern and suggest a general class of mixed models. A convenient definition is:

TABLE 1. Mean bias $\sum \tilde{S}_j(t)/n$, $t = 1, 100, 200, 300, 365$

Day	1	100	200	300	365
Centred	0	0	0	0	0
Forward	0	−0.3	−0.7	−1.3	−1.6
Backward	0.1	0.3	0.2	0.1	0

Definition A *centred mixed model* is a mixed model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$ for data $\boldsymbol{Y}$, $m \times n$, $\boldsymbol{y} = \text{vec}(\boldsymbol{Y})$, such that $\boldsymbol{X}$ and $\boldsymbol{V} = \text{var}(\boldsymbol{y})$ have the form

$$\boldsymbol{X} = \mathbf{1}_n \otimes \boldsymbol{X}_0, \quad \boldsymbol{V} = \boldsymbol{I}_n \otimes \boldsymbol{V}_0 - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} \otimes \boldsymbol{\psi}_0. \tag{14}$$

All our earlier examples, including, perhaps surprisingly, the mixed model (8), are examples of centred mixed models. For model (8) we have

$$\boldsymbol{V}_0 = \sigma_e^2\boldsymbol{I}_m + \boldsymbol{T}_0\boldsymbol{\Sigma}_0\boldsymbol{T}_0^{\mathrm{T}} + \sigma_v^2\breve{\boldsymbol{T}}\breve{\boldsymbol{T}}^{\mathrm{T}}, \quad \boldsymbol{\psi}_0 = -n\sigma_u^2\boldsymbol{T}\boldsymbol{T}^{\mathrm{T}}. \tag{15}$$

Thus, even although model (8) is not centred in the usual sense of the word, the model structure is that of a centred mixed model. For the centred version of model (8) $\boldsymbol{V}_0$ remains the same and $\boldsymbol{\psi}_0 = \sigma_v^2\breve{\boldsymbol{T}}\breve{\boldsymbol{T}}^{\mathrm{T}} - n\sigma_u^2\boldsymbol{T}\boldsymbol{T}^{\mathrm{T}}$. The original mixed models for the one-way layout and the linear by linear model are trivial examples of centred mixed models with $\boldsymbol{\psi}_0 = 0$.

Computation is an issue with centred mixed models. As far as we are aware, centred mixed models cannot be fitted with `lme` since the required variance structure is not available. However, the structure (14) does admit an efficient estimation scheme. The key result is that a variance function $\boldsymbol{V}$ of the form (14) admits a closed form inverse

$$\boldsymbol{V}^{-1} = \boldsymbol{I}_n \otimes \boldsymbol{V}_0^{-1} + \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} \otimes (\boldsymbol{V}_0 - \boldsymbol{\psi}_0)^{-1}\boldsymbol{\psi}_0\boldsymbol{V}_0^{-1}. \tag{16}$$

This result enables us to obtain all the necessary quantities in the estimating equations with computations of size m instead of size $N = mn$. For example, the variance components are chosen by maximising the residual likelihood

$$-\tfrac{1}{2}\log|\boldsymbol{V}| - \tfrac{1}{2}\log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}| - \tfrac{1}{2}\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1})\boldsymbol{y} \tag{17}$$

and we can show that

$$|\boldsymbol{V}| = |\boldsymbol{V}_0|^{n-1}|\boldsymbol{V}_0 - \boldsymbol{\psi}_0| \tag{18}$$

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X} = n\boldsymbol{X}_0^{\mathrm{T}}(\boldsymbol{V}_0 - \boldsymbol{\psi}_0)^{-1}\boldsymbol{X}_0 \tag{19}$$

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y} = \boldsymbol{X}_0^{\mathrm{T}}(\boldsymbol{V}_0 - \boldsymbol{\psi}_0)^{-1}\boldsymbol{Y}\mathbf{1}_n. \tag{20}$$

The estimate of the fixed effect $\boldsymbol{\beta}$ follows from (19) and (20):

$$\hat{\boldsymbol{\beta}} = \frac{1}{n}\left(\boldsymbol{X}_0^{\mathrm{T}}(\hat{\boldsymbol{V}}_0 - \hat{\boldsymbol{\psi}}_0)_0^{-1}\boldsymbol{X}_0\right)^{-1}\boldsymbol{X}_0^{\mathrm{T}}(\hat{\boldsymbol{V}}_0 - \hat{\boldsymbol{\psi}}_0)^{-1}\boldsymbol{Y}\mathbf{1}_n. \tag{21}$$

Similar formulae exist for the remaining term in the residual likelihood and the estimates of the random effects, but are omitted in this paper. The computational formulae (18) through (21) simplify in the trivial case when $\boldsymbol{\psi}_0 = 0$ and are of interest in their own right.

4 Conclusions

We have (a) defined a new class of mixed models, centred mixed models, (b) given a number of examples, (c) indicated efficient computation and (d) used such a model to remove bias in an important example in the analysis of longitudinal data. Future work includes the calculation of confidence intervals and the application of the ideas to B-spline bases and to unbalanced data.

Acknowledgments: I am grateful to Maria Durban and the Spanish Ministry of Science and Innovation (project MTM2011-28285-C02-02) for financial support, and to Janet Heffernan, Viani Djeundje, Paul Eilers and Maria Durban for useful comments on early drafts of this paper.

References

Durban, M., Harezlak, J., Wand, M.P. and Carroll, R.J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**, 1153 – 1167.

Djeundje, V.A.B. and Currie, I.D. (2010). Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. *Electronic Journal of Statistics*, **4**, 1202 – 1224.

Heckman, N., Lockhart, R. and Nielsen, J.D (2013). Penalized regression, mixed effects models and appropriate modelling. *Electronic Journal of Statistics*, **7**, 1517 – 1552.

R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press.

Spatial Modelling using GAMLSS

Fernanda De Bastiani[1], Mikis Stasinopoulos[2], Robert Rigby[2], Audrey H.M.A. Cysneiros[1]

[1] Statistics Department, Federal University of Pernambuco, Recife/PE, Brazil
[2] STORM, London Metropolitan University, London N7 8DB, UK

Abstract: This paper explores the possibilities of fitting spatial data within the GAMLSS framework, to model any or all of the parameters of a non-exponential family distribution for the response variable. It uses kriging, tensor product and thin plate splines for spatial modelling to a small data example and shows that the use of different distributions with GAMLSS helps the modelling of the data. The potential of using spatial analysis within GAMLSS is discussed.

Keywords: GAMLSS; kriging; semi-parametric regression; tensor product splines; thin plate splines.

1 Introduction

One of the features in spatial statistics is the interdependence of data, in the sense of the first law of geography: "Everything is related to everything else, but near things are more related than distant things" Tobler (1970). Geostatistics is where a response variable (and potentially explanatory variables) are measured at points in space. Important work by Krige (1951) and Matheron (1963) laid the foundation for the field of geostatistics where some of the first methods for modelling spatial dependence were proposed, see Schabenberger and Gotway (2005) for more details. The methodology developed thereafter is referred in the literature as "kriging".
Alternatives to kriging in geostatistics, are the smoothing techniques popularised by Hastie and Tibishirani (1990) and by the P-spline approach of Eilers and Marx (1996). The P-spline models were extended to smoothing spatial data, requiring the use of tensor product and row-wise Kronecker product (Eilers and Marx, 2003; Currie *et al.,* 2006; Eilers *et al.,* 2006; Lee, 2010). Thin plate regression splines smoothers are another strong candidate since they are invariant to rotation of the covariate space, Wood (2006).
The study of discrete spatial variation, where the variables are defined on discrete domains, such as regions, regular grids or lattices, are studied

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

by the Markov Random Field theory (MRF). Extensive theoretical and practical details are provided by Rue and Held (2005).
The focus of this paper is to provide spatial modelling facilities within the GAMLSS framework, Rigby and Stasinopoulos (2005). This would allow the fitting of a variety of different distributions, rather than restricting to the exponential family distribution used within the generalised linear models (GLM) framework and allow spatial modelling of all parameters of the response variable distribution. Therefore these models are able to cope with heterogeneity in the variance and/or skewness or/and kurtosis in the data. An example of using MRF in GAMLSS is given in Rigby *et al.* (2013) while an `R package` (R core team, 2014) to incorporate MRF within `gamlss`, to model any or all of the parameters of a distribution for the response variable, is given by De Bastiani *et al.* (2014). In this paper we are concentrating on a comparison of thin plate splines, tensor product splines and kriging within GAMLSS. Section 2 describes the GAMLSS methodology, Section 3 shows the data used while Section 4 provides the analysis. Conclusions are given in section 5.

2 The GAMLSS methodology

GAMLSS provides a very general and flexible system for modelling a response variable. The distribution of the response variable is selected by the user from a very wide range of distributions available in the `gamlss` package in `R` including highly skewed and kurtotic continuous and discrete distributions. The `gamlss` package includes distributions with up to four parameters, denoted by μ, σ, ν and τ, which usually represent the location (e.g. mean), scale (e.g. standard deviation), and skewness and kurtosis shape parameters, respectively. All the parameters of the response variable distribution can be modelled using parametric and/or nonparametric smooth functions of explanatory variables, thus allowing modelling of the location, scale and shape parameters. Specifically, a GAMLSS model assumes that, for $i = 1, 2, \ldots, n$, independent observations Y_i have probability (density) function $f_Y(y_i|\theta^i)$ conditional on $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$ a vector of four distribution parameters, each of which can be a function to the explanatory variables. Rigby and Stasinopoulos (20005) define the original formulation of a GAMLSS model as follows. For $k = 1, 2, 3, 4$, let $g_k(.)$ be a known monotonic link function relating the distribution parameter θ_k to predictor η_k. Then we set

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \tag{1}$$

where h_{jk} is a smooth nonparametric function of variable X_{jk}, where, for example, $\boldsymbol{\theta} = (\theta_{k1}, \theta_{k2}, \ldots, \theta_{kn})^\top$ is a vector of length n. The advantage of modelling spatial data within GAMLSS is that different distributions beside the exponential family can be fitted and also it is possible, if needed, to model spatially any or all the parameters of the distribution.

3 Description of the chemical properties of soil data

The data were collected by the Laboratory of Spatial Statistics in western Paran (Brazil) in a commercial area of grain production in Cascavel City. The data refer to the agricultural year 2010/2011 and reference an area of $167.35ha$. The purpose of the study is to analyse the chemical contents and properties of the soil. Here the Phosphorus [P] ($mg\ dm^{-3}$) composition of the soil is modelled. The marginal distribution (ignoring the spatial coordinates) of the sample indicate some positive skewness in the distribution.

4 Comparison of different methods

The phosphorus data provide a simple example of spatial data with no covariates. The location parameter μ for each fitted distribution is modelled spatially by i) a bivariate thin plane spline ii) a bivariate tensor product spline and iii) using kriging. This was achieved by building interfaces between the `gamlss()` function and the packages `mgcv`, `SAP` and `fields`. All three packages provide 'prior weights' for their fitting procedures something that can be utilised within the iterative backfiting algorithm of `gamlss()`. The other parameters of the fitted distributions in Table 1 were fitted as constants. The smoothing parameters were chosen by 'local REML" procedure which is a Penalised Quasi Likelihood (PQL) approach in which a normal approximation of the likelihood is used at each backfitting iteration of the algorithm to estimate the random effect part of the model. Tensor products were fitted by both `mgcv` and the `SAP` packages, but here we report only the second since the results were very similar. Note that all the distributions fitted are two parameter distribution apart from the t family and BCCT which have three parameters.

It is clear from the Table 1 that the normal or t symmetric distributions are rejected in favour of skew alternatives. The reverse Gumbel, gamma or the LOGNO seem to fit best resulting to lower AIC or BIC. All three distributions provide good residuals diagnostics in terms of worm plots (not showing here).

Figure 1 shows contour and surface plots for the fitted μ assuming the LOGNO distribution which has the lowest AIC and BIC. The left column plots show the contours for μ while the right columns show the fitted surfaces for μ. The rows of Figure 1 show the mgcv-thin plate, SAP-tensor product and field-kringing respectively. The fitted contours for thin plate and tensor product are similar within the spatial range of the data, but differ considerably in the extrapolated regions. The kriging method has a more wiggly surface than the two other methods and this is more evident in the part where there is a gap in the dataset as can be seen is Figure 1.

5 Conclusions

We have shown that it is possible to model spatial data within the GAMLSS model by interfacing GAMLSS with other available `R` packages, `mgcv`, `SAP`,

	Distribution	df	GD	AIC	BIC
ga - REML thin plate	Normal	7.2373	522.669	537.143	555.154
	t-student	5.7888	525.437	537.010	551.416
	Gamma	7.3319	505.559	520.222	538.469
	Inverse Gaussian	7.6225	510.077	525.322	544.292
	Reverse Gumbel	8.5803	499.761	*516.921*	538.274
	BCCG	9.6918	499.080	518.464	542.583
	LOGNO	7.4846	503.255	518.224	*536.851*
sap - REML tensor product	Normal	7.0965	522.666	536.859	554.519
	t-student	6.1372	525.657	537.931	553.204
	Gamma	6.8079	505.514	519.130	536.072
	Inverse Gaussian	6.6682	511.212	524.549	541.144
	Reverse Gumbel	7.6323	502.372	*517.636*	536.630
	BCCG	7.5979	504.848	520.044	538.953
	LOGNO	6.8470	503.964	517.658	**534.698**
fields-REML kriging	Normal	11.5273	513.293	536.348	565.035
	t-student	7.0489	522.683	536.781	554.323
	Gamma	12.1397	494.277	518.556	*548.768*
	Inverse Gaussian	12.0504	503.245	527.346	557.335
	Reverse Gumbel	16.2939	486.066	518.653	559.203
	BCCG	15.0974	487.485	517.680	555.252
	LOGNO	13.2785	489.653	**516.210**	549.255

TABLE 1. Summary of the fitted models, showing the effective degrees of freedom (df) for the spatial part of the model, the global deviance (GD), the AIC and BIC. It highlights that the LOg-Normal has the lowest AIC and BIC.

and `field`. The resulting methodology allows fitting of non Exponential family response variable distributions and also allows spatial modelling not only of the location parameter of the response variable distribution, but also other parameters of the distribution. An advantage of this approach is more accurate fitted centiles. Further study is required to established the properties of such models.

6 Acknowledgments

Capes - Process number: 5925-13-4 - and Facepe - Process number: IBPG-0858-1.02/11, for providing financial support for De Bastiani. Laboratory of spatial statistic for the data. Paul Eilers for providing us with a version of `SAP` package.

References

Currie, I.D., Durbn, M., Eilers, P.H.C. (2006) Generalized linear array models with applications to multidimensional smoothing, *Journal of the Royal Statistical Society*, Series B **68**, 1-22.

De Bastiani, F., Stasinopoulos, D.M., Rigby, R., Voudouris, V. (2014) *gamlss.spatial: Package to fit spatial data in gamlss*, R package version 0.1.

Eilers, P.H.C. and Marx, B.H. (1996) Flexible smoothing with b-splines and penalties (with comments and rejoinder). *Statist. Sci*, **11**, 89-121.

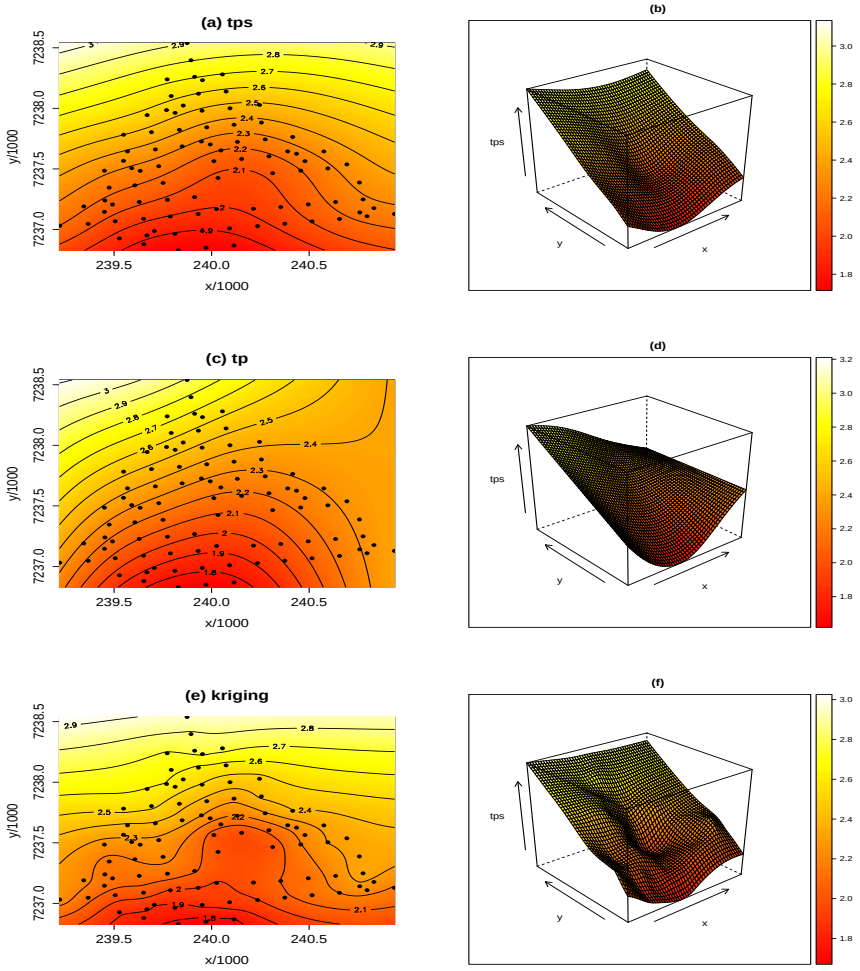


FIGURE 1. Contour and surface plots for fitted μ for the phosphorus data: a) contour and b) surface using mgcv-thin plate, c) contour and d) surface using SAP-tensor product, e) contour and f) surface using field-kringing.

Eilers, P.H.C. and Marx, B.H. (2003) Multidimensional calibration with temperature interaction using two-dimensional penalized signal regression, *Chemometrics and Intell Lab Sys*, **66**, 159-174.

Eilers, P.H.C., Currie, I.D., Durbá, M. (2006) Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, **50** (1), 61-76.

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

Krige, D.G. (1951) A statistical approach to some basic mine valuation

problems on the Witwatersrand, *J. of the Chem., Metal. and Mining Soc. of South Africa* **52** (6), 119-139.

Lee, D.J. (2010) *S*mothing mixed model for spatial and spatio-temporal data, Ph.D. Thesis, Department of Statistics, Universidad Carlos III de Madrid, Spain.

Matheron, G. (1963) Principles of Geostatistics. *Economic Geology*, **58**, 1246-1266.

R Core Team (2014) R: *A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rigby, R.A. and Stasinopoulos, D.M. (2005) Generalized additive models for location, scale and shape, (with discussion). *Appl. Statist.*, **54**, 507-554.

Rigby, R.A. and Stasinopoulos, D.M. (2013) Selecting the smoothing parameters within centile estimation modelling. *Statistical Methods in Medical Research.*

Rue and Held (2005) *G*aussian markov random fields: theory and applications, Chapman & Hall, USA.

Schabenberger, O. and Gotway, C. (2005) *Statistical Methods for Spatial Data Analysis*, Chapman & Hall, London.

Tobler, W.R. (1970) A computer movie simulating urban growth in the detroit region. *Econom. Geogr.*, **46**, 234-240.

Wood, S.N. (2006) *Generalized Additive Models-An Introduction With R*, In: Texts in Statistical Science, Chapman & Hall.

Refining the structure of a pathway with a view to prediction of gene silencing effects

Vera Djordjilović[1], Monica Chiogna[1], Maria Sofia Massa[2], Chiara Romualdi[3]

[1] Department of Statistical Sciences, University of Padua, Italy
[2] Department of Statistics, University of Oxford, UK
[3] Department of Biology, University of Padua, Italy

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

Abstract: We propose a new approach for combining biological knowledge with gene expression data to infer a network of genes. The goal is that of making accurate predictions of the effects of gene silencing. The main advantage of this approach relies on its simplicity: although prior information is included, there is no need to specify a prior distribution on the space of all network structures. The approach is illustrated by predicting the effect of the knockdown of the nkd gene in a fruit fly.

Keywords: gene silencing; causal effects; biological pathways, K2 algorithm.

1 Introduction

Molecular pathways underlie the basic functions of a living cell. They feature genes, gene products and other small molecules working together to achieve a particular biological effect. One example is shown in Figure 1. Although possibly imprecise, they represent our up-to-date knowledge on most cell processes. One question pertaining to pathways is the importance of individual genes that participate in it. Are they essential for the pathway activation, or the cell can work its way around them to deliver the signal? What happens with other participants of the pathway if one gene is switched off? In order to answer these questions scientists perform experiments called gene silencing. Gene silencing is a method for suppressing a particular gene to a minimal non-lethal level, helping us learn more about the function of that gene. Although advantageous, this procedure is very expensive. In the present work, we propose a method for prediction of the effects of silencing (knockdown) that could, under certain conditions, lead to more effective experimental design and thus considerable savings of time

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

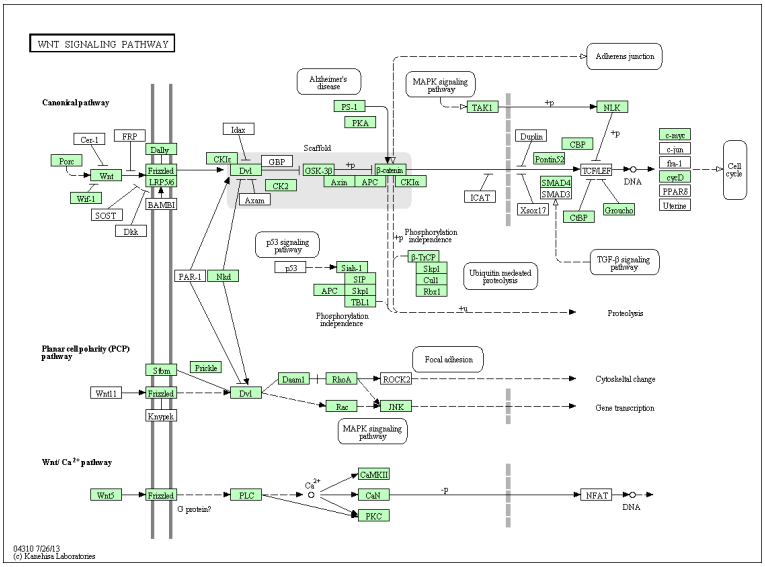


FIGURE 1. WNT signalling pathway in Drosophila, taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG)

and money. To do that, we need gene expression values (under normal conditions) for the set of genes of interest. In the first step, we combine these data with biological knowledge represented by the pathway to find the Directed Acyclic Graph (DAG) that best describes the dependencies between genes. In the second step, we use this graph to make predictions of the effects of silencing.

2 Modeling the gene knockdown

We model gene expression values (appropriately transformed and normalized) with a multivariate normal distribution

$$(X_1, \ldots, X_p)^{\mathrm{T}} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where p is the number of considered genes; $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_{p\times p} > 0$ are unknown parameters. We assume there is a known directed acyclic graph G whose nodes correspond to $X_1, \ldots, X_p$ and whose arrows depict the relationships between them in the following way. Firstly, the joint density of $\boldsymbol{X}$ factorizes with respect to G; in other words, if Pa_i denotes a set of parents of X_i in G, the joint density can be expressed as the product of p one-dimensional conditional densities:

$$f(x_1, \ldots, x_p) = \prod_{i=1}^{p} f(x_i \mid \mathsf{Pa}_i), \tag{1}$$

where f is a generic term for density function. Secondly, we assume that that arrows represent cause and effect relations between variables. In our

context, this can be interpreted as follows: if $X_i \to X_j$ then a change in expression of gene i causes a change in expression of gene j but not vice versa. This a strong assumption not verifiable mathematically and thus based on subject matter knowledge. Genomics setting is one of the (few) contexts in which such assumption is plausible.
Assume now that we would like to predict the effect of silencing gene k, where $k \in \{1, \ldots, p\}$. Directed acyclic graphs provide good framework for studying effects of interventions (see Pearl, 2000). Here, we give a brief description of this approach modified to accommodate continuous data. Let the external intervention leading to gene silencing (knockdown) be target specific, i.e., it affects directly only the targeted gene. Define he j-th causal effect of knockdown, δ_j, to be a change in the mean of X_j resulting from a unit decrease in the silenced gene X_k. Because the joint distribution is multivariate normal, all conditional distributions are also normal, and thus (1) defines a set of linear regression models. Denote $\boldsymbol{B} = \{\beta_{ij}\}_{i,j=1}^{p}$ the upper triangular matrix of regression coefficients, where β_{ij} is the coefficient of X_i in regression for X_j. We note that such ordering of nodes, so that $\boldsymbol{B}$ is upper triangular, is always possible in directed acyclic graphs. It can be easily verified that the vector of causal effects $\boldsymbol{\delta}$ can be found as k-th row of the matrix

$$\boldsymbol{D} = (\boldsymbol{I} - \boldsymbol{B})^{-1}, \tag{2}$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix.

3 Refining the graphical structure

We saw in the previous section that predicting the effects of knockdown once the underlying graph is known is straightforward. However, there is a great deal of uncertainty about the dependence structure. Although pathways represent our up-to-date knowledge of the cellular processes, they might not provide the optimal basis for the prediction of effects of gene silencing. In fact, if we test the model fit of the graph derived from the pathway using data on expression level of its genes, we are likely to find a poor fit, implying that the pathway is not well supported by experimental data. As a remedy we propose a method for combining the gene expression data with the information provided by the pathway. We build the proposal upon the idea of the learning algorithm K2 (Cooper and Herskovits, 1992). The K2 algorithm, belonging to the class of score based learning algorithms, searches for the graph that maximizes the posterior probability among all DAGs with a specified topological ordering. Topological ordering of a directed graph is an ordering of its nodes such that for every directed edge $X_i \to X_j$, X_i comes before X_j. The K2 algorithm takes as an input in addition to the data, the ordering of variables. In our application this represents an opportunity to include available prior information. To do that, we transform the pathway into a DAG and then pass its (non-unique) ordering to the algorithm. We note that K2 is designed for discrete variables, and to the best of our knowledge, no generalization to continuous variables

has been proposed. Although it is possible to discretize our observations, here we consider a different approach.
Score based learning algorithms consist of two independent components: a score function that evaluates each structure with respect to the data, and a search strategy employed to explore the space of possible structures. In the case of the K2, the score function is the K2 criterion while the search strategy is a one step greedy search. Since K2 criterion is defined only for discrete variables we replace it with the criterion applicable to continuous data, namely BIC criterion. The BIC criterion belongs to the Bayesian scoring metrics family and can be seen as an asymptotic approximation of the posterior probability of the structure. Houghton (1998) proved that it is a consistent scoring criterion. It contains an explicit penalization to guard against over-fitting. This penalization is sometimes found to be too stringent in the context of network inference (Yu et al. 2004) and in the next Section we address this issue. As for the other component of the algorithm, the search strategy, the simple greedy search can be substituted with a more elaborate heuristic search method. However, we opted to keep it since it has a straightforward implementation in the context of a search space restricted to DAGs with a specified topological ordering.

4 Drospohila melanogaster experiment

We applied this approach to data provided by the Biology Department of the University of Padua. They performed an experiment in which they silenced `nkd` gene of the WNT pathway (Figure 1) in the fruit fly. The data consists of two sets of 15 observations of 12 genes, the first set corresponding to the treatment (knockdown) group and the second set corresponding to the control group. This experiment provided an excellent opportunity to access the performance of our approach, since we were able to compare model based predictions with observed effects of gene silencing.
Using gene expression data of the control group and the topological ordering of genes according to the WNT pathway we applied the structure learning algorithm. In order to address the issue of too stringent penalization of the BIC criterion, but also the issue of the small sample size we adopted a bootstrap approach. We sampled 2000 samples with replacement from the original data and then estimated the structure for every sample using the proposed algorithm. This allowed us to assign an empirical measure of uncertainty to every plausible edge (an edge is plausible when it is in line with the topological ordering) by counting how many times out of 2000 it was discovered by the algorithm. On the basis of this result we constructed an "average" DAG, which consists of all the edges that were discovered at least $c\%$ of times, where c is an appropriately chosen cutoff level. Obviously, the cutoff level controls the number of edges in the resulting DAG. Subject matter considerations tell us that networks of genes are expected to be sparse, and in this particular case the number of edges is expected to approximately match the number of genes. The choice $c = 50\%$ leads to a structure with 11 edges, shown in Figure 2.

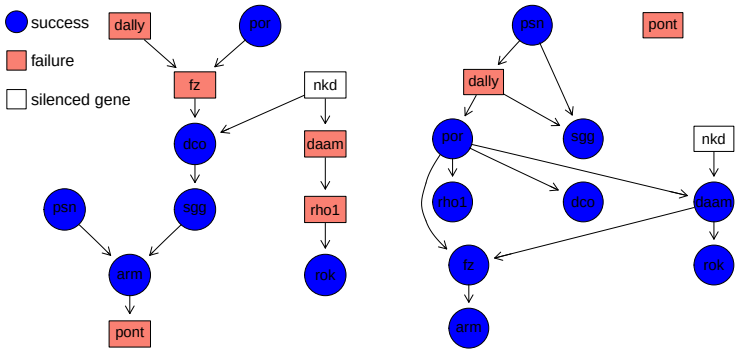


FIGURE 2. Success of predictions of mean expression values after the silencing of the `nkd` gene based on (left) DAG derived directly from the pathway and (right) DAG refined by the proposed algorithm.

We used the average DAG as a basis for making predictions of the effects of silencing of `nkd`. The causal effect on every individual gene was computed according to (2). Since we predict mean expression values of genes after the knockdown, we consider prediction a success when the predicted value lies in the 95% confidence interval for the mean computed on the basis of the knockdown group. Figure 2 shows both the graph derived directly from the pathway, and the one obtained by the proposed algorithm. By refining the structure of the graph we were able to significantly improve our predictions: the original graph yields 6 successful predictions while the refined graph led to 9 successful predictions of gene expression levels after the knockdown.

5 Discussion

In this work we propose a method to refine the structure of the signaling pathway. We consider a simple learning algorithm, inspired by the well-known K2 algorithm, that combines the prior information about the structure with observed gene expression measurements. Our approach is driven by a specific problem: we aim to find the graph that allows for the most accurate predictions of effects of gene silencing. We note that this graph does not necessarily provide a good description of an underlying biological mechanism (due to possible unobserved factors or the local nature of our approach). That said, we believe the structures found by the proposed algorithm offer a rough but very useful sketch of an underlying mechanism and can be used to find new hypothesis to be tested or as a guidance for future silencing experiments. In addition to that, this approach can signal a possible inconsistency in the representation of molecular pathways. For instance, in this study, we noted that the levels of the `dally` gene were extremely high after silencing when compared to the control group. Such behavior could not be expected neither on the basis of available prior in-

formation (Figure 1) nor on the basis of the refined DAG. This led us to look for a possible explanation in the literature, and we found that there is a feedback loop targeting this gene which was not depicted in the KEGG representation. In other words, when the WNT pathway is active it propagates a signal to initiate the transcription of the `dally` gene (among others) which leads to its high expression levels.

References

Cooper, G. F., and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, **9**, pp. 309 – 347.

Haughton, D. M. (1998). On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, **16(1)**, pp. 342 – 355.

Pearl J. (2000). *Causality: models, reasoning and inference, vol. 29*. Cambridge Univ Press.

Yu, J., Smith, V., Wang, P., Hartemink, A., and Jarvis, E. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, **20(18)**, pp. 3594 – 3603.

Smooth semi-parametric adjustment of rate differences, risk differences and relative risks

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

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

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

Abstract: New computational methods have recently been developed that allow stable fitting of constrained GLMs with bounded non-canonical link functions, such as the log link binomial model. By employing B-splines, we can extend these approaches to allow for semi-parametric adjustment of rate differences, risk differences and relative risks. These methods provide alternatives to standard fitting methods, resulting in greater stability for accommodating the required parameter bounds. They also provide a straightforward way to accommodate additional restrictions such as monotonic regression functions. We demonstrate an application to data from a clinical trial of oxygen supplementation in premature infants.

Keywords: Generalized additive model; Semi-parametric model; Rate difference; Risk difference; Relative risk.

1 Introduction

Rate differences, risk differences and relative risks are often useful effect measures in biostatistical settings, and their analogues also have broad applicability in other areas of statistics. However, in order to adjust for covariates we must use a constrained generalized linear model (GLM) with a non-canonical link where the fitted means are restricted to a bounded interval. These GLMs include the log link binomial, and identity link Poisson and binomial models. Common fitting methods based on Fisher scoring and other Newton-type algorithms can fail to converge to the maximum likelihood estimate (MLE) in this situation.
It is therefore useful to have more stable methods for fitting these models. Combinatorial EM (CEM) algorithms have recently been developed for these GLMs, allowing stable computation of the MLE. Using B-splines, we extend these methods to generalized additive models (GAMs), where

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

continuous covariates can have a semi-parametric relationship with the outcome. This approach leads to greater stability for accommodating the required parameter bounds, and allows additional model constraints such as monotonic regression functions.

2 Method

The GLM with link function g is extended to a GAM by the introduction of C continuous covariates that affect $g(\mu)$ through the unspecified functions $f_1, \ldots, f_C$. We restrict our estimate of each f_c to the space defined by a chosen set of basis functions, such that

$$\hat{f}_c(w) = \sum_{d=1}^{D_c} \gamma_{cd} B_{cd}(w).$$

The basis functions we use here are the B-splines of order 3, which are strictly non-negative. Thus if all of the coefficients are non-negative, $\hat{f}_c(w)$ will be non-negative for all w; and likewise if the coefficients are non-positive, the curve will always be non-positive. The B-splines are normalized such that $\sum_d B_{cd}(w) = 1$, which means that we must apply an identifiability constraint $\gamma_{ct_c} = 0$ for some t_c.

When $C = 0$, methods have been developed for estimating the MLE for identity link Poisson (Marschner, 2010), log link binomial (Marschner and Gillett, 2012) and identity link binomial GLMs. The methods are all CEM algorithms (Marschner, 2014), which will always converge to the MLE. With these methods we are also able to restrict certain coefficients to be non-negative or non-positive.

CEM algorithms require that the parameter space is partitioned into distinct subspaces, and use an EM algorithm to find the constrained MLE within each. One of these constrained MLEs will be the overall MLE. For these GAMs, we partition the parameter space based on the index of the smallest or largest B-spline coefficient, which can be achieved by setting a particular $\gamma_{ct_c} = 0$ and restricting the remaining coefficients to be non-negative or non-positive. We repeat this process for all possible choices of t_c and find the constrained MLE for each, one of which will coincide with the overall MLE.

A sufficient condition for $\hat{f}_c$ to be monotonically non-decreasing is that the sequence of B-spline coefficients is non-decreasing. To apply a monotonicity constraint to any of these models, we can reparameterize the smooth curve such that we are estimating the increments between successive coefficients, and can constrain these to be non-negative or non-positive, as required.

3 Application

The BOOST-NZ Study (Darlow et al., 2014) was a randomized trial in premature infants, comparing the effects of different target ranges for oxygen saturation (SpO_2).

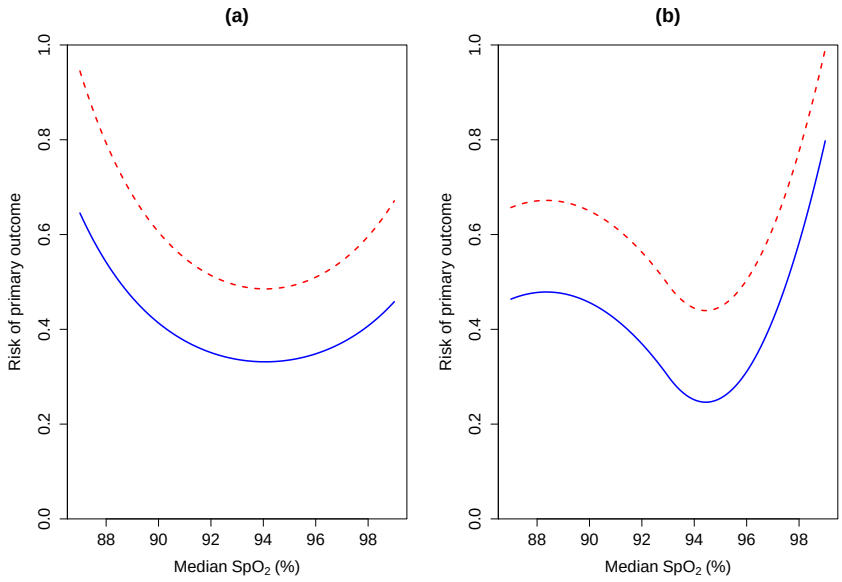


FIGURE 1. Risk of primary outcome by median SpO_2 and randomized treatment (blue solid = low target, red dashed = high target) in BOOST-NZ, under (a) log link and (b) identity link binomial models.

Both high and low levels of oxygen are associated with mortality and other complications, so the primary outcome of the study was death or major disability at two years of age. Unadjusted analysis of the primary outcome showed a relative risk of 1.16 (95% CI 0.90–1.50) and a risk difference of 0.06 (95% CI -0.04–0.17), with lower risk in the low-target group.

We use the methods outlined in Section 2 to adjust these effect measures for the actual level of oxygen that the infant received. Each infant's median SpO_2 level whilst receiving supplementary oxygen was entered as the semi-parametric covariate into each model, and the results are shown in Figure 1. The adjusted analyses show that the minimum risk is associated with an SpO_2 close to 94%. The adjusted effect of randomized treatment is a relative risk of 1.46 (95% CI 1.04–2.07) and a risk difference of 0.19 (95% CI 0.06–0.33). The confidence intervals for these parameters were estimated using a normal approximation.

For the outcome of mortality, the risk of death decreases as the SpO_2 level increases, and so we can restrict the semi-parametric curve to be monotonically non-increasing. The unadjusted effect of treatment is a relative risk of 1.08 (95% CI 0.65–1.78) or a risk difference of 0.01 (95% CI -0.06–0.09) in favour of the low-target group.

The results of the adjusted analyses are shown in Figure 2. The adjusted relative risk is estimated to be 1.89, and the adjusted risk difference is 0.04. The estimates from these models are on the boundary of their respective parameter spaces, so we must estimate confidence intervals using boot-

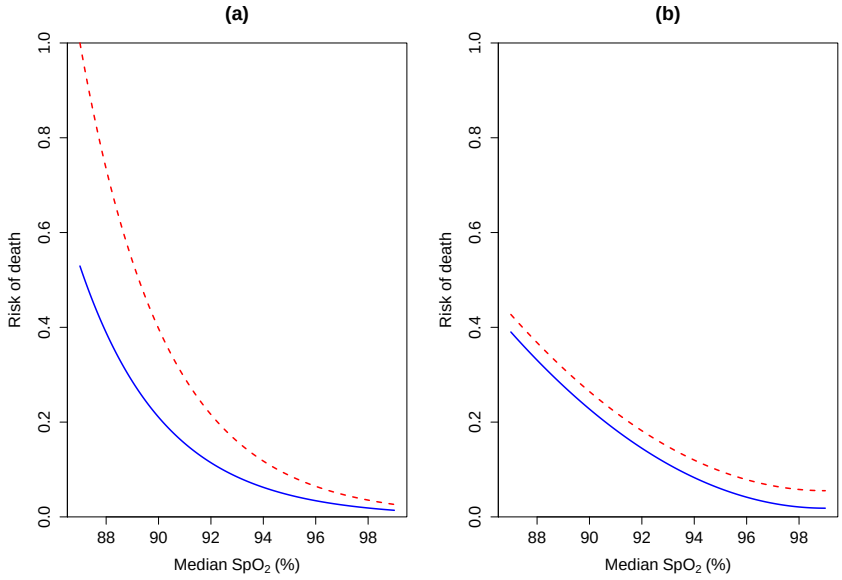


FIGURE 2. Risk of death by median SpO_2 and randomized treatment (blue solid = low target, red dashed = high target) in BOOST-NZ, under (a) log link and (b) identity link binomial models.

strap resampling. Importantly, the algorithm will converge to the MLE in every bootstrap sample, eliminating bias due to non-convergence. From 1000 bootstrap samples, we estimate the 95% confidence intervals to be 1.10–2.86 for the relative risk, and -0.03–0.10 for the risk difference.

4 Other methods

We compared our approach with other methods for fitting GAMs that have been implemented in R, and were able to show that our method has advantages over existing methods in some contexts.

The most notable existing methods are implemented in the `gam` (Hastie, 2013), `mgcv` (Wood, 2011) and `gamlss` (Rigby and Stasinopoulos, 2005) packages. The fitting procedures underlying these approaches each employ a Newton-type algorithm, which is not guaranteed to converge to the MLE unless step-size optimization is performed.

In fact, of these packages, only `mgcv` incorporates automatic step-halving if the potential update of the parameter estimates moves outside the parameter space. This method reported convergence in all 1000 bootstrap samples for the analysis in Figure 1, for both the log and identity links. However, in some cases `mgcv` converged to sub-optimal parameter estimates, particularly when the MLE was on the boundary of the parameter space. Furthermore, `mgcv` is unable to accommodate the monotonicity constraint for the analysis depicted in Figure 2.

The `gamlss` package allows the user to specify the step size for updating the parameter estimates and also offers the option to use step-halving if the deviance increases at a particular iteration. However, the method terminates with an error if the parameter estimates move outside the parameter space, making it inappropriate for automated model-fitting such as bootstrapping. It failed to converge in 52 of the 1000 bootstrap samples using the log link, and did not converge in any of the samples when we used the identity link.

The `gam` package does not include either step-halving or any check for the validity of the parameter estimates. As such, it may fail to converge or converge to a solution outside the parameter space, which occurred in 844 and 963 of the 1000 bootstrap samples, using the log and identity links respectively.

A difference with these methods is that they maximize a penalized likelihood, allowing greater flexibility in the number and positioning of the knots while discouraging large fluctuations in the resulting smooth curve. Penalized likelihood could be incorporated into our methods by a similar approach to that used by Marschner and Gillett (2012; Supplementary materials), but this would add substantially to the computational load.

Aside from its stability, another benefit of our approach is that it is straightforward to impose monotonicity constraints on selected smooth curves. If it is appropriate to assume monotonicity, this can reduce the spurious fluctuations in the estimated curve, and possibly increase the efficiency of the parameter estimates in the model.

The `GMBBoost` (Leitenstorfer and Tutz, 2007) and `GMonBoost` (Tutz and Leitenstorfer, 2007) functions employ the technique of likelihood boosting to apply a monotonicity constraint to smooth functions in maximizing a penalized log-likelihood. The current implementation of both, however, only allows canonical link functions, and therefore cannot be used to fit the models considered in this paper.

Acknowledgments: Special thanks to Prof. Brian Darlow, University of Otago, New Zealand for permission to present the BOOST-NZ data. This research was supported by the Australian Research Council (DP110101254).

References

Baker, S.G. (1994). The multinomial-Poisson transformation. *The Statistician*, **43**, 495 – 504.

Darlow, B.A. et al. (2014). Randomized controlled trial of oxygen saturation targets in very preterm infants: two year outcomes (BOOST-NZ). *Journal of Pediatrics* (in press), DOI: 10.1016/j.jpeds.2014.01.017.

Hastie, T. (2013). gam: Generalized Additive Models (*R package version 1.09*). URL `http://CRAN.R-project.org/package=gam`

Leitenstorfer, F. and Tutz, G. (2007). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics*, **8**, 654 – 673.

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

Marschner, I.C. (2014). Combinatorial EM algorithms. *Statistics and Computing* (in press), DOI: 10.1007/s11222-013-9411-7.

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

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

Tutz, G. and Leitenstorfer, F. (2007). Generalized smooth monotonic regression in additive modelling. *Journal of Computational and Graphical Statistics*, **16**, 165 – 188.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, **73**, 3 – 36.

Structured Additive Regression (STAR) models applied in the analysis of breast cancer risk in central Portugal

Elisa Duarte[1], Bruno de Sousa[2], Carmen Cadarso-Suarez[1], Vitor Rodrigues[3], Thomas Kneib[4]

[1] Unit of Biostatistics, Department of Statistics and Operations Research, School of Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain
[2] Faculty of Psychology and Education Sciences, University of Coimbra, Coimbra, Portugal
[3] Faculty of Medicine, University of Coimbra, Coimbra, Portugal
[4] Institute of Statistics and Econometrics, Dept. of Economics, Georg-August-Universität Göttingen, Göttingen, Germany

E-mail for correspondence: duarte.elisa@gmail.com

Abstract: The aim of this study is to analyze how some of the variables considered as risk factors will relate to the probability of having breast cancer. The data set for this analysis was provided by the Portuguese Cancer League (LPCC) and includes the breast cancer diagnosis, the year of birth, breast cancer family history, age of menarche and menopause, reproductive factors and socioeconomic and geographical factors as covariables. Structured Additive Regression (STAR) models were used in order to combine this wide range of covariates and to simultaneously explore possible spatial correlations. The findings of the study shows that early menarche and late menopause ages increase the risk of the disease. This result is in line with recent studies that argue that early menarche and late menopause can increase breast cancer risk by extending women's time span of reproductive years. Regarding the fixed effects we point out the effect of the family history, showing women with sisters, mother or daughters that had breast cancer constitute a risk group.

Keywords: STAR Models; Breast Cancer Risk; Screening Program.

1 Introduction

Understanding the causes of breast cancer is critical when studying the disease. When cancer is developed, it is usually due to the presence of one

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

or a combination of a multitude of risk factors. Some of the risk factors are intrinsic to woman, like age, menarche and menopause ages, family and personal history. Others factors are associated to the environment of a woman and her life choices. The data set for this analysis was provided by the Portuguese Cancer League (LPCC) and consists of women who attended the Breast Cancer Screening Program in central Portugal,for the period between 1990 and 2010. The aim of the study is to analyze how some risk factors will relate to the probability of having breast cancer, using STAR models. The covariates in the study are: year of birth, menarche age, menopause age, pregnancy, nursing status, the use of oral contraceptives, the municipality purchasing power index (PPI) and the spatial correlations of woman place of residence. In addition, we also considered the interaction between menopause and menarche ages.

2 Case Study

The central region of Portugal represents approximately 25% of the population of Portugal. A total of 78 municipalities were considered in this study. The database has women born after 1920, with screening age between 44 and 70. There are 212517 women (76%) who reached menopause. A total of 275753 women (99%) with diagnostic NO (no breast cancer) and 2618 women (only 1%) with diagnostic YES. Oral contraceptives, pregnancy and nursing status are binary variables. Only 7% of women were never pregnant. In this study, 55% of women breastfed and 53% of women used anovulatory. Other important variable is the breast cancer family history. This is coded with levels from 0 to 3 and introduced in the analysis as dummy variables. Level 0 (88%), means the woman has no direct related family with breast cancer, level 1 (6%), if the relatives with this disease were aunties and grandmothers, level 2 (3%) for sisters, and level 3 (3%) for mother or daughters. The quantitative variables in this study were: age of menarche, with a range 8-18 years old and mean equal 13.2 (SD = 1.8) years old; age of menopause varies between 20 and 59 years old and with a mean equals to 48.2 (SD = 5.3) years old; birth year, between 1920 and 1965, with mean of 1949 (SD = 9.3). The municipality code where a woman lives was used for spatial effects analysis. PPI is expressed as an index, with 100 being the municipality's baseline. Municipalities with values under 100 represent regions with lesser economical power.

3 Statistical Methodology

Structured Additive Regression (STAR) models is the framework chosen in this study since it is of the utmost importance to work with models that not only are flexible enough to deal with different and complex structures of data sets, but also are able to consider a multitude of covariates while exploring possible spatial and temporal correlations. In this study, the general structured additive predictor is used to perform a geoadditive mixed

model of the univariate exponential family with binomial response. The geoadditive mixed model terms are fixed effects, continuous and spatial covariates, and can be represented as

$$\eta_i = f_1(x_{i1}) + \ldots + f_k(x_{ik}) + f_{spat}(s_i) + u_i'\gamma,$$

where $f_j(x_{ij}), j = 1, \ldots, k$, are smooth functions of continuous covariates, and $f_{spat}(s_i)$ is a spatially correlated effect of the location s_i where the observation belongs to. The spatial effect is separated into a spatially structured and a spatially unstructured part, thus the predictor equation takes the following form:

$$\eta_i = f_1(x_{i1}) + \ldots + f_k(x_{ik}) + f_{str}(s_i) + f_{unstr}(s_i) + u_i'\gamma.$$

The non-linear effects are estimated based on Bayesian cubic P-splines (Lang and Brezger (2004)). Markov random fields (MRF) and i.i.d. random effects are the prior structures for the estimation of spatially structured effects and spatially unstructured effects, respectively (Fahrmeir and Lang (2001a)). Using an empirical Bayes approach, both fixed effects and random effects are modeled using random variables with appropriate priors and are estimated using a penalized likelihood method in combination with Restricted Maximum Likelihood (REML) for estimating the random effects variances (Fahrmeir et al.(2004), Kneib (2005)).

4 Results

The geoadditive predictor used to perform this analysis is:

$$\begin{aligned}\eta_i \;=\; & \gamma_0 + \gamma_1(famhist1_i) + \gamma_2(famhist2_i) + \gamma_2(famhist3_i) \\ & + \gamma_4(pregnancystatus_i) + \gamma_5(oralcontraceptives_i) \\ & + \gamma_6(nursingstatus_i) + f_1(birthyear_i) + f_2(PPI_i) \\ & + f_3(menopause_i) + f_4(menarche_i) + f_{3|4}(menopause, menarche) \\ & + f_{str}(municipality_i) + f_{unstr}(municipality_i)\end{aligned} \tag{1}$$

The findings for the fixed effects are presented in Table 1. As expected, the use of oral contraceptives increases the risk of breast cancer while pregnancy reduces this risk. However, nursing status appears as a risk factor, while the literature shows it as a protective factor. Women with sisters, mother or daughters with breast cancer constitute a risk group. Figure 1 shows the non-linear effects of the year of birth, menopause age and the purchasing power index. The risk of breast cancer decreases for younger women and increases for late menopause. The PPI effect has a U shape, indicating municipalities with less and higher economical power as a risk factor. The non-linear effect of menarche was not significant. An interesting result has been the interaction effect found between menopause and menarche ages. The corresponding surface and contour plot (Figure 2) indicate a higher risk

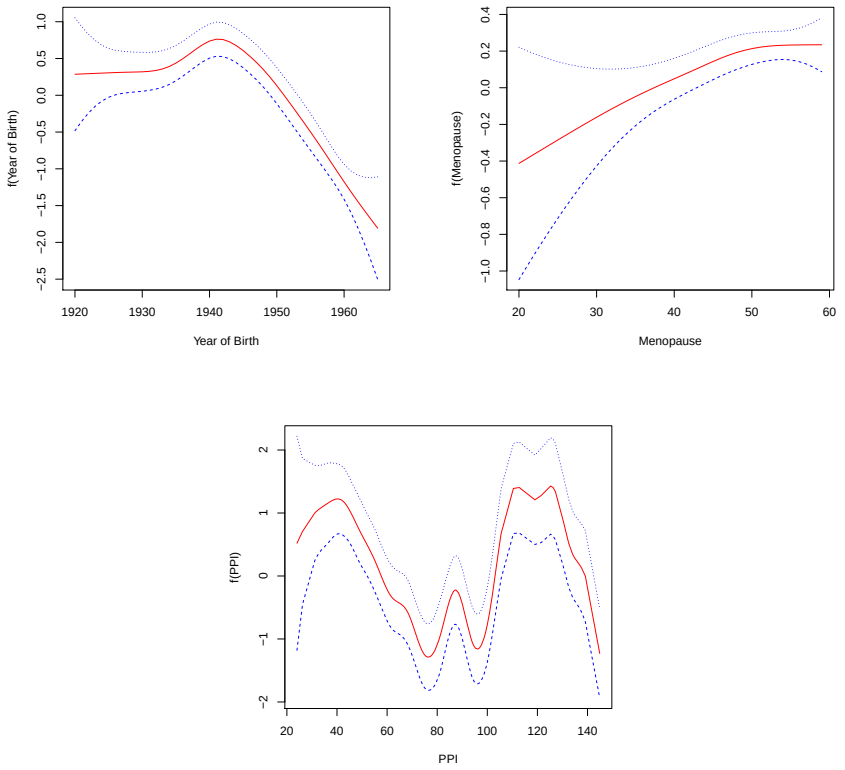


FIGURE 1. Non linear effects of year of birth (top left), menopause age (top right) and PPI (lower).

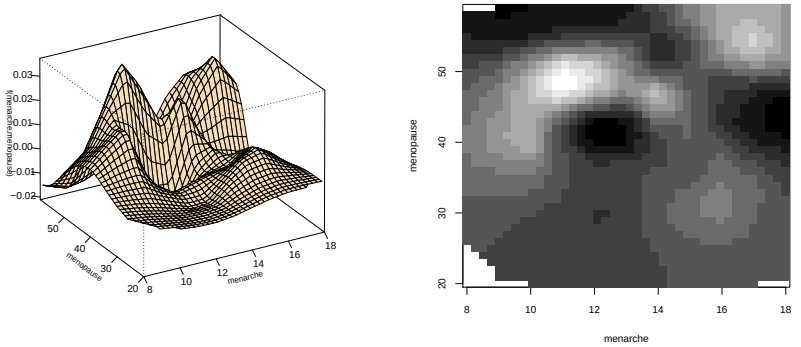


FIGURE 2. Posterior mode estimates for the interaction effect of menopause and menarche.

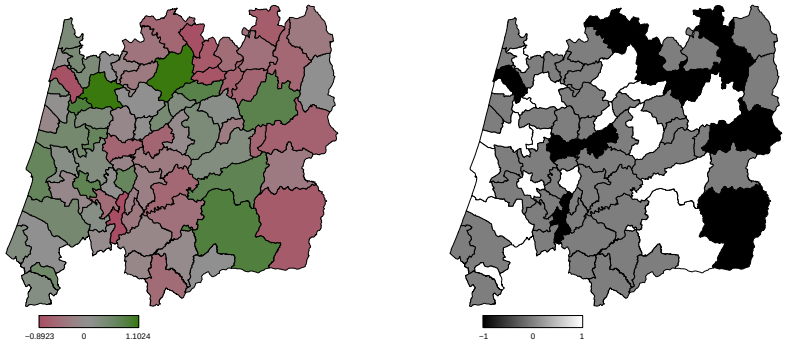


FIGURE 3. Unstructured spatial effects: posterior mode estimates (left), posterior 95 % probabilities (right).

of breast cancer for women with early menarche and late menopause. This result concurs with the premise that women with longer fertility periods may have a higher risk of breast cancer. Although the structured spatial effects show a marked increase in breast cancer risk along the east-west direction (results not shown), the corresponding posterior probability map indicates that such effects are not significant. The unstructured random effects map of Figure 3, points out several municipalities with significant higher risk.

TABLE 1. Estimates, standard deviations (SD), p-values and credible intervals (CI) of fixed effects.

Variable	Estimates	SD	p-value	CI 95%
Oral Contraceptives	0.117	0.051	0.021	(0.018, 0.217)
Nursing status	0.177	0.054	0.001	(0.072, 0.283)
Pregnancy status	-0.343	0.086	<0.001	(-0.512, -0.174)
Family History 1	0.173	0.100	0.085	(-0.024, 0.369)
Family History 2	0.536	0.093	<0.001	(0.353, 0.719)
Family History 3	0.499	0.116	<0.001	(0.270, 0.727)

5 Final comments

The findings of this analysis shows the importance of the screening data, in the study of breast cancer risk factors. In general, the results concurs and strengthen the premises supported by the literature. However some results went against this, namely nursing status. Considered as a protective factor in the literature, is shown in our findings as a risk factor. This interpretation must be carefully taken. Another issue that arose with this

analysis is that low risk of breast cancer can be associated with municipalities with moderate and rich purchasing power. The bigger accessibility of non-screening diagnostic clinics in richer municipalities influence the participation patterns of women, a self-selection bias that can influence the results. In future analysis, we could address this problem by including in the model other variables, such as the participation and detection rates.

Acknowledgments: This work was financed by POPH-QREN, shared by the European Social Fund and national funds MCTES - Portuguese Ministry of Science, Technology and Higher Education through the research project SFRH/BD/64761/2009, and by Spanish Ministry of Research and Innovation through the project MTM2011-28285-C02-01.

References

Fahrmeir L. and Lang S. (2001a). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Fields Priors. *Journal of the Royal Statistical Society, Series C*, **50**, 201 – 220.

Fahrmeir L., Kneib T. and Lang S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 731 – 761.

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

Kneib T. (2005). *Mixed model based inference in structured additive regression.* PhD thesis, Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universität München.

RNA Sequencing and Zipf's Law

Paul H. C. Eilers[1,2], Marco C. A. M. Bink[1]

[1] Biometris, Wageningen, The Netherlands
[2] Erasmus MC, Rotterdam, The Netherlands

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

Abstract: We describe a model for observed frequencies of "read" counts in RNA sequencing, which resembles Zipf's law. We extend it to a Poisson mixture with a latent power law. The model is illustrated on data for soybean, *C. elegans* and humans.

Keywords: Power law, composite link model

1 Introduction

Nowadays, many statisticians will be familiar with basic facts from genetics. DNA, organized in chromosomes, carries genetic information in the form of genes, which is transcribed to RNA, for further translation into proteins downstream. The RNA molecules are characteristic for the gene they belong to and so we can measure the activity of genes by measuring the concentration of their RNA. This activity generally varies over different tissues and time.

A modern technique in this field is high-throughput RNA sequencing, abbreviated as RNAseq. It catches short RNA fragments, called "reads" and determines their sequence of nucleotides, which are mirror images of the familiar A, C, G and T "letters" of the DNA alphabet. If the sequences of a gene is known, the number of matching reads can be counted. This is done for all known genes. It is assumed that RNA concentrations are proportional to the counts. This description of the technology cuts several technical corners, but it will be adequate for our presentation.

For the analysis of RNAseq reads we must use generalized linear models for counts, with proper allowance for over-dispersion. The majority of genes shows low counts (less than 10). While studying these low counts we observed remarkably stable patterns. They are the subject of this paper.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

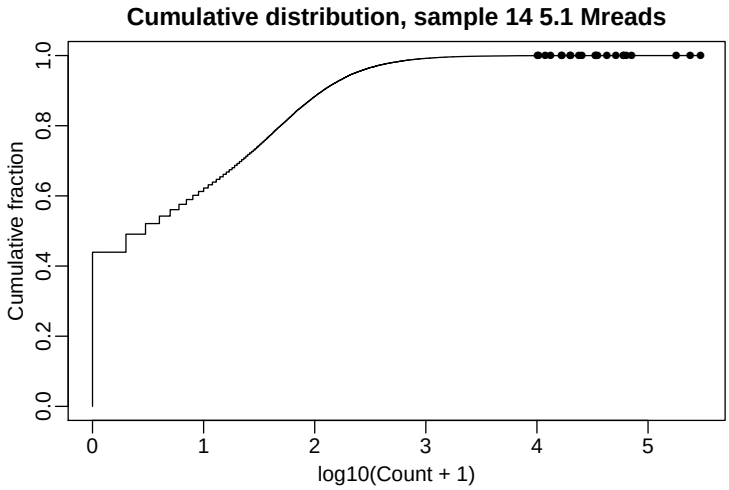


FIGURE 1. Cumulative distribution of the counts in one sample of soybean. The dots indicate the 20 largest counts.

2 Data and models

The total number of reads in an RNAseq experiment can be many millions (M). Our data consist of 14 samples, with 66210 genes, from different parts of the soybean plant (Severin et al, 2010). Among the 14 samples the largest total is 5M, the smallest 1M. Other studies can have hundreds of millions of reads per sample.

Many counts are (very) small and almost half of the genes have zero counts. This is illustrated by the cumulative distribution in Figure 1. The percentage of genes having counts zero, 10 or less, or 100 or less are shown in Figure 2. They show little systematic variation with sample size.

Figure 3 presents histograms for counts from zero to ten, showing a rapid decay. If we plot them on double logarithmic axes, for all 14 samples, we get almost straight lines, quite close to each other.

This pattern indicates that the sum of the frequencies of low counts does not vary much. Indeed the smallest sum is 36516 while the largest is 49528, a ratio of 1.36, where the total counts vary by a factor 5.

The decreasing frequencies remind us of Zipf's law (Zipf, 1932; Young, 2014). Zipf discovered that in natural language the frequencies of words are approximately inversely proportional to their rank. A generalization states that $p_k = ck^{-a}$, for integer $k > 0$, with c the normalizing constant. It follows immediately that $\log p = \log c + a \log k$, a straight line with slope a. In our case, a naive estimate of a can be obtained by linear regression of the logs of the frequencies on the logs of the counts (excluding zero counts). When we do this with a model with a common slope and separate intercepts for each sample (excluding the zero counts), we find $a = -0.73$ (standard error 0.018).

One way to model discrete distributions is to interpret them as a mixture

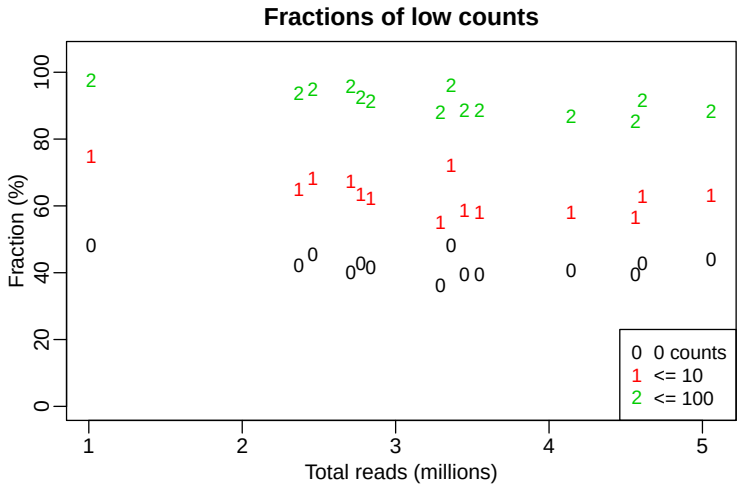


FIGURE 2. Percentages of genes showing low counts (see legend).

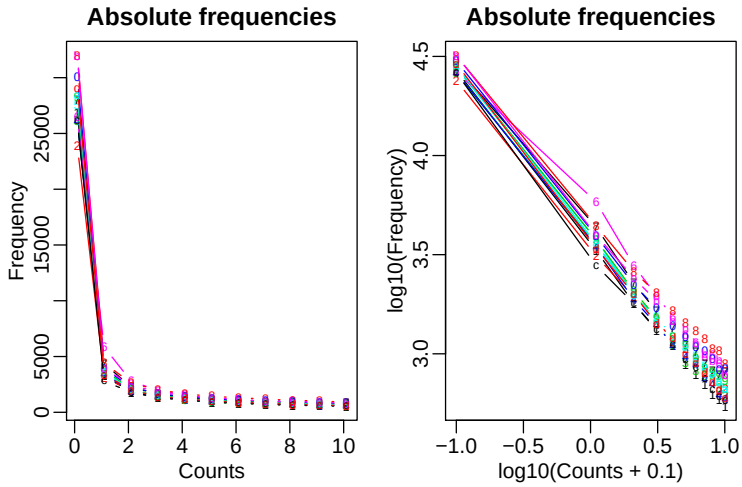


FIGURE 3. Histograms of counts up to 10, on linear (left) and logarithmic scales (right), for all 14 Soybean samples (marked by different colors and symbols).

of Poisson distributions; we will follow that approach here. Let λ be the expected count for an arbitrary gene. We introduce a latent density, following a power law, and imagine a two-stage sampling process. First λ is sampled from $f(\lambda) = c\lambda^{\alpha}$. It determines the expected value of a Poisson distribution, of which the count of reads is sampled. The theoretical distribution of counts is a mixture of Poisson distributions, with mixing density $f(\lambda)$:

$$p_k = c \int_0^{\infty} \lambda^k e^{-\lambda} \lambda^{\alpha} / k! d\lambda.$$

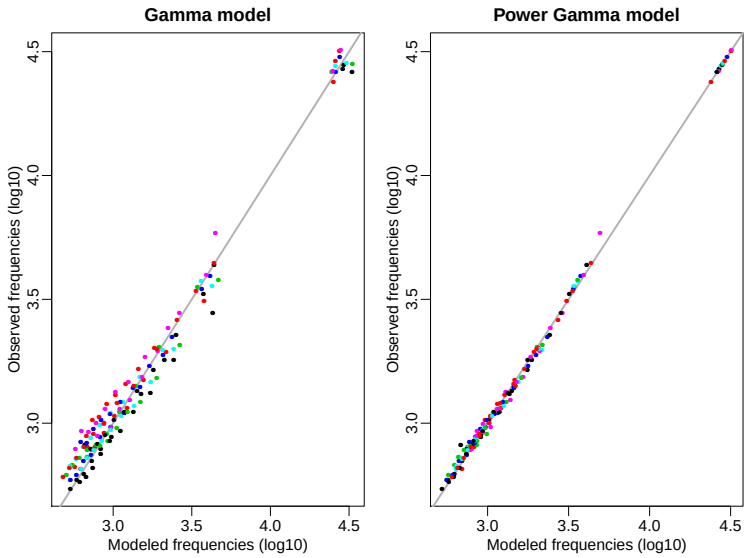


FIGURE 4. Left: fit of the Gamma model to the observed frequencies of counts. Right: fit of the Power Gamma model. In gray the 1:1 line.

The gamma function is defined as $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ and $k! = \Gamma(k+1)$, so we find that

$$p_k = c\Gamma(k+\alpha+1)/\Gamma(k+1).$$

It would be convenient if the normalizing constant c could be determined by integrating λ^α from zero to infinity, but the integral diverges. Because we only study a limited range of values for k, we first set $c = 1$ and normalize the vector p afterwards to sum to 1 over this range.
This model fits quite well. Results are shown in the left panel of Figure 4. The value of α_j was determined for each sample separately by minimizing the deviance

$$D_j = 2\sum_{jk} y_{jk}\log(y_{jk}/\mu_{jk}),$$

where y_{jk} is the frequency of k counts in sample j and $\mu_{jk} = p_k \sum_l y_{jl}$. We find that the value of $\hat{\alpha}$ varies over a small range (-0.87 to -0.82) and shows a strong negative correlation with the total number of reads in a sample.
A generalized linear model with Poisson response and $\log p$ as explanatory variable gives an excellent fit. This amounts to having

$$p_{jk} = c[\Gamma(k+\alpha_j+1)/\Gamma(k+1)]^{\beta_j}.$$

We call it the Power Gamma model. The right panel of Figure 4 shows the much better fit of this model. The value of $\hat{\beta}$ varies between 0.75 and 0.90, with one exception (it is 1.05 for sample 6).
These results are not limited to this Soybean data. In figure 5 we show results for the worm *C. elegans*. The data were found in the Recount database

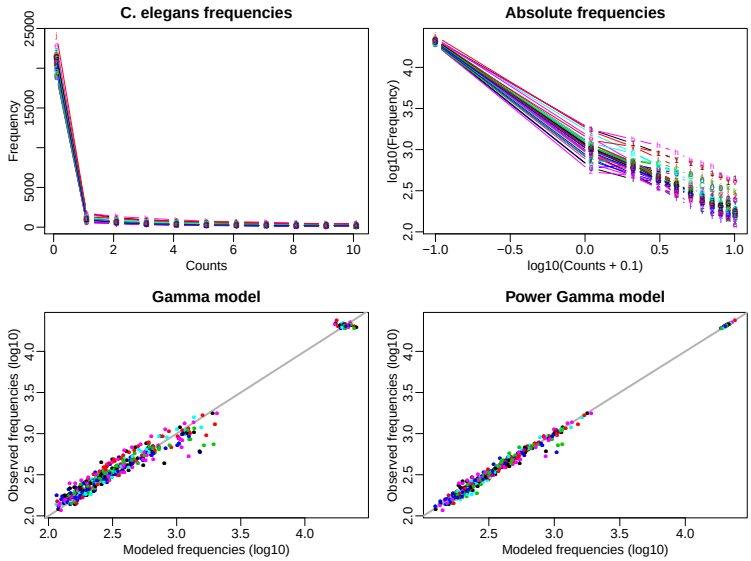


FIGURE 5. Summary of data and results for *C. elegans*.

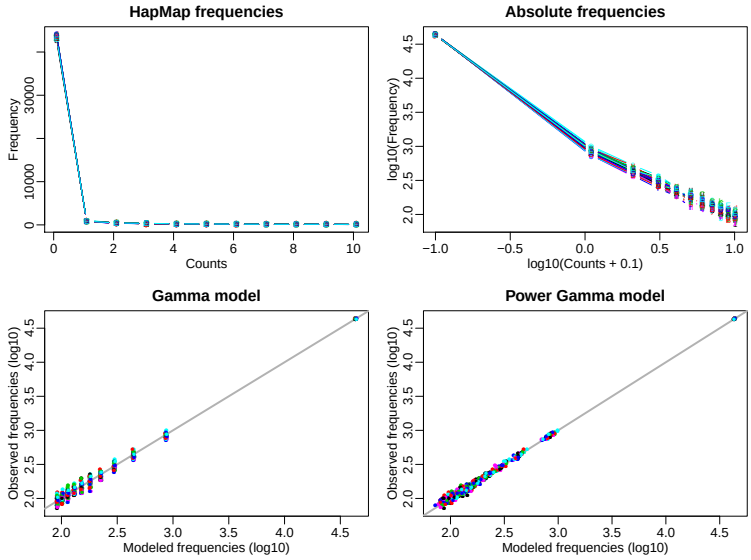


FIGURE 6. Summary of data and results for humans (HapMap).

(Frazee et al. 2011). The total number of reads varied from below 1 to almost 60 million. Figure 6 shows results for human data, based on HapMap cell lines (Cheung et al. 2010), also obtained from ReCount. Here the total number of reads varied from 3 to almost 9 million. The patterns we find are the same as for the soybean data and the power-Gamma model gives an excellent fit in both cases. We have investigated several other data sets,

and got very similar results in each case.

3 Discussion

We have presented an accurate model, inspired by Zipf's law, for the frequencies of low read counts in RNAseq. The model fits well to many data sets. An important issue that has to be resolved, is the mechanism behind the observed distributions. Is it caused by artifacts in the sequencing pipeline, or is it a biological phenomenon, typical for low gene expression levels? We are working on that.

References

Cheung V.G. et al. (2010) Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biology* **8**. doi: 10.1371/journal.pbio.1000480.

Frazee A.C., Langmead B., Leek J.T. (2011) ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics* 12, 449.

Severin A.J. et al. (2010) RNA-Seq Atlas of *Glycine max*: A guide to the soybean transcriptome. *BMC Plant Biology*, **10**, 160.

Zipf G.R. (1932)*Selected Studies of the Principle of Relative Frequency in Language* Harvard University Press.

Young C. (2012) Who is afraid of George Kingsley Zipf? *Significance*, **10:6**, 29–34.

Forecasting cancer mortality figures in Spanish provinces with an ANOVA-type P-spline model

J. Etxeberria[12], M.D. Ugarte[1], T. Goicoa[13], A. Militino[1]

[1] Public University of Navarre, Pamplona, Spain.
[2] Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Spain.
[3] Research Network on Health Services in Chronic Diseases (REDISSEC), Spain.

E-mail for correspondence: jaione.etxeberria@unavarra.es

Abstract: Statistical models for cancer mortality or incidence projections provide invaluable help to epidemiologists and public health managers to evaluate, plan and improve the distribution of resources for cancer prevention, treatment and research. Not for nothing, cancer still remains a major public health concern. Alternative tools have been proposed in the literature to forecast cancer burden. These techniques include age-period-cohort models or joint-point regression models, but much of this work does not include the spatial component and space-time interactions. Very recently P-spline models have been incorporated into the disease mapping toolkit and they have been proved to be a very attractive alternative for smoothing and also forecasting cancer mortality or incidence counts. In this work, an ANOVA-type P-spline model is considered to obtain cancer mortality projections. This model considers explicitly additive smooth terms for space, time, and space-time interactions, splitting the final projections into different components: one spatial, another one temporal and a final term accounting for the contribution of the spatio-temporal interaction to the total projection. An advantage of P-spline models is that they can be reformulated as mixed effects models (or generalized mixed effects models in this context), so that well settled estimation and prediction theory can be applied. The methodology will be illustrated to forecast cancer mortality risks in 50 Spanish provinces.

Keywords: Cancer Mortality; Forecasting; P-spline models.

1 Introduction

Noncommunicable diseases such as cardiovascular diseases, cancer and chronic obstructive pulmonary disease, account for 80% of deaths in the European Region. Cancer is the second leading cause, accounting for nearly

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

20%, and prevention is the only measure to reduce the impact of that diseases. In this context, projections of future cancer incidence and mortality figures play a key role as they are essential to make recommendations for the allocation of prevention services and social programs in national or regional level. Forecasting of cancer mortality trends may be also useful to assess how much progress have done public health interventions (advertising against smoking, screening programs, promoting healthy life-styles, etc.) in reducing cancer rates. Many health agencies such as World Health Organization, European Cancer Registry (EUROCARE) or even the Spanish Ministry of Health, Social Services and Equality use projections of cancer mortality based on statistical models such as temporal Poisson log-linear models, Jointpoint Regression, etc. to set out health strategies for oncoming years.

The inclusion of the spatial component in this forecasting models allows to identify the spatial patterns (maps) of a disease and to show how this spatial pattern is going to evolve with time. For example, Schmid and Held (2004) provide stomach cancer mortality projections using age-period-cohort models including spatial correlation. Very recently Ugarte et al.(2012a) consider a three-dimensional P-spline model to project prostate cancer mortality counts in 50 Spanish provinces.

The goal of this paper is to assess the suitability of an ANOVA-type P-spline model to provide projections of cancer mortality counts and compare the results with different conditional autoregressive models (CAR), P-splines models, and a combination of both in terms of their predictive capacity. This model was initially proposed by Lee (2010) and Lee and Durbán (2011) to estimate smoothed ozone levels in Europe. To technical details, the readers are referred to those references. In this work we focus on extending the model to provide a new forecasting approach in space-time disease mapping. The extension of this model is based on the mixed model reformulation of the ANOVA-type P-spline model requiring some matrix algebra. This model allows to split the projections into a smooth trend common to all regions, a smooth spatial surface constant along the time period, and smooth interaction terms representing the region-specific time evolution of the risks or counts. The mean squared error (MSE) of the predictions and appropriate prediction intervals are also derived allowing the assessment of the coverage probabilities. The technique will be illustrated with Spanish prostate cancer mortality data during the period 1975-2008 because this is the data set used in other papers and comparisons among models will be feasible.

2 Extended ANOVA-type P-spline model

In this work, our interest lies in estimating and forecasting risks and counts for each province ($s = 1, \ldots, S$). Let us define the extended time period for observed and forecast values as $t^* = 1, \ldots, T+1, T+2, \ldots, T+p$, where p are the number of years to forecast. Then, conditional on the unknown

relative risk r_{st^*}, the number of deaths C_{st^*} is assumed to be Poisson distributed with mean $\mu_{st} = e_{st^*} r_{st^*}$, where e_{st^*} is the expected number of deaths calculated on the basis that the sth province in time t behaves as the whole country in the studied period. Then

$$C_{st^*} | r_{st^*} \sim Poisson(\mu_{st^*} = e_{st^*} r_{st^*}), \quad \log \mu_{st^*} = \log e_{st^*} + \log r_{st^*}. \quad (1)$$

Using the extended ANOVA-type model, the log-risk for observed and forecast values are modelled as

$$u_{st^*} = \log r_{st^*} = \delta + f_s(x_1, x_2) + f_t(t^*) + f_{st}(x_1, x_2, t^*) = \mathbf{B}^* \theta^*. \quad (2)$$

where δ can be interpreted as the logarithmic of the overall risk; $f_s(x_1, x_2)$ represents the smooth spatial effects constant along the period; $f_t(t^*)$ is a extended temporal trend common to all areas, and $f_{st}(x_1, x_2, t^*)$ is the extended interaction term that can be interpreted as the specific temporal trend for each area (see for example Ugarte et al., 2012b). In these expressions x_1 and x_2 are the coordinates of the centroid of the ith small area (longitude and latitude respectively), t^* is the time (for observed and forecast values), and f_i, $i = s, t, st$ are smooth functions to be estimated using P-splines with B-spline bases. $\mathbf{B}^*$ is the extended B-spline basis and θ^* is a vector of coefficients.

To ensure that f_i, $i = s, t, st$ are smooth functions, the P-spline approach places penalties on the coefficients θ^*. The extended penalty matrix $\mathbf{P}^* = diag(\mathbf{0}, \mathbf{P}_s, \mathbf{P}_t^*, \mathbf{P}_{st}^*)$ is defined by a block-diagonal matrix whose components are penalties for the two-dimensional spatial component, the one dimensional time component and the three-dimensional component (space-time interactions). The key point in this process is to choose an extended transformation matrix preserving the original transformation matrix $\mathbf{T}$ used to fit the data (see for more detail Lee and Durbán, 2011; Ugarte et al., 2012b). Based on the transformation matrix $\mathbf{T}$ given in the previous citations, and extended transformation matrix $\mathbf{T}^*$ is proposed is this work. Using this extended transformation matrix and the results in Gilmour et al.(2004) about predictions with mixed models, the generalized mixed model reformulation of the extended ANOVA-type P-spline model can be obtained.

3 Illustration

To illustrate results, Spanish prostate cancer mortality data from 1975 to 2008 are considered. This data set has been described elsewhere (see Ugarte et al., 2012a; Ugarte et al., 2012b; Etxeberria et al., 2013) to study different disease mapping models in terms of smoothing and forecasting. This facilitate comparisons with the ANOVA-type P-spline model considered in this work.

The Dawid-Sebastiani score (Dawid and Sebastiani, 1999) is used to compare the P-spline ANOVA type model with a pure interaction P-spline

model and spatio-temporal CAR models in three, two and one-year-ahead predictions. For three year ahead predictions, the score for the pure interaction P-spline model outperforms the P-spline ANOVA type model. These models also provide better scores for two-year ahead predictions, but similar values for one-year-ahead projections. In general, the P-spline ANOVA type model produces an overcoverage that is not observed with a pure interaction P-spline model and CAR models. This can be attributed to wider prediction intervals because of higher mean squared error. The main reason to explain this higher mean squared error is that the number of variance components (smoothing parameters) in the P-spline ANOVA type model is higher. However, the model allows to decompose the prediction into different parts: one spatial, one temporal common to all regions, and one specific for each area. Hence it is possible to detect if the contribution of each part make the risks/counts prediction increase or decrease. This is relevant for health researchers and authorities to make hypothesis about factors in specific regions that contribute to increase or decrease the risks.

4 References

Dawid, A.P. and Sebastiani, P. (1999). Coherent Dispersion Criteria for Optimal Experimental Design. *The Annals of Statistics* **27**, 65-81.

Etxeberria J., Goicoa T., Ugarte M.D., and Militino A.F. (2013). Evaluating space-time models for short-term cancer mortality risk predictions in small areas. Submitted to *Biometrical Journal*,

Gilmour A., Cullis B., Welham S., Gogel B., and Thompson R. (2004). An efficient computing strategy for prediction in mixed linear models. *Computational Statistics and Data Analysis*, **44**, 571-586.,

Lee D.J. (2010). Smoothing mixed models for spatial and spatio-temporal data Ph-D Thesis, Carlos III University, Madrid.,

Lee D.J. and Durbán M. (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, **11**, 49-69.,

Schmid, V. and Held, L. (2004). Bayesian extrapolation of space-time trends in cancer registry data. *Biometrics*, **60**, 1034-1042.,

Ugarte. M.D., Goicoa, T., Etxeberria, J. and Militino, A.F. (2012a). Projections of cancer mortality risks using spatio-temporal P-spline models. *Statistical Methods in Medical Research* **21**, 545-560.,

Ugarte. M.D., Goicoa, T., Etxeberria, J. and Militino, A.F. (2012b). A P-spline ANOVA type model in space-time disease mapping. *Stochastic Environmental Research and Risk Assesment* **26**, 835-845.,

Change-point estimation in piecewise constant regression models with random effects

Salvatore Fasola[1], Vito M.R. Muggeo[1]

[1] Dipartimento di Scienze Statistiche e Matematiche, University of Palermo, Italy

E-mail for correspondence: salvatore.fasola@unipa.it

Abstract: We propose an iterative algorithm to estimate change-points in general regression models. The algorithm avoids grid search to obtain maximum likelihood estimates, and thus it guarantees moderate computational time regardless of the sample size and the number of change-points to be estimated. Furthermore, it allows estimation in random effects models, where grid search is unfeasible. We present the proposed approach in practice by analyzing variations of lung functionality on a sample of transplant recipients.

Keywords: change-points, piecewise constant, grid search algorithm.

1 Introduction

Change-point detection is an important goal of many statistical analyses with applications in several fields, including Biology, Ecology and Economics. One of the most known applications concerns array-based comparative genomic hybridization (Pinkel et al., 1998), where knowing change-point locations is crucial to identify possibly damaged genes involved in cancer and other diseases; in Economics interest lies in detecting structural changes, namely time points where one or more covariate effects change abruptly. Several approaches have been proposed to address this problem within the likelihood framework, such as segmentation techniques (Olshen et al., 2004), segmented regression (Muggeo and Adelfio, 2011), penalized regression (e.g., Rippe, Meulman and Eilers, 2012), or the well-known dynamic grid search algorithm (Bai and Perron, 2003) having a computational cost equal to $O(n^2)$ for any number of change-points. While grid search is the 'usual' approach, particularly in economics and econometrics, there are at least two issues to be emphasized. First, sample size is still a concern, and, therefore, estimation with huge datasets could represent a practical

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

limitation. Second, and more importantly, grid search cannot be employed when dealing with individual piecewise constant profiles, namely when the underlying regression model includes subject-specific parameters modelled by random effects (e.g. Jackson and Sharples, 2004). We propose an iterative algorithm to carry out estimation of general regression models with unknown change-points. We discuss the presented algorithm in the simple context of a single-break piecewise constant relationship, and apply it to a model with random change-points, where the Bayesian paradigm is usually employed.

2 Methods

For the sake of simplicity, we consider a simple, only fixed effects, model with a shift and relevant change-point only in the mean level. Let Y be the response variable with $\mu_i = E(Y_i|x_i)$, related to the quantitative covariate x through the following regression equation

$$\mu_i = \beta_0 + \beta_1 I(x_i > \psi). \tag{1}$$

At ψ the mean level of Y shifts instantaneously from β_0 to $\beta_0 + \beta_1$: our goal is to set up an iterative algorithm to obtain all the unknown model parameters (β_0, β_1, and ψ) efficiently. Our proposal relies on

$$I(x_i > \psi) \equiv \frac{1}{2}\frac{x_i - \psi}{|x_i - \psi|} + \frac{1}{2},$$

for $x_i \neq \psi$. This identity, when placed in (1), after some simple manipulations gives

$$\mu_i = \beta_0 + \beta_1 z_i(\tilde{\psi}) + \gamma w_i(\tilde{\psi}), \tag{2}$$

where $\gamma = -\beta\psi$. Note the auxiliary (or 'working') covariates are

$$z_i(\tilde{\psi}) = \left(\frac{1}{2} + \frac{1}{2}\frac{x_i}{|x_i - \tilde{\psi}|}\right) \quad \text{and} \quad w_i(\tilde{\psi}) = \left(\frac{1}{2}\frac{1}{|x_i - \tilde{\psi}|}\right), \tag{3}$$

with $\tilde{\psi}$ meaning an approximate value. Notice model (1) has been converted in the simple linear model (2). Formulas above suggest a simple iterative algorithm: starting from initial guesses for the change-point, compute the auxiliary covariates (3), fit the working linear model (2), and update the change-point estimate via

$$\tilde{\psi} = -\frac{\tilde{\gamma}}{\tilde{\beta}}. \tag{4}$$

While the outline of the algorithm is quite simple, there are two main pitfalls that should be warned. First, the (profile) log-likelihood for the change-point is a step function, with typically many local optima; second, x_i values close to $\tilde{\psi}$ may cause computational troubles, since denominators in (3) go to zero. We skip details, but moving the x_is away from the change-point value $\tilde{\psi}$ according to some adjusting factor, allows to circumvent both

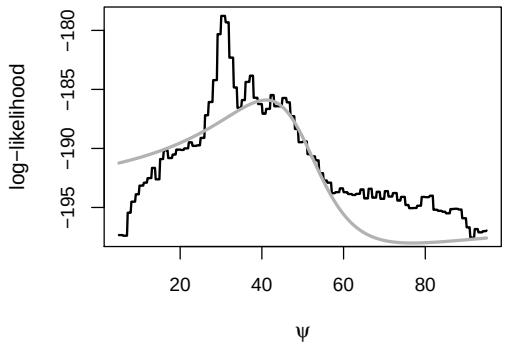


FIGURE 1. Profile log-likelihoods for ψ in a toy dataset. Black line: objective log-likelihood for model (1), with global optimum at 30. Grey line: log-likelihood for the working model (2), with starting value $\tilde{\psi} = 50$, and relevant optimum at 41 to be used to compute the working covariates at the next iteration.

problems, to some extent. Figure 1 tries to portray as the algorithm works in a toy dataset.
Notice how the working linear model (2) leads, at each step, to a smooth objective function, with a unique solution to be used as starting value in the next iteration.

3 Application to random effect modelling

One of the most noteworthy advantages of the aforementioned algorithm is that it straightforwardly extends to piecewise linear mixed models with random change-points; here the grid-search algorithm can not be employed and the only feasible approach appears to be the Bayesian one. We discuss how the proposed algorithm allows inclusion of random effects also in the change-point parameter within the likelihood based framework. For $i = 1, 2, \ldots, n$ subjects each with $j = 1, 2, \ldots, n_i$ measurements, the model reads as

$$\begin{aligned} y_{ij} &= \beta_{0i} + \beta_{1i} I(x_{ij} > \psi_i) + \epsilon_{ij} \\ &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) I(x_{ij} > \psi + p_i) + \epsilon_{ij}. \end{aligned} \tag{5}$$

The subject specific parameters β_{0i}, β_{1i} and ψ_i are given by the sum of fixed and random effects.
Using the idea of the previous section, we transplant the *nonlinear* mixed effect model (5) into a conventional *linear* mixed model

$$y_{ij} = \beta_{0i} + \beta_{1i} z_{ij}(\tilde{\psi}_i) + \gamma_i w_{ij}(\tilde{\psi}_i) + \epsilon_{ij} \tag{6}$$

that is fitted iteratively. At each iteration, change-point estimates are updated as in (4), for each subject. At convergence, change-points can be

expressed as linear parameters of an additional LMM; using the new auxiliary covariate

$$w'_{ij} = -\tilde{\beta}_{1i} w_{ij}(\tilde{\psi}_i)$$

the final model is

$$y_{ij} = \beta_{0i} + \beta_{1i} z_{ij} + \psi_i w'_{ij} + \epsilon_{ij}, \tag{7}$$

where in the z_{ij}s and w'_{ij}s the dependence on previous estimates has been omitted. This allows to obtain subject specific change-point and also the relevant variance parameter estimates.

Jackson and Sharples (2004) analyzed data from $n = 204$ patients receiving lung transplant: in the first months after the transplant patients have an high risk of complications, such as rejection episodes and infections, and thus lung conditions, evaluated via the forced expiratory volume in 1 second (FEV_1), need to be monitored constantly. For each subject, different measurements are available with decline patterns being smooth or changing suddenly. Unlike Jackson and Sharples (2004) relying on the Bayesian paradigm, we model such data in a likelihood based framework and apply the aforementioned algorithm to fit a regression model with random effects in the regression and change-point parameters as well. More specifically, the model we fit takes the form (5), where y_{ij} is the j-th FEV_1 measurement (as baseline percentage) for patient i, and x_{ij} is the month at which the measurement is taken after the transplant. For simplicity we assume the random effects for intercept (b_{0i}), shift (b_{1i}) and changepoint (p_i) to be Gaussian and independent, namely

$$\begin{bmatrix} b_{0i} \\ b_{1i} \\ p_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_\psi^2 \end{bmatrix} \right),$$

and, as usual, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ independently. We focus only on subjects having an abrupt change in their FEV_1 profiles; Table 1 shows fixed effect and variance parameter estimates for the fitted model. Fixed parameter estimates indicate that high values of FEV_1 immediately after the transplant are followed by an important drop ($\hat{\beta}_1 = -42.88$) occurring, on average, after 40 months. However, the variance parameters emphasize considerable heterogeneity among subjects especially in time of occurrence of dropping (via the variance component σ_ψ^2) and relevant amount of dropping (via σ_1^2). Such heterogeneities are well appreciated in Figure 2 that illustrates observed trajectories and relevant fitted profiles for some subjects under study: the quite different locations of changepoints and mean levels after the changepoints themselves reflect the high variance estimates reported in Table 1.

4 Conclusions

We have discussed an iterative algorithm to fit models with change-points. While the proposed algorithm is quite general and can be exploited even

TABLE 1. Parameter estimates for the piecewise constant regression model with random change-points (5) fitted iteratively via the working model (6).

Parameter	Estimate	St.Err.
β_0	97.89	2.78
β_1	−42.88	5.00
ψ	40.02	5.44
σ_0^2	9.30	-
σ_1^2	17.00	-
σ_ψ^2	18.42	-

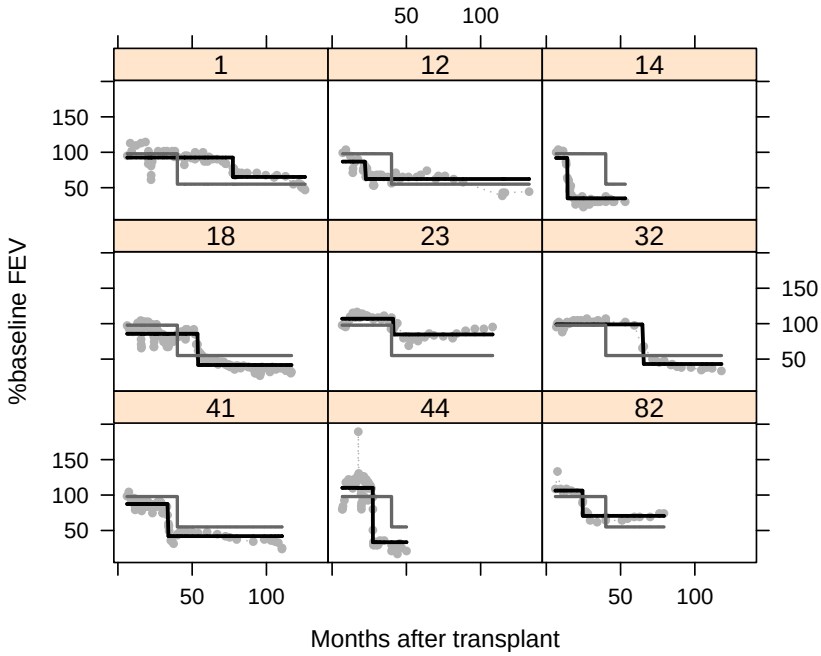


FIGURE 2. Observed and fitted piecewise constant profiles for some patients under study. The grey lines represent the fixed effect estimates and the black lines the subject specific estimates.

for the 'simple' regression models in the spirit of Bai and Perron (2003), we have discussed it within the mixed model framework where no paper previously published in literature addresses the problem from a likelihood view. Interestingly, the proposed algorithm is able to provide fixed effects and subject-specific predictions of the change-points in a likelihood based framework. The algorithm appears to work reasonably well in practice, but there are some aspects to be investigated, in particular computations of standard errors for all model parameter estimators. For instance, the stan-

dard errors reported in Table 1 are based on the usual information matrix: some simulations have shown that such information-based standard errors do not work adequately for the change point estimator, and thus other methods should be exploited. For instance, the bootstrap could represent a possible alternative. Testing for the non-zero change-point variance component or for the existence of change-point itself, also represent challenging topics to be investigated in detail.

References

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, **18**, 1 – 22.

Jackson, C. H. and Sharples, L. D. (2004). Models for longitudinal data with censored changepoints. *Applied Statistics*, **53**, 149 – 162.

Muggeo, V.M.R. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, **27**, 161 – 166.

Olshen, A.B. et al. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557 – 572.

Pinkel, D. et al. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, **20**, 207 – 211.

Rippe, R.C.A., Meulman, J.J., and Eilers, P.H.C. (2012). Visualization of genomic changes by segmented smoothing using an L_0 penalty. *PloS one*, **7**, e38230.

Composite smooth estimation of the state price density implied in option prices

Gianluca Frasso[1], Paul H.C. Eilers[2]

[1] Institut des sciences humaines et sociales, Univ. de Liége, Belgium.
[2] Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands.

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

Abstract: We propose a new semi-parametric approach for the estimation of the state price density (SPD) implied in option prices. Our procedure is based on the penalized composite link model (PCLM) and ensures smooth and arbitrage-free estimates.

Keywords: No arbitrage conditions, PCLM, Option pricing, SPD.

1 Introduction

In computational finance the state prices density (SPD) is a fundamental tool in option pricing tasks. It describes the perceived uncertainty about the future value of an option contract. There is no way to observe the SPD directly. Instead it has to be estimated from quoted option prices.
We propose a semi-parametric approach. The observed option prices can be described as expected values of possible pay-offs at maturity weighted by the unknown SPD. We model the logarithm of the latter as a smooth function while matching the expected values of (possible) pay-offs with the observed prices. This leads to a special case of the penalized composite link model (Eilers, 2007).

2 Composite smooth estimation of the SPD

Imagine that we offer you the 'option' to buy a given asset (e.g. a stock) at a future date T (maturity) paying a given price k (strike price). Intuitively, if $s < k$ at T you will not exercise the option given that the value of the underlying asset is lower than the price you originally agreed to pay for it. Of course the situation is opposite in the case $s > t$ (positive pay-off). How

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

much (c) would you pay today for this contract giving you the right to buy an asset of uncertain value in the future (at time to maturity $\tau = T - t$)? Intuitively this price should take into account this uncertainty about the pay-off $s - k$ at expiration date (a part form the the cost of money). In other words the fair price of this contract should be equal to the discounted expected value of the possible pay-offs computed under an appropriate density function $f(s)$: the state price density. More formally:

$$c = \exp(-r\tau) \int_0^\infty (s-k)^+ f(s) ds, \tag{1}$$

where r is the interest rate guaranteed by a risk-free asset (e.g. a saving account).
Unfortunately we do not observe $f(s)$ directly but only the prices c quoted on the market. The idea is to estimate $f(s)$ starting from the (c, k) pairs observed at a given time. This is an inverse problem and it is ill-conditioned, meaning that the data do not uniquely determine $f(s)$. We assume that $f(s)$ is positive and smooth. To estimate it we suppose that $f_j = f(u_j) = \exp(\eta_j)$ and that the ith observed option price follows the model:

$$\begin{aligned} y_i &= \mu_i + \epsilon_i = \boldsymbol{G} \exp(\boldsymbol{\eta}) + \boldsymbol{\epsilon} \\ &= \textstyle\sum_j (u_j - k_i)^+ \exp(\eta_j) + \epsilon_i, \end{aligned} \tag{2}$$

with ϵ_i i.i.d. random variables with null mean and constant variance $\sigma^2 I$ and $u = \{u_1, ..., u_M;\ u_j \in [\min(\boldsymbol{k}) - \gamma, \max(\boldsymbol{k}) + \gamma]\}$ is a fine grid of spot prices (γ is a constant that useful to enlarge the set of plausible pay-offs). In Eq. (2), $\boldsymbol{G}$ is a composition matrix defined as:

$$\boldsymbol{G} = \begin{bmatrix} (u_1 - k_1)^+ & (u_2 - k_1)^+ & \cdots & (u_M - k_1)^+ \\ (u_1 - k_2)^+ & (u_2 - k_2)^+ & \cdots & (u_M - k_2)^+ \\ \vdots & \cdots & \ddots & \vdots \\ (u_1 - k_N)^+ & (u_2 - k_N)^+ & \cdots & (u_M - k_N)^+ \end{bmatrix}.$$

In addition, we need to ensure that $\sum_{j=1}^M \hat{f}_j = 1$. To incorporate this constraint, we extend the $\boldsymbol{G}$ matrix with a row vector of large constants ξ (e.g. 10^5), and add an extra element to the vector of observed call prices also having value ξ.
The vector $\boldsymbol{\eta}$ can be estimated via (penalized) IRWLS (Eilers, 2007) by minimizing:

$$S = \|\boldsymbol{c} - \boldsymbol{\mu}\|^2 + \lambda \|\boldsymbol{D}\boldsymbol{\eta}\|^2,$$

where $\boldsymbol{D}$ is a matrix forming third order differences. The mean function can be linearized in a neighborhood $\widetilde{\mu}_i$ as follows:

$$\mu_i = \widetilde{\mu}_i + \sum_j \frac{\partial \widetilde{\mu}_i}{\partial \eta_j} \Delta\eta_j = \widetilde{\mu}_i + \sum_j G_{ij} \exp(\widetilde{\eta}_j) \Delta\eta_j.$$

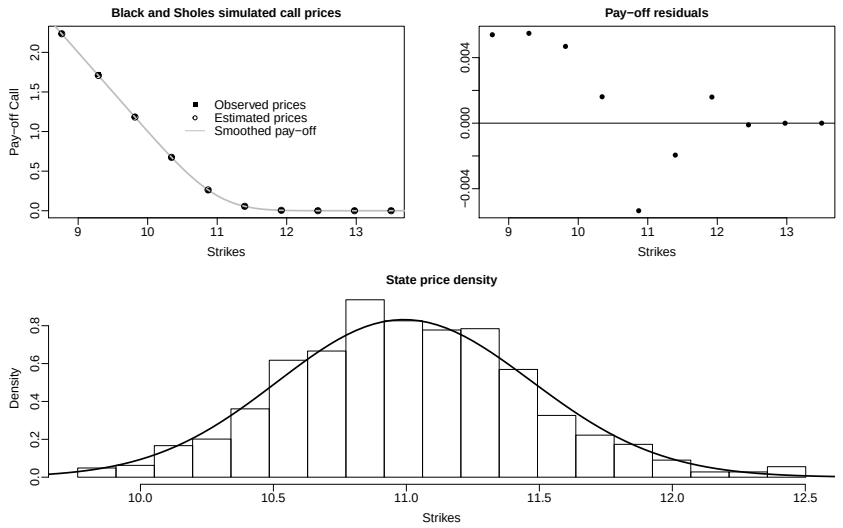


FIGURE 1. Composite smooth estimation of the SPD for a set of prices of an European vanilla call option simulated under the B&S model.

The estimates are then obtained by iteratively solving the set of normal equations:

$$\left(\widetilde{\boldsymbol{E}}^\top \widetilde{\boldsymbol{E}} + \lambda \boldsymbol{D}^\top \boldsymbol{D}\right) \boldsymbol{\eta} = \widetilde{\boldsymbol{E}}^\top \left(\boldsymbol{c} - \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{E}}\widetilde{\boldsymbol{\eta}}\right),$$

where $\boldsymbol{E} = \boldsymbol{G}\,\mathrm{diag}\,(\exp(\boldsymbol{\eta}))$ and a tilde as in $\widetilde{\boldsymbol{\eta}}$ indicates an approximation to the solution. The estimation process usually converges quite fast (less than 10 iteration in most cases). The roughness penalty $\lambda\|\boldsymbol{D}\boldsymbol{\eta}\|^2$ ensures smoothness of $\boldsymbol{\eta}$. The parameter λ tunes the degree of smoothness of the final fit and can be selected by exploiting the link between penalized regression and mixed models (Ruppert et al., 2003). Indeed, taking $\lambda = \sigma_\epsilon^2/\sigma_{PEN}^2$ the optimal balance between goodness of fit and degree of smoothing can be found through an EM-type algorithm, i.e. by updating till convergence

$$\sigma_\epsilon^2 = (N - ED)^{-1}\,\|\boldsymbol{c} - \boldsymbol{\mu}\|^2 \text{and } \sigma_{PEN}^2 = (ED)^{-1}\,\|\boldsymbol{D}\boldsymbol{\eta}\|,$$

with the effective dimension defined as $ED = \mathrm{tr}[(\boldsymbol{E}^\top\boldsymbol{E} + \lambda\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{E}^\top\boldsymbol{E}]$ (in analogy with Hastie and Tibshirani, 1990).

Figure 1 shows the estimates obtained for a set of call prices simulated under the Black and Scholes (1973) model (with $k \in [\$8.02, \$14.61]$ and an observed spot price $s = \$10$ and $\tau = 180$ days).

The model in Eq. (2) guarantees estimates that are consistent with the no-arbitrage conditions (Harrison et al., 1981):

- Positiveness of the estimated pay-off: $\hat{\boldsymbol{c}} \geq 0$.
- Monotone decay: $\dfrac{\partial \hat{\boldsymbol{c}}(k)}{\partial k} < 0$.
- Convexity: $\dfrac{\partial^2 \hat{\boldsymbol{c}}(k)}{\partial k^2} \geq 0$.

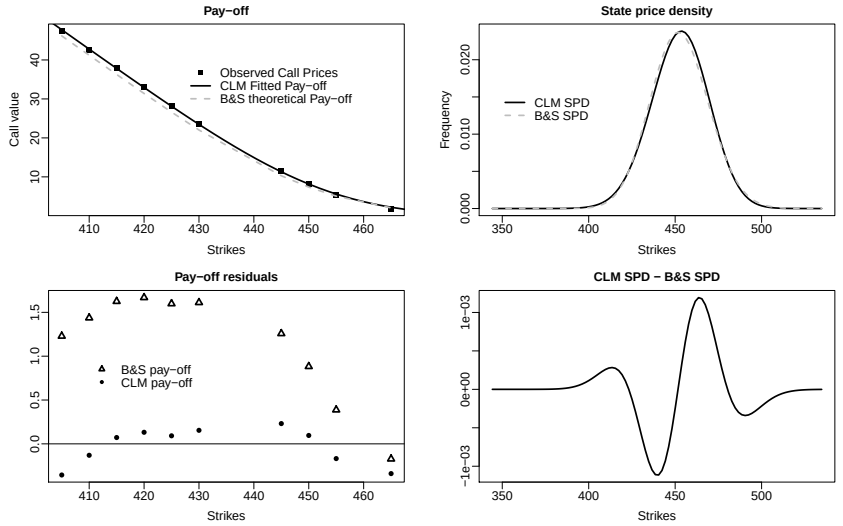


FIGURE 2. Smoothed pay-offs and extracted SPDs of the S&P 500 call option with maturity 53 days. The estimated call prices and state price density are compared with the ones consistent with the Black and Scholes log-normality hypothesis.

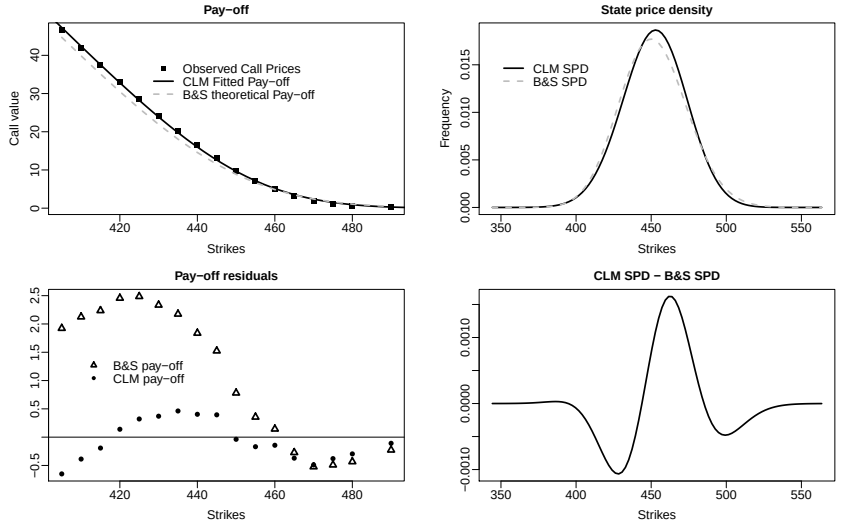


FIGURE 3. Smoothed pay-offs and extracted SPDs of the S&P 500 call option with maturity 80 days. The estimated call prices and state price density are compared with the ones consistent with the Black and Scholes log-normality hypothesis.

- Proper density: $\sum_{j=1}^{M} \exp(\eta_j) = 1$.

The prices estimated using the model in Eq. (2) are positive by definition. The monotonicity condition is also satisfied. Indeed, for two strike prices $k_i \leq k_{i+1}$ the matrix $\boldsymbol{G}$ is such that the number of non-zero elements is higher in row i than in row $i+1$. The vector $\hat{\boldsymbol{c}}$ is computed as positive linear combination of strictly convex functions and this ensures the convexity of the estimated pay-off. The fourth condition requires that the estimated SPD is a proper density. This is imposed by augmenting the convolution matrix and the vector of observed prices with a large ξ vector (constant) as described above.

3 A real data example

In this section we apply of our model to estimate the SPDs implied in the observed prices of two S&P 500 call options (Carmona, 2004). Figure 2 and figure 3 show the results obtained for contracts expiring in 53 and 80 days (values of the index equal to \$449.3 and \$446.79). The estimates have been obtained using a grid of 100 possible spot prices ($\boldsymbol{u}$). The estimated call prices and state price density are compared with the ones extracted under the Black and Scholes (1973) log-normality assumption. It clearly appears that the model properly fits the observed data. For a larger time to maturity the quality of the call prices approximation obtained following the Black and Scholes model is lower.

4 Discussion

We have introduced a new semi-parametric approach for the estimation of the state price density implied in option prices. Our proposal takes advantage from the penalized composite link model framework. The parametrization of the regression model, together with the definition of a composition matrix consistent with the option pay-off function, ensures estimated prices consistent with the theoretical no-arbitrage constraints. We have shown the efficiency and flexibility of the proposed approach using real and simulated data.

Our approach offers many opportunities for future research. The regression framework can be extended to estimate bivariate multivariate SPDs, using tensor products P-splines. The second dimension could be time to maturity or it could be used to model intra-day variation.

Instead of a full two-dimensional density estimate, it is possible to model a "mother density" which is scaled and shifted. We have investigated this idea for different times to maturity, with promising results.

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

References

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. Journal of political economy, 81(3):637.

Harrison, J. M. and Pliska, S. R. (1981). Martingales and stochastic integrals in the theory of continuous trading. Stochastic Processes and their Applications, 11(3):215 - 260.

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

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

Carmona, R. (2004). Statistical Analysis of Financial Data in S-Plus. Springer Text in Statistics.

Eilers, Paul H.C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. Statistical Modelling, 7(3):239 - 254.

Modeling Binary Functional Data with Application to Animal Husbandry

Jan Gertheiss[1,2], Verena Maier[3], Engel F. Hessel[1], Ana-Maria Staicu[4]

[1] Department of Animal Sciences, Georg-August-Universität Göttingen, Germany
[2] Center for Statistics, Georg-August-Universität Göttingen, Germany
[3] Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany
[4] Department of Statistics, North Carolina State University, Raleigh, USA

E-mail for correspondence: jan.gertheiss@agr.uni-goettingen.de

Abstract: We propose a functional logistic regression approach allowing us to model binary but functional measurements by assuming latent but smooth subject-specific profiles. The method also allows to incorporate additional covariates. We use the method to analyze feeding records of pigs.

Keywords: Categorical Functional Data; Generalized Additive Models; Marginal Models; Pig Fattening.

1 Introduction

We observe a group of 127 pigs over a period of about 100 days. On a very dense grid of time points, it is recorded when each pig is feeding. The data is coming from the PIGWISE project funded by the European Union within the ICT-AGRI 2010 call for transnational research projects. The objective of the project is to "optimize the performance and welfare of fattening pigs using High Frequent Radio Frequency Identification (HF RFID) and synergistic control on individual level".
HF RFID is used to record feeding times of the pigs. More precisely, HF RFID antennas were installed above the troughs to identify feeding pigs fitted with one or two passive RFID tags on their ears. The HF RFID system at the trough registered the presence of the tags when they came within range of the antenna (Madelyne et al., 2014). This leads to binary functional data for each pig and day, as for each time point it is recorded whether the pig is feeding or not. Let $y_{ij}(t) = 1$ if pig i is feeding at time t at day j, and 0 otherwise.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

Movements of the pig during feeding, however, can move the ear tag in and out of the range of the antenna. For these reasons, consecutive RFID registrations of an ear tag will display irregular time gaps between readings (Madelyne et al., 2014). Therefore the data were downsampled to consecutive sampling intervals of 10 seconds. As sometimes a pig is just passing by the trough but not feeding, a pig was considered as feeding when the respective tag was registered at least twice within a 10 second interval. So on a very dense and regular grid of time points $t_1, t_2, t_3, \ldots, t_{8640}$ binary observations $y_{ij}(t_r) \in \{0, 1\}$ are available for pig i across day j, saying whether the pig is feeding ($y = 1$) or not ($y = 0$).
In addition to the feeding data, there are measurements such as temperature and humidity available that may influence the pigs behavior. One important objective of the data analysis is to find pig-specific feeding profiles telling us when a certain pig is typically feeding. (1) These profiles can be used for summarizing and illustrating the data observed. (2) They are a potential basis for further data analysis and smart usage of the HF RFID technology. On the one hand, this technology is intended for monitoring pigs and identifying pigs showing unusual feeding behavior since this may indicate problems such as sickness. For identifying "unusual" behavior, however, we first need to known the usual feeding behavior of a pig. On the other hand, once the feeding profiles are obtained, we can use them as a basis for further data analysis, for example, for clustering pigs or comparing groups of pigs.
For analyzing the feeding data, we propose a functional logistic regression approach allowing us to model the binary but functional measurements by assuming underlying smooth pig-specific profiles. The method also allows to incorporate additional covariates (such as temperature and humidity).

2 Methods

In the simplest case, the probability $\pi_{ij}(t)$ that pig i is feeding at time t at day j is modeled as

$$\pi_{ij}(t) = \frac{\exp\{\eta_{ij}(t)\}}{1 + \exp\{\eta_{ij}(t)\}}, \tag{1}$$

with $\eta_{ij}(t) = \alpha_i(t)$, and $\alpha_i(t)$ denoting the underlying profile of pig i of interest. By extending $\eta_{ij}(t)$, however, the method also allows to incorporate additional functional covariates such as temperature $x_j(t)$ and humidity $u_j(t)$, or non-functional covariates. Here, the specification we propose has the form

$$\eta_{ij}(t) = \alpha_i(t) + \beta_{1i} x_j(t) + \beta_{2i} u_j(t) + \beta_{12i}\{x_j(t)u_j(t)\}.$$

The interaction term is included as the effect of temperature may be different for different values of humidity. Since each pig may react differently to changes in temperature and humidity, the corresponding regression coefficients β_{1i}, β_{2i}, and β_{12i} are pig-specific.

The observations made at time points t_r are now used to estimate the pig-specific profiles $\alpha_i(t)$ and effects of temperature and humidity. The observations made for one pig are of course dependent. According to the theory of generalized estimation equations (Liang and Zeger, 1986; Zeger and Liang, 1986), however, we can use a working independence assumption to estimate the parameters, as we are just interested in the marginal effects of the latent profiles (and other covariates) on feeding behavior. For instance, we are not going to predict whether pig i is feeding at time t given it was feeding (or not) at time t'. Primary interest is in estimating the profiles. Those are assumed as smooth and equal for each day and hence equal at the beginning (0h/12 a.m.) and at the end of the day (24h/12 p.m.). For function $\alpha_i(t)$, we therefore use a cyclic cubic regression spline where the ends of the function match up to the second derivative (see, e.g., Wood, 2006). Estimation is carried out by R package `mgcv` (Wood, 2006).

3 Simulation Studies

Before applying our method to the data it is tested in a small simulation study. In each simulation run, we generate a true underlying profile using (similar to Tutz and Gertheiss, 2010)

$$\alpha(t) = \sum_{k=1}^{5} (b_k \sin(t\pi(5 - b_k)/150) - m_k),$$

with $t \in [0, 300]$, $b_k \sim U(0, 5)$, and $m_k \sim U(0, \pi)$. Then binary data points $y_j(t_r)$ are generated at equidistant points $t_r \in [0, 300]$, $r = 1, \ldots, 300$, using probabilities $\pi_j(t) = \exp\{\alpha(t)\}/(1+\exp\{\alpha(t)\})$. We consider three different sample sizes $n = 1, 10, 100$, that is, $j = 1, \ldots, n$.

We compare our spline method to a very simple method just using interpolated relative frequencies of 1s at time points t_r, and a smoothed version of these frequencies using a smoothing spline with 20, 40, or 80 degrees of freedom. Performance of the different methods is evaluated via the squared error

$$\int_0^{300} \left(\hat{\pi}(t) - \frac{\exp\{\alpha(t)\}}{1 + \exp\{\alpha(t)\}} \right)^2 dt,$$

where $\hat{\pi}(t)$ is the (feeding) probability at time t estimated by the method considered. This procedure of profile and data generation, model fitting and evaluation is repeated 50 times. Figure 1 shows the errors obtained for the different methods. It is seen that the errors produced by our spline method (SME) are usually very small. Simple relative frequencies are much worse (see RFE), performance of the smooth version heavily depends on the degrees of freedom (see SRFE20, SRFE40, SRFE80). The penalty parameter for our spline method is chosen via the Un-Biased Risk Estimator (UBRE; Craven and Wahba, 1978).

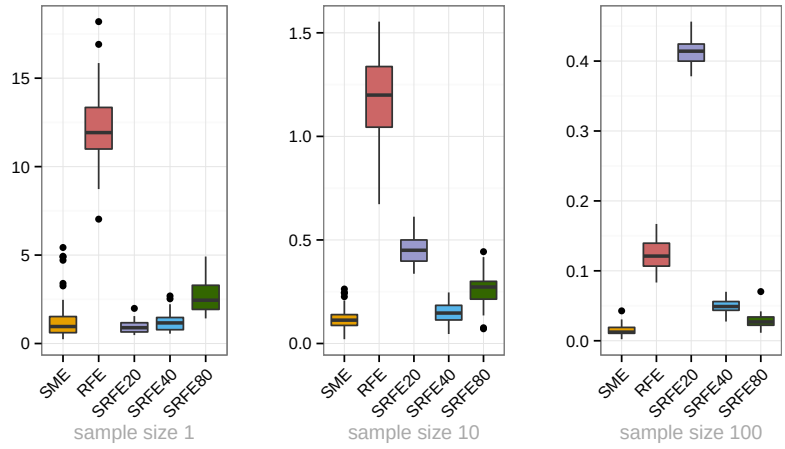


FIGURE 1. Mean squared error with respect to the latent profile estimated by different methods.

4 Application to Animal Husbandry

For our real data, UBRE cannot be used for determining the penalty parameter because it assumes that the observations are independent. With our data this leads to far too small penalty parameters and hence extreme under-smoothing. Similar problems arise with methods like GCV or REML. We hence use pig-specific K-fold cross-validation on a daily basis. More precisely, we use folds of entire days, and for each day j in the test set the integrated Brier score (IBS)

$$\int_{0\text{h}}^{24\text{h}} (y_{ij}(t) - \hat{\pi}_{ij}(t))^2 \, dt, \tag{2}$$

is calculated, with $\hat{\pi}_{ij}(t)$ being the estimated (marginal) probability that pig i is feeding at time t (given in seconds since 0h) at day j. For each pig minimization of the IBS is done separately, producing pig-specific smoothing parameters.

For one specific pig, Figure 2 shows the estimated (marginal) probabilities of feeding at time $t \in [0\text{h}, 24\text{h}]$ at day 93–102. The differences in the curves result from differences in temperature and humidity (note, the latent profile is assumed to be equal for each day). The profile and the effects of temperature and humidity were estimated using the data from days 1–92 only. On the x-axis the actually observed feeding times are marked. We see that the probability curve is in relatively good accordance with these observed values. The pig shown here usually feeds in the morning and afternoon, which is the somewhat natural feeding behavior of pigs.

In order to compare the method proposed to some alternative and simpler approaches, we fit the feeding probabilities for each pig and day 93–102 (using days 1–92 only). For each pig i and day j, we calculate the integrated

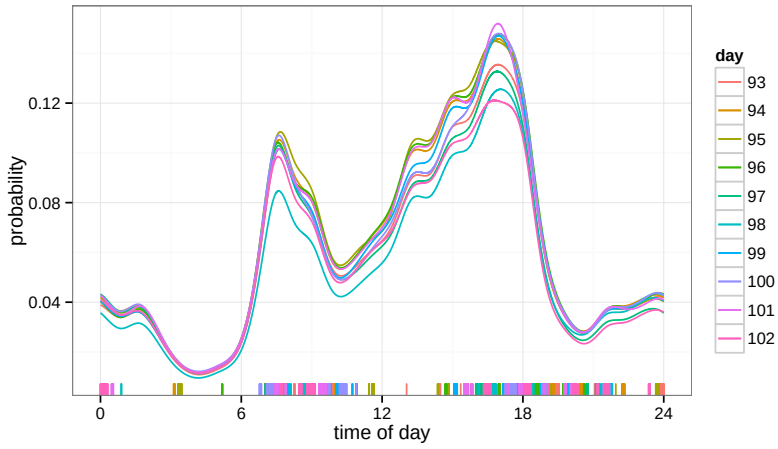


FIGURE 2. Estimated (marginal) probabilities for feeding as a function of time for one specific pig and ten days, together with observed feeding times marked at the x-axis.

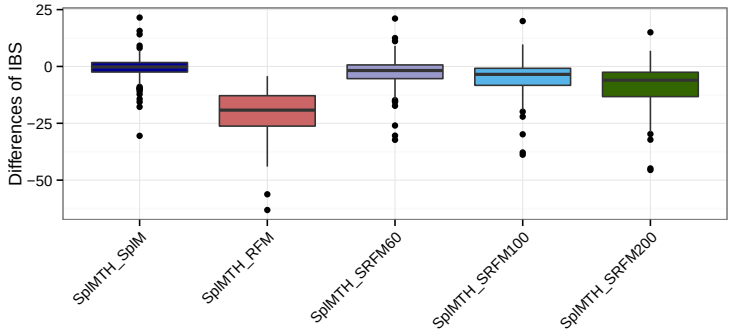


FIGURE 3. Prediction performance in terms of the integrated Brier score of the proposed method compared to some alternative approaches.

Brier score (2). The score is then averaged across days to have one score per pig. Figure 3 summarizes the pigwise differences between our spline method with additional predictors temperature and humidity (SplMTH) and competing simpler methods. These methods are a spline method as specified at (1) but without temperature and humidity (SplM), in analogy to the simulation studies, a very simple method just using interpolated relative frequencies of feeding at time points $t_0, t_1, \ldots, t_{8640}$ (RFM), and a smoothed version of these frequencies using a smoothing spline with 60, 100, or 200 degrees of freedom (SRFM60, SRFM100, and SRMF200, respectively). We see that our method is distinctly superior to the simpler methods, but accuracy does not really benefit from including temperature and humidity.

Acknowledgments: The results presented are generated in the framework of the ICT-AGRI era-net project PIGWISE "Optimizing performance and welfare of fattening pigs using High Frequent Radio Frequency Identification (HF RFID) and synergistic control on individual level" (Call for transnational research projects 2010). The German contribution was funded by the German Federal Office for Agriculture and Food (BLE).

References

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377 – 403.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13 – 22.

Maselyne, J., Saeys, W., De Ketelaere, B., Mertens, K., Vangeyte, J., Hessel, E.F., Millet, S., and Van Nuffel, A. (2014). Validation of a High Frequency Radio Frequency Identification (HF RFID) system for registering feeding patterns of growing-finishing pigs. *Computers and Electronics in Agriculture*, in press.

Tutz, G. and Gertheiss, J. (2010). Feature extraction in signal regression: a boosting technique for functional data regression. *Journal of Computational and Graphical Statistics*, **19**, 154 – 174.

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

Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121 – 130.

Comparing the predictive performance of different regression models for longitudinal and time-to-event data

Ipek Guler[1], Christel Faes[2], Francisco Gude[3], Carmen Cadarso-Suárez[1]

[1] Unit of Biostatistics, Department of Statistics and Operations Research. University of Santiago de Compostela, Spain
[2] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Belgium
[3] Clinical Epidemiology Unit. Hospital Clinico Santiago de Compostela, Spain

E-mail for correspondence: ipek.guler@usc.es

Abstract: Many follow-up studies produce different types of outcomes including both longitudinal measurements and time-to-event outcomes. Commonly, it is of interest to study the association between them. To estimate this association, an extended version of the Cox model with longitudinal covariates (Anderson and Gill, 1982) or a two-stage approach (Self and Pawitan, 1992) can be used. However, these techniques have several limitations, including the possibility of biased estimations. To solve these limitations, joint modeling approaches for longitudinal and survival data were proposed in the recent statistical literature (Rizopoulos, 2010; Phillipson et al, 2012). This paper compares the predictive accuracy performances of these modeling approaches to study the longitudinal and time-to-event survival processes together. The predictive performance of these models is assessed using time-dependent ROC curves (Heagerty et al, 2005). All the statistical approaches were applied to a biomedical database on Primary Biliary Cirrhosis data.

Keywords: Joint model; longitudinal data; time-dependent ROC curves; time-dependent AUCs.

1 Introduction

There exist various methods to study the association between a longitudinal outcome and the survival process in the literature. The earliest methods are the extended Cox model (Anderson and Gill, 1982) and a two-stage approach method (Self and Pawitan, 1992). Although these methodes have

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

advantages in terms of fast computing, they also have several limitations. The extended Cox model assumes that the covariates are external and not related to the failure mechanism (Kalbfleisch and Prentice, 2002); also, this model does not take into account the measurement error of the longitudinal process. In the two-stage approach no survival information is used for the longitudinal process such that informative drop-out is not accounted for. If the main interest is on the association between the longitudinal and survival data, joint models are required to feature this correlation. Joint models have gained increasing attention over the last two decades, especially in biomedical investigations. In this paper, we compare the predictive performances of two recent approaches of joint modeling by Rizopoulos (2010) and by Philipson et al (2012) with alternative methods like the extended Cox model (Anderson and Gill, 1982) and the two-stage approach (Self and Pawitan, 1992).
All models were used to analyse the data of a study on Primary Biliary Cirrhosis (Fleming and Harrington, 1991). The objective is to analyse the effect of the longitudinal measures of serum bilirubin (*serBilir*) on the patient survival. The database includes a status indicator of 312 patients (*status2*), a treatment indicator if the patient have a "placebo" or "D-penicil" (*drug*) and an indicator of hepatomegaly. The trajectories of serum bilirubin are shown in Figure 1.

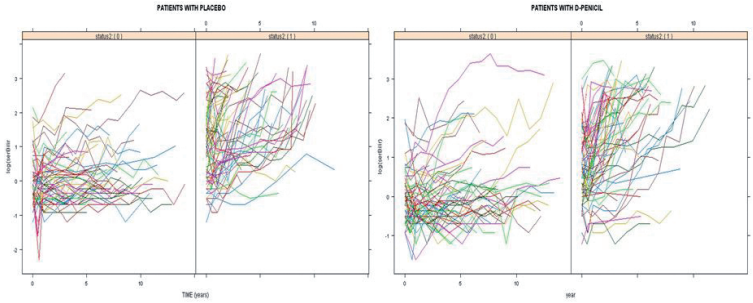


FIGURE 1. Longitudinal trajectories for failed and censored patients separately for patients with placebo and with D-penicil.

2 Statistical models

In this section we introduce four different regression models to study the longitudinal process with time to event survival for the PBC dataset, considering a model selection according to predictive accuracy.

2.1 The extended Cox Model (Anderson and Gill, 1982)

One of the models that allows incorporation of time dependent covariates into the survival model is an extension of the Cox Model introduced by Anderson and Gill, 1982. In our biomedical study, this model can be described

as

$$h_i(t) = h_0(t)R_i(t)\exp(\gamma_1 \text{drug}_i + \gamma_2 \text{hepatomegaly}_i + \alpha \log(\text{serBilir})_i(t))$$

where γ is $R_i(t)$ is a left continuous at risk process with $R_i(t) = 1$ if subject i is at risk time at time t and $R_i(t) = 0$ otherwise, $\text{serBilir}_i(t)$ is encoded using (start-stop) notation. These points indicate the time intervals of the recorded serum bilirubin measurements.

2.2 The two-stage approach (Self and Pawitan, 1992)

This model is described in two stages:

(i) A linear mixed effects model is fitted to estimate the true longitudinal process (Pinheiro and Bates, 2000)

$$\begin{aligned} \log(\text{serBilir})_{ij} &= \text{Intercept} + \beta_1(\text{drug}(t_{ij})\text{year}(t_{ij})) \\ &\quad + \beta_2(\text{hepatomegaly}(t_{ij})) + U_{0i} + U_{1i}t_{ij} + \epsilon_i(t_{ij}) \quad (1) \end{aligned}$$

where t is year, β_1 and β_2 is the fixed effects coefficients, U_{0i} is the random intercept and U_{1i} is the random slope parameter.
(ii) The random effects of the model (1) are incorporated into the survival sub-model as covariates in the following manner

$$h_i(t) = h_0(t)\exp\gamma_1 \text{drug}_i + \gamma_2 \text{hepatomegaly}_i + \gamma drug_i + \alpha_0 U_{0i} + \alpha_1 U_{1i}\text{year}_i$$

where U_{0i} is the random effect of intercept and U_{1i} is the random effect of slope.
The main advantages of two-stage models include fast computing and existing software. However they may lead biased inference because of estimating parameters in the first stage based only observed covariate data. That is, the trajectories of the longitudinal data who experience an event may be very different from those who do not.

2.3 Joint Model I (Rizopoulos, 2010)

This model focuses on the survival model in which an individual's survival is influenced by a longitudinal outcome that is measured with error. In this approach the longitudinal sub-model is the same as the two-stage approach (1) and the survival sub-model is expressed as:

$$h_i(t) = h_0(t)\exp(\gamma_1 \text{drug}_i + \gamma_2 \text{hepatomegaly}_i + \alpha \log(\text{serBilir})_i(t))$$

where $\log \text{serBilir}_i(t)$ is true (unobserved) value of longitudinal outcome.

2.4 Joint Model II (Phillipson et al, 2012)

In this model an informative censoring is assumed for the longitudinal variable and the focus is on the both processes. The longitudinal sub-model is described by the two-stage approach (1) and the survival sub-model is expressed as:

$$h_i(t) = h_0(t)\exp(\gamma_1 \text{drug}_i + \gamma_2 \text{hepatomegaly}_i + \gamma drug_i + \alpha_0 U_{0i} + \alpha_1 U_{1i}\text{year}_i)$$

3 Results

The extended Cox model is fitted by Anderson Gill model (1982), two-stage approach is fitted by a linear mixed effects model (Pinheiro and Bates, 2000) and a survival Cox model (Cox, 1972). The joint modelling approach of Rizopoulos (2010) and Philipson et al (2012) are fitted to the dataset with their implementations through the JM package in R written by Rizopoulos (2010) and joineR package written by Philipson et al (2012). Furthermore, the linear predictors of each model at time t is used to calculate the incident sensitivity and dynamic specificity to compute the ROC curves and the area under curve for each time point (Heagerty et al, 2005). This calculation is implemented through risksetROC package in R.

The results obtained for each regression model, given in Table 1, show a statistically significant association between the longitudinal and survival process with different coefficients. According to these results the patients with higher serum bilirubin tend to have a worse survival. However, as we can observe in this table, the two-stage model over-estimates the effect of the slope ($\alpha_1 = 6.29$) as compared with the joint model approach ($\alpha_1 = 1.39$) and has a lower discrimination after a certain time point (Figure 3). The extended Cox model has lower discrimination at all time points and a decrease in standard errors of coefficients compared with the others as a result of ignoring the longitudinal process measured with error leading to biases estimates.

Joint model approaches show better discrimination according to time dependent AUC values (Figure 3) in comparison with the extended Cox model and the two-stage approach. Also the log likelihood values of two joint model approaches are less than those obtained from both the extended version of the Cox model and the two-stage approach.

4 Discussion

This study showed that when the longitudinal data and survival process are correlated, a joint model approach is most appropriate to analyse this relationship in comparison with the extended Cox model and the two-stage approach.

Also for the joint model, we used two different approaches implemented in software R, JM by Rizopoulos (2010) and joineR by Philipson et al (2012). The joineR package permits to analyse the intercept and slope effect to survival separately and JM package has several options to study the longitudinal and survival sub-models with more flexibility. Further study is required to analyse the longitudinal trajectories with flexible linear mixed models and to compare different models for the survival process (e.g., Weibull, piecewise and spline method) already implemented in JM package.

Acknowledgments: This work is partially financed by Spanish Ministry of Economy and Competitiveness (MTM 2011-28285-C02-01).

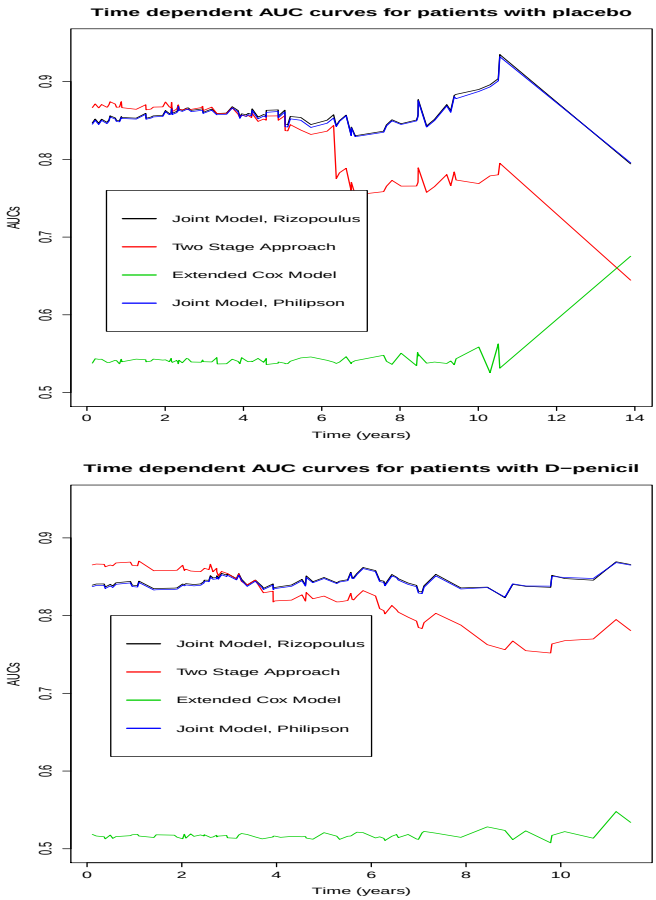


FIGURE 2. Time dependent AUCs for each model separately for patients with placebo and with D-penicil.

References

Anderson, P.K., Gill, R.D. (1982). Cox's Regression Model for Counting Process: A Large Sample Study. *Annals of Statistics* **10**, 1100 – 1120.

Heagerty, P.J., Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, **61**(1), 92 – 105.

Heagerty, P.J., Saha-Chaudhuri, P. (2012). risksetROC: Riskset ROC estimation from censored survival data.

Kalbfleisch, J., Prentice, R. (2002). *The Statistical Analysis of Failure Time Data.* Hoboken, New Jersey: John Wiley and Sons.

Philipson, P., Sousa, I., Diggle, P., Williamson, P., Kolamunnage-Dona, R., Henderson, R. (2012). joineR: Joint modelling of repeated measurements and time-to-event data.

Longitudinal Process				
	Cox Model	Two-stage	Joint Model I	Joint Model II
	Coef. (SE)	Coef. (SE)	Coef. (SE)	Coef. (SE)
Intercept	0.56(0.08)	0.56(0.08)	0.55(0.08)	0.40(0.06)
Drug	0.13(0.11)	0.13(0.11)	0.13(0.11)	0.14(0.11)
Year	0.17(0.01)	0.17(0.01)	0.19(0.01)	0.18(0.01)
Drug*Year	0.004(0.02)	0.004(0.02)	0.009(0.02)	0.0008(0.02)
Loglikelihood	−1534.73	−1534.737	-	-
Event Process				
Drug	0.15(0.18)	0.06(0.15)	0.05(0.16)	0.10(0.20)
Hepatomegal	0.41(0.24)	0.56(0.17)	0.66(0.17)	0.72(0.17)
serBilir (α)	1.07(0.09)		1.21(0.08)	
(α_0)		0.96(0.10)		1.13(0.09)
(α_1)		6.29(0.68)		1.39(0.19)
Loglikelihood	−705.7366	−736.1549	−1960.943	−2420.385

TABLE 1. Fitted models of different approaches.

Pinheiro, J., Bates, D. (2000). *Mixed Effects Models in S and S-plus.* Springer.

Rizopoulos, D. (2010). JM: an R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, **35**(9), 1 – 33.

Self, S., Pawitan, Y. (1992). Modelling a marker of disease progression and ofset of disease.In:*AIDS Epidemiology:Methodological Issues of Eds. Jewell, N.P., Dietz, K. Farewell, V.T. Boston*: Birkhauser.

Measures of Interaction for Relationships Among Dichotomous Variables

Karl Heiner [1], John Hinde [2]

[1] State University of New York at New Paltz, USA
[2] The National University of Ireland Galway, Ireland

E-mail for correspondence: kwheiner@aol.com

Abstract: In this paper we model data from a 2^6contingency table. The categorical variables have to do with individuals with HIV. Practitioners are interested in predicting an outcome, so we have fit logistic regression models that include interaction terms. For the practitioners, the interpretation of the interaction terms can be difficult. Interval estimates for these interpretations are seen to require unreasonable assumptions and approximations. We suggest a method for developing confidence intervals for the common measures of interaction. For these data, we find that models that measure relationships among all of the variables (i.e., log-linear models) provide a better description.

Keywords: loglinear models; logit models; interaction measures.

1 Introduction

The purpose of this paper is to demonstrate the application and interpretation of various models for describing the relationship among variables associated with care of individuals infected with the human immunodeficiency virus (HIV). Models used to describe these data included logistic regression models, log-linear models and Bayesian models for categorical data. Of particular interest is the interpretation of interaction parameters within these models and the relationship of parameters among models. The dichotomous factors in the 2^6 table were age group, gender, the clinical setting in which care was delivered, whether or not a mental health problem was identified by the care giver, whether or not patients consistently adhered to their treatment regimen and whether or not patients experienced viral load suppression for a year.
One naturally might be interested in asking if viral load suppression can be predicted from the other five variables. For example, one may ask whether the presence of a mental health issue is predictive of viral load suppression.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

Logistic regression models are frequently used for this type of analysis. On the other hand, a log-linear model may shed light on the possibility that viral load suppression, or lack thereof, is related to mental health and the possibility that a patient with a mental health issue is less likely to adhere to his treatment regimen which in turn could impact viral load.
The first order interaction terms in a logistic regression model are related to the second order interactions in a log-linear model (Aitkin, et. al., 2009). The preference of interpretation of the interaction terms in a logistic regression model is not consistent among applied researchers (e.g. epidemiologists, economists, educators) and assessing precision of the estimates of the interaction parameters appears to be unsettled.
We review some suggestions that have been offered in the literature and propose what we believe to be a sensible approach. Should we measure interaction on an additive scale, as is typical in fields like engineering and psychology where models predominately employ the identity link, or on the multiplicative scale, so as to be consistent with the interpretation of parameters estimated in logistic regression, which is common in epidemiology, biostatistics and economics.

2 Fitting some Models

In fitting log-linear models, we began with the saturated model and reduced this by removing interaction terms that were not significant. The only factor interacting with the setting variable was age; older individuals were more likely to receive their care in hospitals as opposed to clinics. In subsequent modeling we assumed that setting was essentially an alias for age. Our final model showed that younger individuals were more likely to be male, more likely to reveal a mental health issue and more likely to have a problem adhering to their treatment regimen. Mental health issues were associated with gender, age, problems with adherence, and to a lesser extent with viral load suppression. Adherence problems and viral load suppression were strongly associated, as one might suspect. Figure 1 depicts the final model where we have made some mild epidemiological assumptions about causation (e.g., mental health does not have a causal impact on gender), a practitioner may conclude that good adherence facilitates viral load suppression, mental health issues may lead to adherence problems and that women and elderly individuals living with HIV are less likely to have mental health issues, the later less likely to have adherence problems.
A logistic regression model was fit specifying viral load suppression as the response variable and the other five dichotomous variables and their first order interactions as predictors. Three interaction terms were significant, gender:age, gender:setting and age:mental health. It is common for practitioners to exponentiate the main effect parameters to obtain odds ratios and to a somewhat lesser extent and perhaps not so wise, to interpret these odds ratios as risk ratios.
The interpretation of odds ratios as risk ratios is apparently justified by the fact that when the response variable is rare, these two ratios will be

approximately the same. How rare does the response indicator have to be in order for this approximation to apply and how much difference does the distinction matter in some applied fields, e.g. public health administration? Here viral load suppression indicator is not rare, being yes slightly more than half of the time.

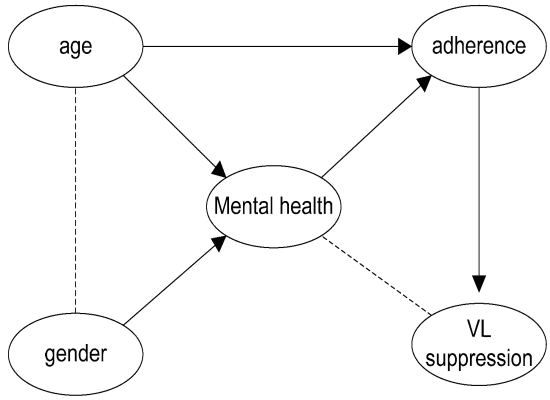


FIGURE 1. Graphical representation of log-linear model fit

3 Measures of Interaction in Logistic Regression

Suppose we have two potentially predictive dichotomous factors and we wish to express the probability of the presence of some outcome (a third indicator variable that we wish to predict) under each of the four combinations of the levels of the two factors. We may call the probability of the indicator variable (response variable) being one at the i^{th} level of factor A and the j^{th} level of factor B, p_{ij}.

3.1 Multiplicative Measures of Interaction

Define the risk for the first and second levels of factor A to be $(p_{11} + p_{12})/(p_{11}+p_{12}+p_{21}+p_{22})$ and $(p_{21}+p_{22})/(p_{11}+p_{12}+p_{21}+p_{22})$, respectively, and similarly for factor B. The risk ratio comparing the risk at the first and second level of A is then $(p_{11} + p_{12})/(p_{21} + p_{22})$.
The odds in favor of a 1 on the response variable at the first and second levels of factor A would be $(p_{11} + p_{12})/(1 - (p_{11} + p_{12})$ and $(p_{21} + p_{22})/(1 - (p_{21} + p_{22})$. The odds ratio comparing the odds at the first and second level of A is $((p_{11} + p_{12})/(1 - (p_{11} + p_{12}))/((p_{21} + p_{22})/(1 - (p_{21} + p_{22}))$ which is approximately the risk ratio if the p_{ij}'s are very small.
Within level one and two of factor B, the risk ratios would be respectively p_{11}/p_{21} and p_{12}/p_{22} and the ratio of risk ratios $(p_{11}p_{22})/(p_{21}p_{12})$,

and similarly for ratios of odds ratios, $(\mathrm{OR}_{11}\mathrm{OR}_{22})/(\mathrm{OR}_{21}\mathrm{OR}_{12})$. These are multiplicative measures of interaction.

3.2 Additive Measures of Interaction

Kalilani and Atashhili (2006) review some approaches for measuring interactions, e.g. linear contrasts (LC), interaction contrast ratios (ICR) or equivalently relative excess risk due to interaction (RERI), attributable proportion due to interaction (AP) and a synergy index (S) measuring the ratio between combined effect and individual effects. ICR, AP, and S are considered additive measures in as much as that they are ratios involving linear contrasts. These authors agree with the argument that interactions measured on an additive scale are more associated with biological interactions than interactions measured on a multiplicative scale. They point out that in epidemiology, interaction describes the extent the joint effect of two factors differs from their independent effects and that interaction effects are the remainder of subtracting the marginal effects from the joint effect in an additive measure.
The linear contrast (LC) or interaction contrast (IC) is $LC = IC = (p_{11} - p_{21}) - (p_{12} - p_{22})$ which is the way we think about interactions in experimental design. The interaction contrast ratios (ICR) or equivalently relative excess risk due to interaction (RERI) is $ICR = IC/\mathrm{RR}_{11} = \mathrm{RR}_{22} - \mathrm{RR}_{21} - \mathrm{RR}_{12} + 1$. The attributable proportion due to interaction (AP) is defined to be $AP = IC/\mathrm{RR}_{22} = (\mathrm{RR}_{22} - \mathrm{RR}_{21} - \mathrm{RR}_{12} + 1)/\mathrm{RR}_{22}$ and the synergy index (S) measuring the ratio between combined effect and individual effects is given by $S = (\mathrm{RR}_{22} - \mathrm{RR}_{11})/((\mathrm{RR}_{21} - \mathrm{RR}_{11}) + (\mathrm{RR}_{12} - \mathrm{RR}_{11})$. The amount by which the product of the individual factors has to be multiplied by to obtain the joint effect is a multiplicative measure. In public health, epidemiologists attempt to predict a disease from a number of factors. Many argue that in this setting, these effects are best viewed on an additive scale. Here, the interaction would be defined to be the difference in joint effect differences. Based on simulations, they emphasize that the odds ratios should not be used to estimate the risk ratios, as they sometimes are when calculating the interaction contrast. This is the main point of the Kalilani and Atashhili (2006) paper. Zhang and Yu (1998) discuss the extent to which the odds ratio over-estimates the risk ratio when outcome proportions are not small.

4 Precision of Interaction Measures

Using an alternative parameterization, but assuming odds ratios (ORs) are approximately risk ratios (RRs), Hosmer and Lemeshow (1992) find that Var(ICR) and Var(AP) are obtainable. They propose using Wald type intervals, but caution that the coverage properties of these intervals have not been studied for the methods proposed. Somewhat more recently, Assmann, Hosmer, Lemeshow and Mundt (1996) use four methods of constructing confidence intervals in logistic regression, a delta method and

three bootstrap methods. They found that one of the bootstrap methods (using percentiles of the bootstrap distribution) had very good coverage properties. This work also assumed that ORs are approximately RRs.
Chu, et al. (2011) demonstrate the use of Markov Chain Monte Carlo (MCMC) in obtaining confidence intervals for RERI. For the following example, from data collected on individuals living with HIV in the US, we have obtained estimates for the interaction and probability intervals using MCMC. For comparison purposes, we have obtained bootstrap estimates using SAS as suggested by Assman et al. (1996).
Collapsing the data from our study into a 2×2 table with cell proportions $\{\{0.678, 0.672\}, \{0.606, 0.405\}\}$, our modeled proportions using MCMC were virtually identical, $\{\{0.679, 0.071\}, \{0.606, 0.405\}\}$. Defining the interaction contrast, LC, as above, the MCMC readily produced an estimate of -0.1927 and a 95% probability interval (-0.3242, -0.0446). We also defined the interaction measures RERI and AP, obtaining estimates and probability intervals. We then constructed bootstrap confidence intervals for RERI and AP using SAS and the approaches studied by Assman, et. al. (1996). These estimates that assume that odds ratios are approximately risk ratios differed considerably from those produced by MCMC.

5 Summary

We have explained the use and performance of various approaches in the context of the log-linear models for the HIV data and contrasted these with a Bayesian MCMC approach, which makes the calculation and comparison of alternative interaction summaries straightforward. Using this approach, results (estimates and associated uncertainty) from the model are an easily interpreted, but not misleading, summary for health-care professionals and policy advisers.

References

Aitkin, Murray, Francis, Brian, Hinde, John and Darnell, Ross (2009). *Statistical Modelling in R*. Oxford Statistical Science.

Assman, Susan F., Hosmer, David W., Lemeshow, Stanley and Mundt, Kenneth A. (1996). Confidence Intervals for Measures of Interaction. *Epidemiology*, **7**, (3), 286 – 290.

Chu, H., Nie, L. and Cole, S.R. (2001). Estimating the Relative Excess Risk Due to Interaction, A Bayesian Approach. *Epidemiology*, **22**, (2), 242 – 248.

Greenland, Sander ()1993. Risk versus Additive Relative Risk Models. *Epidemiology*, **4**, (1), 32 – 36.

Hosmer, David W. and Lemeshow, Stanley (1992). Confidence Interval Estimation of Interaction.
Epidemiology, **3**, (5), 452 – 456.

Hunter, Stuart J., Hunter, William G. and Box, George E. P. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery , 2nd Edition*. New York: John Wiley.

Kalilani, Linda and Atashili, Julius (2006). Measuring additive interaction using odds ratios. *Epidemiologic Perspectives & Innovations*, 3:5.

Knol, Mirjam J., VanderWeele, Tyler J., Groenwold, Rolf H., Klungel, et al. (2011). Estimating measures of interaction on an additive scale for preventive exposures. *Eur J Epidemiol.*, **26**, (6), 433 – 438.

Li, Hongyu and Barry, Jay (2012) Interpreting Interactions in Logistic Regression. *StatNews*, **84**, October. Cornell University: Cornell Statistical Consulting Unit.

Zhang, J. and Yu, K.F. (1998). What's the relative risk? *JAMA*, **280**, 1690 – 1691.

A latent variable approach to digit preference

Gillian Heller[1], Lindsay Dunlop[2]

[1] Department of Statistics, Macquarie University, Sydney, Australia
[2] Department of Haematology, Liverpool Hospital, Liverpool, Australia

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

Abstract: Digit preference occurs most commonly as a result of recall inaccuracy, but may also be due to behavioural preference. Inspection of usage of units of blood transfused after stem cell transplantation reveals a strong preference for even numbers, which is due to behavioural preference on the part of prescribing physicians. In the reporting of the age at which smokers quit smoking, strong bias towards round numbers is observed. We conceive of a latent variable which has a smooth distribution, which is transformed via stochastic rules to a discrete variable with probability spikes at preferred digits. We propose a modelling framework based on a latent variable specification and stochastic transformation to the spiked distribution. Specification of the stochastic rules is important to success in accurate modelling of the process.

Keywords: digit preference; heaping; latent variable.

1 Introduction

Response variables with frequency spikes at certain digits are commonly encountered. The best-known example is zero inflation, for which models are well developed. In our first motivating example, the outcome of interest is the number of packed red blood cell (PRBC) units transfused to stem cell transplant patients. A strong preference for even numbers is evident. In the second motivating example, survey respondents who are ex-smokers report the age at which they stopped smoking. A strong rounding preference is observed.

We distinguish between an underlying unobserved variable Z, which has a smooth distribution, and the observed variable N whose distribution is discrete with probability spikes at preferred digits. The mapping of Z to N is stochastic and governed by a model for the digit preference. We specify a model for N which assumes the latent distribution of Z, and the stochastic rules for $Z \Rightarrow N$. Maximum likelihood estimation is implemented for the marginal distribution of N .

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

2 Statistical model for N

We assume a latent variable Z and an observed variable N, where N is determined from Z by some stochastic rules. A smooth discrete or discretized continuous distribution is assumed for Z. We define the matrix C as having entries $C_{ij} = P(N = j \mid Z = i)$. If, for example, odd numbers are rounded up to the next number with probability π, then

$$C = \begin{pmatrix} 1-\pi & \pi & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & \\ 0 & 0 & 1-\pi & \pi & \\ 0 & 0 & 0 & 1 & \\ \vdots & & & & \end{pmatrix}$$

The distribution of N is achieved by the redistribution of probabilities:

$$P_N(n) = \sum_{k=0}^{\infty} C_{kn} P_Z(k) \qquad \text{or} \qquad \mathbf{P_N} = C^{\mathrm{T}}\, \mathbf{P_Z} \ .$$

In order to fully define the distribution of N, it remains to specify the latent distribution $P_Z(z)$. The distribution of N is fully specified, with parameters $(\pi, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters of the distribution of Z. While the likelihood equations may be tractable in particular cases of $P_Z(z)$, in general they are solved numerically.

3 The blood transfusion data

Heller and Dunlop (2012) report on a study of $n = 166$ stem cell transplant patients at a Sydney hospital who received an autologous peripheral blood stem cell transplant. The haematologist made the decision to prescribe two or more units of PRBC per transfusion event, dependent on the patient's clinical condition. The number of units transfused was summed over the patient's entire admission. The left panel of Figure 1 shows the histogram of the total number of units transfused. The following features are apparent: the minimum number of units transfused in those patients transfused is 2; and there is a strong preference for transfusing an even number of units.

4 The smoking data

In the 2001 National Drug Strategy Household Survey conducted by the Australian Institute of Health and Welfare (AIHW 2002), ex-smokers were asked at what age they stopped smoking. People aged $\geq$65 years were analysed ($n = 1,205$). The right panel of Figure 1 shows the histogram of the reported age of quitting smoking. A strong preference towards rounding to 10s, and to a lesser extent to 5s, is observed.

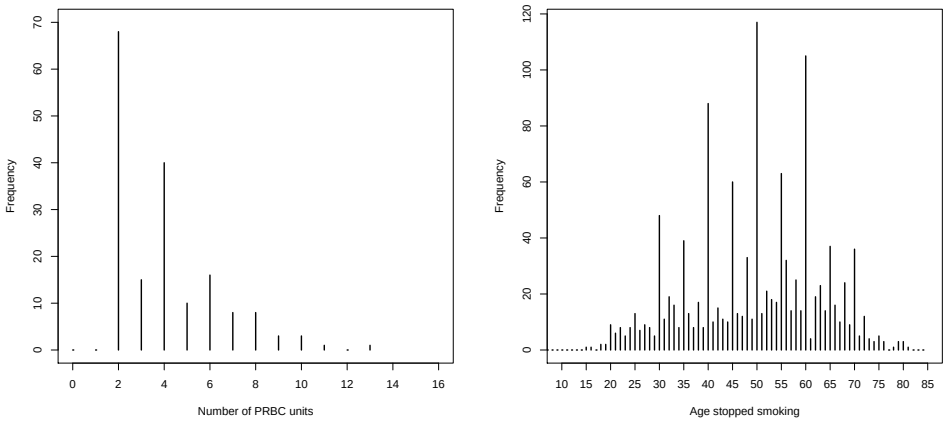


FIGURE 1. Left panel: histogram of number of PRBC units transfused; right panel: histogram of reported age of quitting smoking

5 Statistical model for blood transfusion

We make a distinction between the amount of blood the patient needs and the amount he/she receives. Z is the unobserved number of units needed and the relationship between Z and N, the number of units transfused, is modelled as

1. if one or two units are needed, two units are transfused;
2. if an odd number of units are needed, one more unit than is needed is transfused, with probability π.

We have

$$P_N(n) = \begin{cases} P_Z(1) + P_Z(2) & \text{for } n = 2 \\ \pi P_Z(n-1) + P_Z(n) & \text{for } n = 4, 6, \ldots \\ (1-\pi) P_Z(n) & \text{for } n = 3, 5, \ldots \\ 0 & \text{otherwise} \end{cases}$$

and

$$C = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 1-\pi & \pi & 0 & 0 & \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 1-\pi & \pi & \\ \vdots & & & & & & & \end{pmatrix}$$

For the latent distribution of Z we specify the discretized Gamma distribution parametrized as:

$$P_Z(z) = \int_{z-1}^{z} \frac{w^{-1}}{\Gamma(\nu)} \left(\frac{w\nu}{\mu}\right)^{\nu} e^{-w\nu/\mu}\, dw\ , \qquad z = 1, 2, \ldots\ .$$

Maximum likelihood estimation was performed numerically, using the R function `nlm` for optimization. MLEs are $\hat{\mu} = 3.12$, $\hat{\nu} = 1.71$, $\hat{\pi} = 0.37$. A strong tendency to round up is indicated. In a comparison with the zero truncated Poisson (ZTP) and zero truncated negative binomial (ZTNB) distributions, the estimates of the mean of the latent distribution in the latter two models are positively biased. In addition, standard errors for the mean based on the ZTP and ZTNB are an order of magnitude greater than that of the proposed model. The fitted distribution of the number of units N is shown in Figure 2, as well as the fitted latent Gamma distribution of the volume of blood. The fitted distribution of N replicates the observed pattern well, supporting the choice of the discretized Gamma distribution for Z.

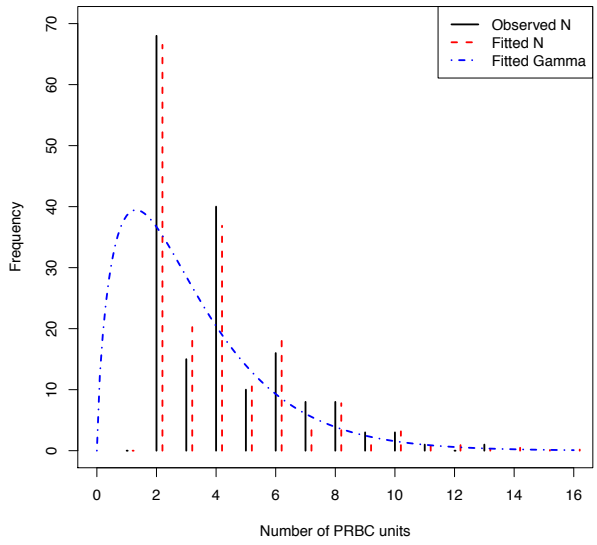


FIGURE 2. Observed and fitted distribution of PRBC units transfused (N); and underlying Gamma distribution of the latent volume of blood.

6 Statistical model for age of quitting smoking

Z is the unobserved age (in years) at which the subject quit smoking and N is the reported age. The relationship between Z and N is modelled as

1. if Z ends in 7, 8, 9, 1, 2 or 3, round to 0 with probability π_1;
2. if Z ends in 4 or 6, round to 5 with probability π_2.

The C matrix has entries $C_{ij} = \mathrm{P}(\text{age } j \text{ reported as } i)$:

$$
\begin{array}{c|ccccccccccccccccccccc}
 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 & \cdots & 85 \\
15 & 1 & \pi_2 & & & & & & & & & & & \\
16 & & 1-\pi_2 & & & & & & & & & & & \\
17 & & & 1-\pi_1 & & & & & & & & & & \\
18 & & & & 1-\pi_1 & & & & & & & & & \\
19 & & & & & 1-\pi_1 & & & & & & & & \\
20 & & & \pi_1 & \pi_1 & \pi_1 & 1 & \pi_1 & \pi_1 & \pi_1 & & & & \\
21 & & & & & & & 1-\pi_1 & & & & & & \\
22 & & & & & & & & 1-\pi_1 & & & & & \\
23 & & & & & & & & & 1-\pi_1 & & & & \\
24 & & & & & & & & & & 1-\pi_2 & & & \\
25 & & & & & & & & & & \pi_2 & 1 & & \\
\vdots & & & & & & & & & & & & &
\end{array}
$$

Specifying a discretized Weibull distribution for Z results in the fitted distribution shown in Figure 3.

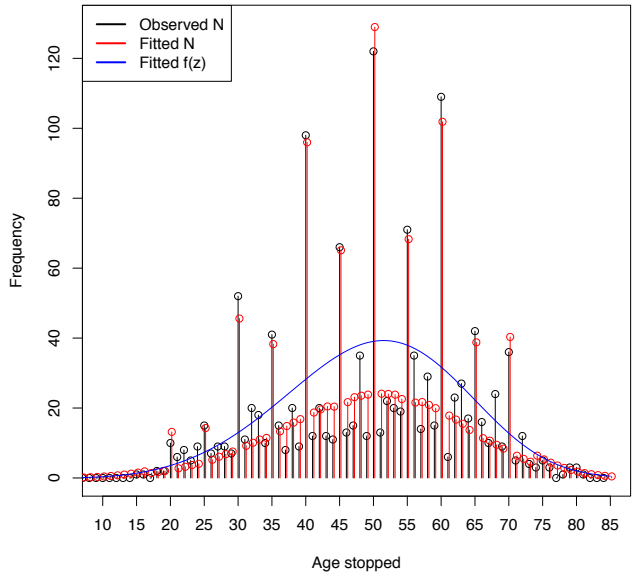


FIGURE 3. Observed and fitted distribution of reported age (N); and underlying Weibull distribution of age.

7 Discussion

Wang and Heitjan (2008) propose a Bayesian model for heaped cigarette count data in a situation more complex than our examples, where digit preference is either for a multiple of 5 or 10, or the size of a cigarette pack (20), and the heaping is partially due to misreporting and partially due to people tending to smoke whole numbers of cigarette packs. Camarda et al (2008) and Eilers and Borgdorff (2004) adopt a penalized likelihood approach to model distributions observed with digit preference.

We have shown that the proposed model accurately reproduces the observed variation in both data sets. Observed digit preference may be due to misreporting, or genuine behavioural preference. In the PRBC data, there was no misreporting and the digit preference observed was due to the tendency for physicians to prescribe at least two units at a time. In the survey data, digit preference was due to recall bias. Whatever the reason for the digit preference, it is important to reflect this accurately in the specification of the C matrix. It is a fairly simple matter to implement more intricate patterns of digit preference using this method.

References

Eilers P.H.C. and Borgdorff, M.W. (2004). Modeling and correction of digit preference in tuberculin surveys, *International Journal of Tuberculosis and Lung Disease*, 8(2):232 – 239.

Camarda C.G., Eilers P.H.C. and Gampe J. (2008). Modelling general patterns of digit preference, *Statistical Modelling*, 8(4):385401.

Australian Institute of Health and Welfare (2002). *2001 National Drug Strategy Household Survey: first results.* Drug statistics series no. 9. Cat. no. PHE 35. Canberra: AIHW.

Heller G.Z. and Dunlop, L.C. (2012). A modelling approach for blood units transfused after stem cell transplantation, *Statistics in Medicine*, **31** (28): 3649 – 3655.

A latent variable approach to derive consensus GDPs

Jesus Crespo–Cuaresma[1], Martin Feldkircher[2], Bettina Grün[3], Paul Hofmarcher [3], Stefan Humer [1]

[1] WU Vienna, Department of Economics, Austria
[2] Oesterreichische Nationalbank, Austria
[3] JKU Linz, Department of Applied Statistics, Austria

E-mail for correspondence: paul.hofmarcher@jku.at

Abstract: The Penn World Tables (PWT) are the most prominent source for international comparable Gross Domestic Product (GDP) data and are frequently updated. Recently, several authors have pointed out the sensitivity of econometric results to revisions of these data (e.g., Johnson et al. 2013). In this paper we propose a model framework based on a latent variable specification to derive consensus GDP series based on a large set of the six most recent PWT vintages. This approach allows us to take into account the variability associated with the different PWT revisions in a formal and consistent way and to quantify the corresponding uncertainty of real GDP figures.

Keywords: Gross Domestic Product (GDP); consensus; Bayesian estimation.

1 Motivation

International income data from the Penn World Tables (PWT) data base is widely used among economists. The PWT provide purchasing power parity (PPP) adjusted GDP (and GDP components) figures which are essential for conducting cross-country comparisons.
Repeated revisions due to updates of national income data (Base Year), new International Comparison Program (ICP) price data and changes in the underlying methodology to estimate purchasing power parities (PPPs) result in substantial alterations of PWT's real GDP figures. Moreover, Ciccone & Jarocinski (2010) and Johnson et al. (2013) show that studies on the determinants of economic growth, e.g. Sala-i-Martin et al. (2004) and Fernandez et al. (2001), are sensitive to the used PWT vintage.
In this paper we propose a Bayesian model framework based on a latent variable specification (see e.g., Grün et al., 2013) to derive a consensus

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

GDP (henceforth CGDP) series based on the heterogenous information of the different PWT vintages. For the latent CGDPs our framework assumes both a inter-temporal correlation within the single country's GDP series as well as a correllation between the countries. The former is achieved via latent AR-processes while for the latter correlation we make use of Bayesian spline regressions (see e.g. Lang and Brezger, 2004). In addition, based on the estimated CGDPs, our framework allows for validating the single PWT vintages by analyzing the mean/variance structure of the consensus deviations for the single PWT vintages.

1.1 The latent consensus GDP model

For PWT version j, we denote the observed GDP of country i at time t, with $Y_{ij}(t)$. Due to different input data (ICP, Base Year) and heterogenous modelling approaches, $Y_{ij}(t)$ might vary for fixed i, t between the PWT versions j. Therefore we assume a latent "true" CGDP, $Y_i^*(t)$, estimated from the observed $Y_{ij}(t)$. That is,

$$Y_{ij}(t) = Y_i^*(t) + \epsilon_{ij}(t), \tag{1}$$

with $\epsilon_{ij}(t)$ denoting the error of PWT version j for country i at time t. Following equation 1 probabilistic processes for both the consensus $Y_i^*(t)$ and the error structure $\epsilon_{ij}(t)$ have to be specified. $Y_i^*(\cdot)$, our CGDP estimates can be interpreted as a *more informative* GDP estimate, as it encapsulates the information contained in the various observable, but heterogenous PWT vintages $Y_{ij}(t)$.

To model CGDP, we assume that each $Y_i^*(t)$ follows some country's idiosyncratic development $\nu_i(t)$, (e.g., wars, government actions) plus some macroeconomic factors f, which induce a correlation between the countries' CGDPs. Several approaches might be appropriate for modelling $\nu_i(\cdot)$ and $f(\cdot)$. Here we assume latent $AR(1)$ processes for $\nu_i(t)$ and the more flexible class of cubic penalized spline regressions (P-splines) to handle correlations between the $Y_{ij}(t)$, i.e. to model $f(t)$. The dependency of each single i from the global market f is denoted via α_i. Formally we have

$$Y_i^*(t) = \nu_i(t) + \alpha_i f(t), \tag{2}$$

with $\nu_i(t) = \gamma_i \nu_i(t-1) + \epsilon_i(t)$ and $f(t) = \sum_m \xi_m B_m(t)$. $B_m(t)$ denotes the m-th spline basis function. Following e.g., Fahrmeir et al. (2010) conventional Bayesian P-spline smoothing is based on random walk priors for the (first) differences of ξ, i.e., $\xi_m = \xi_{m-1} + u_m$, $u_m \sim N(0, \tau^2)$.

Next to the CGDP $Y_i^*(t)$ we have to discuss admissible models for the errors $\epsilon_{ij}(t)$. Here we assume that each PWT vintage has its own characteristic error $\mu_j(t)$, which is independent of country i, i.e.,

$$\epsilon_{ij}(t) = \mu_j(t) + \sigma_j \epsilon_{ij}(t), \tag{3}$$

with $\sigma_j \sim N(0, \tau_\sigma)$, $\epsilon_{ij}(t) \sim N(0,1) \quad \forall t$ and $\mu_j(t)$ denoting the PWT vintage specific errors. Again, the time structure of the error function $\mu_j(t)$

is captured via standard AR(1) processes and to guarantee identifiability of our model $\sum_j \mu_j(t) = 0, \ \forall t$ is required.

2 Results

We are now ready to present the most prominent results of our CGDP model framework. We apply our model framework to demeaned GDP data from the Penn World Table Version 6.1 to 8.0. Table 1 summarizes the main characteristics of these versions but also displays the estimated averaged mean errors of the different PWT vintages.

TABLE 1. Properties of the different PWT vintages. The last row displays the mean values (time averaged) of the PWT deviations $\mu_j(t)$ from the estimated consensus

PWT	6.1	6.2	6.3	7.0	7.1	8.0
From	1950	1950	1950	1950	1950	1950
To	2000	2004	2007	2009	2010	2011
Base Year	1996	2000	2005	2005	2005	2005
Countries	168	188	188	189	189	167
ICP	1996	*2002*	*2002*	2005	2005	1970–2005
$\mu_j(\cdot)$	592.80	−206.35	−227.78	198.80	167.85	−403.09

Based on the estimated CGPD values the largest deviations are observed for PWT version 6.1 followed by the latest release PWT 8.0.
Additionally we can inspect the estimated CGDP as well as the CGDP compositions in terms of $\alpha_i f(t)$ and the idiosyncratic developments $\nu_i(t)$. Figure 1 illustrates these decompositions for the USA and Botswana. While for the US we find the global market as a driver for the GDP pattern and the idiosyncratic process is mainly a shift of the GDP values, the picture looks vica-versa for Botswana. Here we find that the CGDP is mainly driven by the idiosyncratic $\nu_i(t)$ process.

Acknowledgments: Paul Hofmarcher's research is supported by the Oesterreichische Nationalbank under the Jubiläumsfond grant 14663.

References

Ciccone, C., and Jarocinski, M. (2010). Determinants of economic growth: will data tell? *American Economic Journal: Macroeconomics*, **2**, 222 – 246.

Fahrmeir, L., Kneib, T., and Konrath, S. (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, **20**, 203 – 219

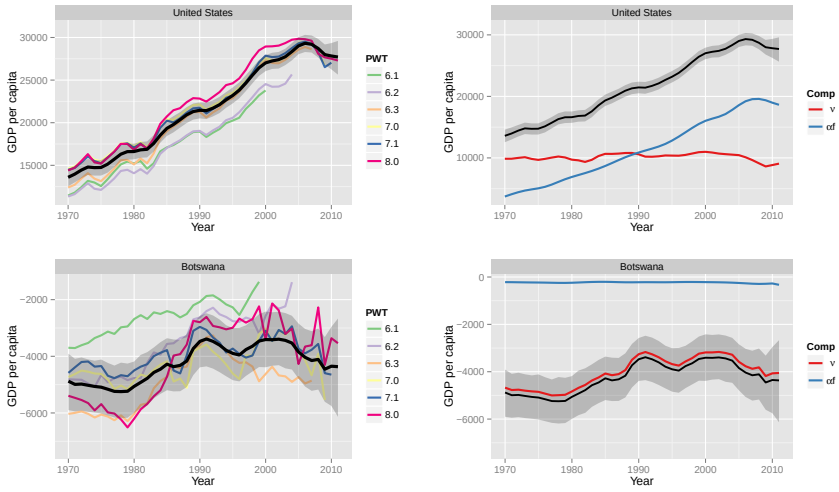


FIGURE 1. Demeaned GDP estimates of the different PWT vintages and the estimated consensus GDP (black) for the USA (top) and Botswana (bottom). On the right we find the discussed CGDP decompositions.

Fernandez, C., Ley, E. and Steel, M.F.J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, **16**, 563 – 576.

Grün, B., Hofmarcher, P., Hornik, K., Leitner, C., and Pichler, S. (2013). Deriving consensus ratings of the big three rating agencies. *Journal of Credit Risk*, **9**, 75 – 98.

Johnson, S., Larson, W., Papageorgiou, C., and Subramanian, A. (2013). Is newer better? Penn World Table Revisions and their impact on growth estimates. *Journal of Monetary Economics*, **60**, 255 – 274.

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

Sala-i-Martin, X., Doppelhofer, G. and Miller, R.I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review*, **94**, 813 – 835.

Bivariate Gaussian Distributional Regression: An Application on Diabetes

Nadja Klein,[1] Francisco Gude,[2] Carmen Cadarso-Suárez,[3], Thomas Kneib[1]

[1] Georg-August University Goettingen, Germany
[2] Hospital Clínico Universitario de Santiago, Spain
[3] Universidad de Santiago de Compostela, Spain

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

Abstract: So far, different markers reflecting the blood sugar levels are analysed separately and used for the diagnosis and control of diabetes. It is known that such markers (glycated proteins) have a high correlation which is expected to be dependent on clinical or biochemical factors. Since neglecting correlations between two proteins or heteroscedasticity of the distribution may lead to misclassification of patients, we aim in identifying important covariates on the variance and correlation parameters with the help of structured additive distributional regression for multivariate responses.

Keywords: correlated responses; penalised splines; glycated proteins.

1 Introduction

Diabetes, a common life-long disease, is characterised by chronic hyperglycemia (high blood sugar) and causes long-term complications like retinopathy (damage of the retina of the eye), nephropathy (kidney disease), or neuropathy (diseases affecting nerves). According to the World Health Organization, it is expected that the number of patients with diabetes will rise to 370 million in the world by 2030 (Wild et al., 2004).
The determination of the glycated protein hemoglobin (*a1c*) has been proposed as a criterion for the diagnosis of diabetes. However, the correlation between blood sugar level and *a1c* is not perfect. In addition to hemoglobin, the serum protein fructosamine (*fru*) in the plasma can also become glycated and can therefore also be used as a marker of blood sugar levels with the difference that *fru* provides an index of glucose control over a period of 2–3 weeks compared to a period of 3 months for *a1c*. Estimating independent models for *a1c* and *fru* may lead to misspecification and thus to the

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

wrong diagnosis. Furthermore, since one of them represents the glycation process inside the red blood cells (RBC) (intracellular) and the other one in plasma (extracellular), any factors that modify RBC survival (such as anemia) will also modify *a1c* test results relative to the true level of blood sugar and thus also its relationship with the other glycated proteins.
So far, most statistical approaches in analysing epidemiological and clinical studies focused on linear predictors following the classical framework of generalised linear models. In all these models, only the mean of a dependent variable is related to covariates, neglecting the potential dependence of higher moments or the correlation between *a1c* and *fru* on covariates. Assuming that the conditional distribution of *a1c* and *fru* is bivariate Gaussian, the framework of Bayesian structured additive distributional regression (Klein et al., 2013a) allows to consider such dependence structures. Their model class is based on the framework of generalised additive models for location, scale and, shape (GAMLSS, Rigby and Stasinopoulos, 2005) where potentially complex parametric distributions can be assumed for response variables. However, since the framework of GAMLSS is currently restricted to univariate responses, we rely on an extended approach for multivariate responses, recently developed by Klein et al. (2013b) in the spirit of distributional regression.
With the purpose of assessing the impact of clinical and biochemical factors on both the mean and the variability of the glycated proteins *a1c* (in %) and *fru* (in μmol/l), we analysed data from 542 adults recruited until December 2013 within a study performed in the municipality of A Estrada in north-western Spain. Available clinical factors are: *age* of the adult in years, hepatic disease (*hep*, 0=no, 1=yes) and kidney disease (*renal*, 0=no, 1=yes). Biochemical ones are: glucose (*glu*, fasting glucose concentrations in mg/dl), serum albumin (*alb*, in g/l) and mean corpuscular volume (*mcv*, f/l in red cells). In addition, a specific objective of the study is to analyse the relationship between both responses according to the levels of *mcv* after adjusting by other potential confounding factors including glucose levels.

2 Bivariate Gaussian Distributional Regression

We assume that the joint conditional distribution of $(a1c_i, fru_i)'$ given all covariates summarised in $\boldsymbol{\nu}_i' = (\mathbf{x}_i, \mathbf{z}_i)'$, $i = 1, \ldots, 542$, follows a bivariate Gaussian distribution. We write $(a1c, fru)' \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and recall that the joint density is characterised by the expectation $\boldsymbol{\mu} = (\mathbb{E}(a1c), \mathbb{E}(fru))' \in \mathbb{R}^2$ and the positive definite covariance matrix with parameters $\sigma_1^2 = \mathrm{Var}(a1c)$, $\sigma_2^2 = \mathrm{Var}(fru)$, and $\rho = \mathrm{Cor}(a1c, fru)$. Following Klein et al. (2013b) we do not only explain the expectations μ_1, μ_2 of the marginal distributions of *a1c* and *fru* as functions of covariates as it is for instance done in seemingly unrelated regression, but also allow the standard errors and the correlation parameters to depend on covariates. Furthermore, we allow for a flexible modelling of continuous covariates to overcome the restrictions of linear predictors. This means that each parameter of the distribution is linked to

an additive predictor structure (Fahrmeir, Kneib, and Lang 2004) formed of covariates as explained below. To ensure possible parameter space restrictions, we choose appropriate link functions and end up with the five predictor equations

$$\eta_i^{\mu_1} = \mu_{1,i}, \quad \eta_i^{\mu_2} = \mu_{2,i}, \quad \eta_i^{\sigma_1} = \log(\sigma_{1,i}), \quad \eta_i^{\sigma_2} = \log(\sigma_{2,i}), \quad \eta_i^{\rho} = \frac{\rho_i}{\sqrt{1-\rho_i^2}}$$

where the corresponding parameter is denoted in the exponent of the predictor. Effects have been selected based on the Deviance Information Criterion (DIC) in combination with expert knowledge from previous studies:

$$\eta_i^{\mu_d} = \beta_0^{\mu_d} + renal_i \beta_1^{\mu_d} + hep_i \beta_2^{\mu_d} + f_1^{\mu_d}(age_i) + f_2^{\mu_d}(glu_i) + f_3^{\mu_d}(alb_i) + f_4^{\mu_d}(mcv_i)$$

$$\eta_i^{\sigma_k} = \beta_0^{\sigma_k} + f_1^{\sigma_k}(glu_i), \quad \eta_i^{\rho} = \beta_0^{\rho} + f_1^{\rho}(mcv_i),$$

$d = 1, 2$. Dropping the parameter index for simplicity, the predictors are an additive composition of an intercept β_0 representing the overall level of the predictor, linear effects x_i (*renal*, *hep*), and functions $f_j(z_{ij})$ reflecting nonlinear effects of continuous covariates z_{ij} (*age*, *glu*, *alb*, *mvc*) with penalised spline approaches. Our justification for choosing the bivariate Gaussian distribution is based on normalised quantile residuals. Estimation is performed with Bayesian inference relying on Markov chain Monte Carlo simulation techniques and iteratively weighted least squares proposals. Further details about the approach are given in Klein et al. (2013b).

3 Results

The effects of covariates on expectations, standard errors and correlation between the responses, can be visually inspected and their functional form identified in the graphs shown in Figure 1 to 4. *(1) Marginal expectations, Figure 1 and 2.* Mean *a1c* concentrations increase almost linearly with *age* while *fru* concentrations only do so for elderly people (> 60 years). Glucose (*glu*) is the main covariate on predicting both *a1c* and *fru* concentrations and the functional form of the effect of glucose levels on both proteins is similar. *mcv* concentrations are slightly and inversely associated with *a1c* concentrations. Albumin (*alb*) levels seem not to be associated with *a1c* concentrations while the effect of albumin levels on *fru* is marked, as expected.
(2) Marginal standard deviations, Figure 3. Variabilities of *a1c* and *fru* are higher at lower and higher glucose levels. A larger variability is expected in both glycated proteins at higher levels of glucose indicating people with diabetes. The high variability in the glycated proteins at lower levels of glucose could also indicate the presence of people with diabetes being treated with anti-diabetic drugs and thus presenting low glucose levels when measured.
(3) Correlation, Figure 4. The correlation between both proteins increases with increasing levels of *mcv*. Starting from a slightly negative correlation we estimate a significantly positive correlation between *a1c* and *fru* for *mcv*

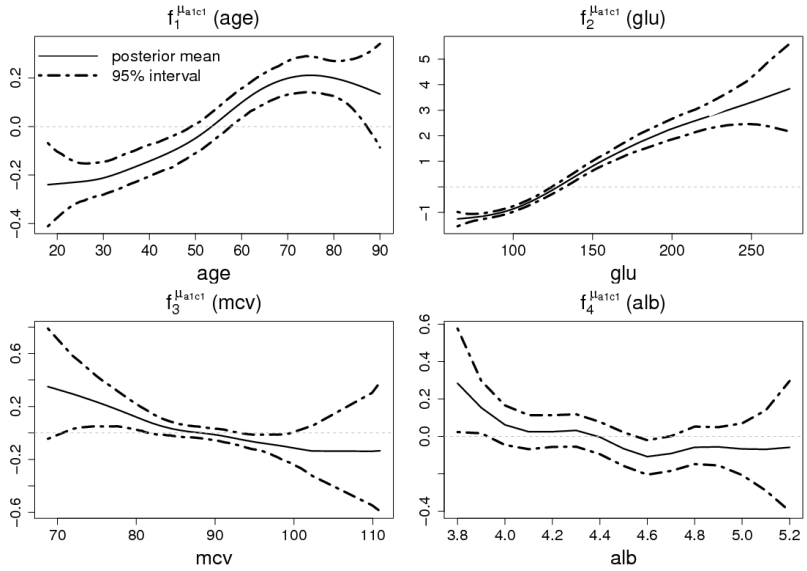


FIGURE 1. Estimated posterior mean effects of *age* (top-left), *glu* (top-right), *mcv* (bottom-left) and *alb* (bottom-right) on μ_{a1c} together with 95% credible interval. Effects are centred around zero.

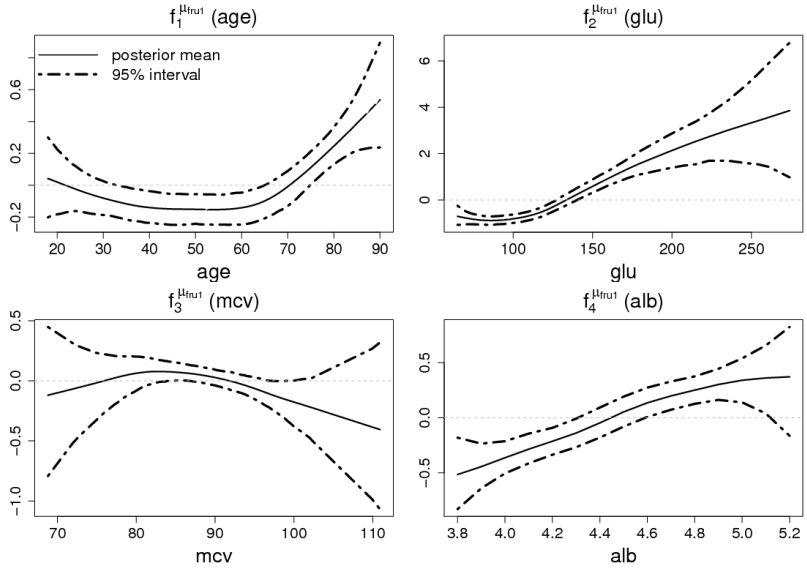


FIGURE 2. Estimated posterior mean effects of *age* (top-left), *glu* (top-right), *mcv* (bottom-left) and *alb* (bottom-right) on μ_{fru} together with 95% credible interval. Effects are centred around zero.

in the range of 85-100 f/l. This supports the hypothesis that the glycation processes are different if these proteins are located inside or outside the red cells. Thus, clinicians should take into account levels of *mcv* when control

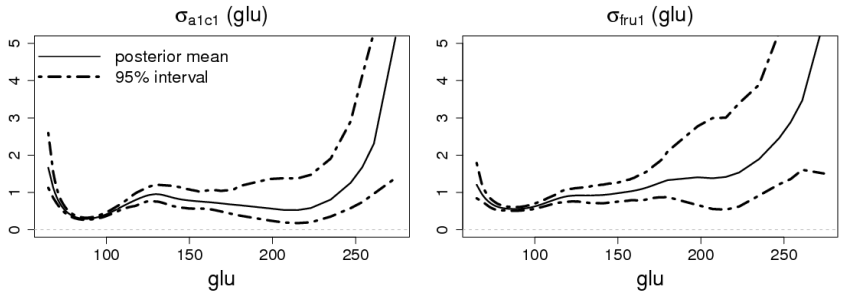


FIGURE 3. Estimated marginal standard errors of σ_{a1c} (left) and σ_{fru} (right) as a function of *glu*. Shown are posterior means together with 95% credible interval.

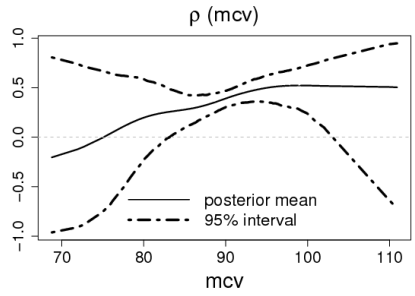


FIGURE 4. Estimated correlation between *a1c* and *fru* as a function of *mcv*. Shown are posterior means together with 95% credible interval.

and diagnosis of diabetes is based on *a1c* values.

4 Discussion

In the present study we provide evidence that besides blood glucose other conditions such as age and albumin concentrations are variables related to *a1c* and *fru*. Furthermore, we find that different mean corpuscular volumes induce different correlations between *a1c* and *fru*. Since concentrations of *a1c* are used for diagnosing diabetes these finding could have clinical implications: *a1c* of patients with lower levels of *mcv* might be overestimated while underestimated for patients with higher levels of *mcv* if the correlations are neglegted.

In future research it will be interesting to incorporate the functional form of glucose levels with repeated measurements over the day and to further relax the distributional assumption or to compare the bivariate normal distribution with other bivariate distributions based on copulas.

Acknowledgments: The work was supported by grants from the Carlos III Health Institute, Spain (RD12/0005/0007, PI11/02219), the Spanish Ministry of Science and Technology (MTM 2011-28285-C02-01) and the

German Academic Exchange Service (DAAD). The work of Nadja Klein and Thomas Kneib was supported by the German Research Foundation (DFG) via the research training group 1644 and the research project KN 922/4-1. The work of Carmen Cadarso-Suárez and Thomas Kneib was supported by the Spanish Ministry of Technology and Innovation Grant.

References

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 731–761.

Klein, N., Kneib, T., and Lang, S. (2013a). Bayesian Structured Additive Distributional Regression. *Tech. Report, http://econpapers.repec.org/paper/innwpaper/2013-23.htm.*

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2013b). Bayesian Structured Additive Distributional Regression for Multivariate Responses. *Tech. Report, http://econpapers.repec.org/paper/innwpaper/2013-35.htm.*

Wild, S., Roglic, G., Green, A., Sicree, R., and King, H. (2004) Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* **27**, 1047–1053.

Regression modelling of misclassified correlated interval-censored data

Arnošt Komárek[1], María José García-Zattera[2], Alejandro Jara[3]

[1] Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
[2] Department of Statistics and Measurement Center MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile
[3] Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile

E-mail for correspondence: arnost.komarek@mff.cuni.cz

Abstract: We propose a flexible modelling approach for correlated (clustered) time-to-event data, when the response of interest can only be determined to lie in an interval obtained from a sequence of examination times (interval-censored data) and, on top of that, the determination of the occurrence of the event is subject to misclassification.

Keywords: Accelerated failure time model; multivariate survival data; misclassification.

1 Introduction

Based on dental data gathered in a longitudinal oral health study, the Signal Tandmobiel® (ST) study (Vanobbergen et al., 2000), we aim at assessing the effect of predictors on the time to caries experience (CE) in the permanent dentition. This motivates our research consisting of developing a regression model for misclassified correlated (clustered) interval-censored data since: (i) events on teeth of the same child are dependent, (ii) due to the setup of the study, the tooth status was assessed yearly and, thus, the time to CE can only be determined to lie in a given interval of time, and (iii) several examiners were involved in the study and their caries classification may not perfectly reflect the tooth's true condition and, therefore, the presence/absence of CE can be misdiagnosed.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

2 Misclassified interval-censored data

Let $T_{(i,j)} \in \mathbb{R}_+$ be the time-to-event (time to CE) for the jth unit (tooth) of the ith subject (child), $i = 1, \ldots, N$, $j = 1, \ldots, J$. Suppose that the occurrence of the event is assessed by using a sequence of subject-specific evaluations. Let $0 < v_{(i,1)} < v_{(i,2)} < \cdots < v_{(i,K_i)} < +\infty$ be the ordered examination times for the ith subject, $i = 1, \ldots, N$. In a regular interval-censored data context, the time-to-event $T_{(i,j)}$ is unobserved but is exactly known to lie in an interval $T_{(i,j)} \in \left(v_{(i,l_{(i,j)}-1)}, v_{(i,l_{(i,j)})}\right]$ obtained from the sequence of examinations, $l_{(i,j)} \in \left\{1, \ldots, K_i + 1\right\}$, where $v_{(i,0)} \equiv 0$ and $v_{(i,K_i+1)} \equiv +\infty$. Such data are obtained if it can be assumed that at each examination, the occurrence of the event is diagnosed by a perfect procedure with a zero probability of both false positive and false negative findings. The follow-up of the (i,j)th unit is then typically terminated at time $v_{(i,l_{(i,j)})}$ when the event was diagnosed.

In the context of the ST study, or any other study where the occurrence of the event is determined by a diagnostic test or examination with imperfect sensitivity and/or specificity, however, all units are examined for the occurrence of the event at all examination times. Therefore, the observed data for the (i,j)th unit is a vector of binary variables $\mathbf{Y}_{(i,j)} = \left(Y_{(i,j,1)}, \ldots, Y_{(i,j,K_i)}\right)^\top$, where $Y_{(i,j,k)}$ indicates (subject to a potential classification error) whether the event was diagnosed at examination at time $v_{(i,k)}$ for having experienced the event ($Y_{(i,j,k)} = 1$) or not ($Y_{(i,j,k)} = 0$), $k = 1, \ldots, K_i$. With perfect classification, the sequence $\mathbf{Y}_{(i,j)}$ would always be monotone in case of a classical non-reversal survival event. Nevertheless, this is not necessarily true in our context due to possible misclassification of the event status. In the following, set $\mathbf{Y}_i = \left(\mathbf{Y}_{(i,1)}^\top, \ldots, \mathbf{Y}_{(i,J)}^\top\right)^\top$, $i = 1, \ldots, N$, which can be characterized as misclassified interval-censored data for the ith subject.

Finally, we shall assume that for each subject and unit, a p-dimensional design vector including endogenous covariates is recorded, $\mathbf{x}_{(i,j)}$, $i = 1, \ldots, N$, $j = 1, \ldots, J$. The main aim here is to develop a regression model for the event times $T_{(i,j)}$ as a function of covariates $\mathbf{x}_{(i,j)}$, where the event times $T_{(i,j)}$ are observed only through vectors of possibly misclassified binary indicators $\mathbf{Y}_{(i,j)}$ of the event status. Finally, let $\mathbf{T}_i = \left(T_{(i,1)}, \ldots, T_{(i,J)}\right)^\top$, $i = 1, \ldots, N$, be the event times of all units of the ith subject. When developing the model, it should also be taken into account that the elements of the random vector $\mathbf{T}_i$ are not necessarily independent due to the clustering.

3 The misclassification model

With respect to the evaluation of the event status, we shall assume that its classification is performed at each examination by one of Q different examiners (Q different diagnostic tests). We denote by $\xi_{(i,k)} \in \left\{1, \ldots, Q\right\}$ the index of the examiner who performed classification of all units of the ith

subject at its kth examination at time $v_{(i,k)}$. Values of $\xi_{(i,k)}$, $i = 1, \ldots, N$, $k = 1, \ldots, K_i$, act as covariates of the misclassification model.
Following García-Zattera et al. (2012), we allow for the fact that different examiners may have different misclassification rates and that they can also vary across different units. In the context of the dental study, this allows, e.g., for the fact that classification of caries on teeth deeper in the mouth is more difficult than on teeth closer to the front. Let $\boldsymbol{\eta}_q = \big(\eta_{(q,1)}, \ldots, \eta_{(q,J)}\big)^\top$, and $\boldsymbol{\alpha}_q = \big(\alpha_{(q,1)}, \ldots, \alpha_{(q,J)}\big)^\top$, $q = 1, \ldots, Q$, be the vectors containing the unit-specific unknown specificities and sensitivities of the qth examiner, respectively. That is,

$$\eta_{(q,j)} = \mathsf{P}\big(Y_{(i,j,k)} = 0 \,\big|\, T_{(i,j)} > v_{(i,k)};\, \xi_{(i,k)} = q\big),$$
$$\alpha_{(q,j)} = \mathsf{P}\big(Y_{(i,j,k)} = 1 \,\big|\, T_{(i,j)} \leq v_{(i,k)};\, \xi_{(i,k)} = q\big),$$

$i = 1, \ldots, N$, $j = 1, \ldots, J$, $k = 1, \ldots, K_i$, $q = 1, \ldots, Q$. All specificities, $\boldsymbol{\eta} = \big(\boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_Q^\top\big)^\top$, and all sensitivities, $\boldsymbol{\alpha} = \big(\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_Q^\top\big)^\top$, are then the unknown parameters of the misclassification model. A simplified version of the misclassification model assuming, e.g., $\eta_{(q,1)} = \cdots = \eta_{(q,J)}$ and $\alpha_{(q,1)} = \cdots = \alpha_{(q,J)}$, for every $q = 1, \ldots, Q$, was also considered.

4 The event times model

A possible way to regress the event times on the covariates, while taking into account dependence stemming from clustering, is a random-intercept accelerated failure time (AFT) model

$$\log\big(T_{(i,j)}\big) \;=\; \mathbf{x}_{(i,j)}^\top \boldsymbol{\beta} \;+\; b_i \;+\; \varepsilon_{(i,j)}, \qquad i = 1, \ldots, N,\; j = 1, \ldots, J,$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown regression coefficients, $b_1, \ldots, b_N$ are i.i.d. random variables with density g_b, and $\varepsilon_{(1,1)}, \ldots, \varepsilon_{(N,J)}$ are i.i.d. random variables with density g_ε, independent of $b_1, \ldots, b_N$. Let $\varphi(\cdot;\, \mu,\, \sigma^2)$ be the density of an $\mathcal{N}(\mu,\, \sigma^2)$ distribution. We shall assume that $g_\varepsilon(\cdot) = \varphi(\cdot;\, 0,\, \sigma_\varepsilon^2)$, where σ_ε^2 is an unknown error variance. To make the model robust against the misspecification of the distribution of the event times $T_{(i,j)}$ implied by assumed forms of g_ε and g_b, we specify the random-intercept density g_b in a flexible way. To this end we exploit a penalized Gaussian mixture (see, e.g., Komárek and Lesaffre, 2008), that is,

$$g_b(\cdot) = \frac{1}{\tau} \sum_{l=-M}^{M} w_l\, \varphi\Big(\frac{\cdot \,-\, \mu}{\tau};\; \kappa_l,\, \zeta^2\Big), \tag{1}$$

where $\kappa_{-M}, \ldots, \kappa_M$ is a fixed and fine grid of equidistant knots centered at $\kappa_0 = 0$, ζ^2 is a known basis variance, $\mathbf{w} = \big(w_{-M}, \ldots, w_M\big)^\top$ is a vector of unknown positive weights that sum up to one, and μ and $\tau > 0$ are unknown location and scale parameter, respectively. Following the recommendations provided by Komárek and Lesaffre (2008), we took $M = 15$, leading to

$2M + 1 = 31$ mixture components, $\kappa_{-M} = -4.5$, $\kappa_M = 4.5$, and $\zeta^2 = 0.2^2$. Note that regularization of estimate of the density (1) is achieved by using a penalty for the mixture weights $\mathbf{w}$ in a mood of penalized smoothing. In summary, the unknown parameters of the event times model are $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \mathbf{w}^\top, \mu, \tau^2, \sigma_\varepsilon^2\right)^\top$.

5 Observed data model and inferential procedure

By considering standard conditional independence assumptions, the misclassification model and the event times model induce a marginal model for observed data, which are the sequences $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ of binary indicators of possibly misclassified event statuses of all units of all subjects. Unknown parameters of the marginal model are the specificities $\boldsymbol{\eta}$ and sensitivities $\boldsymbol{\alpha}$ from the misclassification model, and the parameters $\boldsymbol{\theta}$ from the event times model. Following the results obtained by García-Zattera et al. (2012), we consider the following restrictions on the misclassification parameter space to avoid identification problems: $\eta_{(q,j)} + \alpha_{(q,j)} > 1$ for all $q = 1, \ldots, Q$, $j = 1, \ldots, J$.

The inferential procedure has been implemented in a Bayesian framework using the Monte Carlo Markov chain (MCMC) methods. The software implementation is available in recent versions of the R (R Core Team, 2014) contributed package `bayesSurv` (≥2.3).

6 Simulation study

To illustrate the behavior of the proposed model and to asses the effect of the misclassification process, we conducted a simulation study under the scenarios which mimic to a certain extent the ST data. The time-to-event data were simulated using the following random intercept AFT model

$$\log\left(T_{(i,j)}\right) = \beta_0 + \beta_1 x_{(i,j,1)} + \beta_2 x_{(i,j,2)} + b_i + \varepsilon_{(i,j)}, \qquad i = 1, \ldots, N,\ j = 1, \ldots, 4,$$

where $x_{(i,j,1)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$, $x_{(i,j,2)} \overset{\text{i.i.d.}}{\sim} \text{Bern}(0.5)$, $\beta_0 = 2.00$, $\beta_1 = 0.20$ and $\beta_2 = -0.10$, $\varepsilon_{(i,j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, while considering different values of the error variance σ_ε^2. Furthermore, several settings for the shape of the random effects distribution were considered. In this paper, we show selected results for a scenario in which the random effects $b_1, \ldots, b_N$ followed a Gumbel distribution transformed to have mean zero and variance σ_b^2. The variances σ_ε^2 and σ_b^2 were chosen such that the overall variance $\sigma_b^2 + \sigma_\varepsilon^2$ was constantly equal to 0.1 and different scenarios corresponding to different proportions between the random effects and error variability were considered. In this paper, we show results for settings with ratio $\sigma_b/\sigma_\varepsilon$ being 0.5 and 5. Finally, results for three sample sizes of $N = 500, 1000, 2000$ will be shown.

The true time-to-event were interval-censored by simulating the "visit" times for each subject. We considered $K_i = 10$. The first visit time was

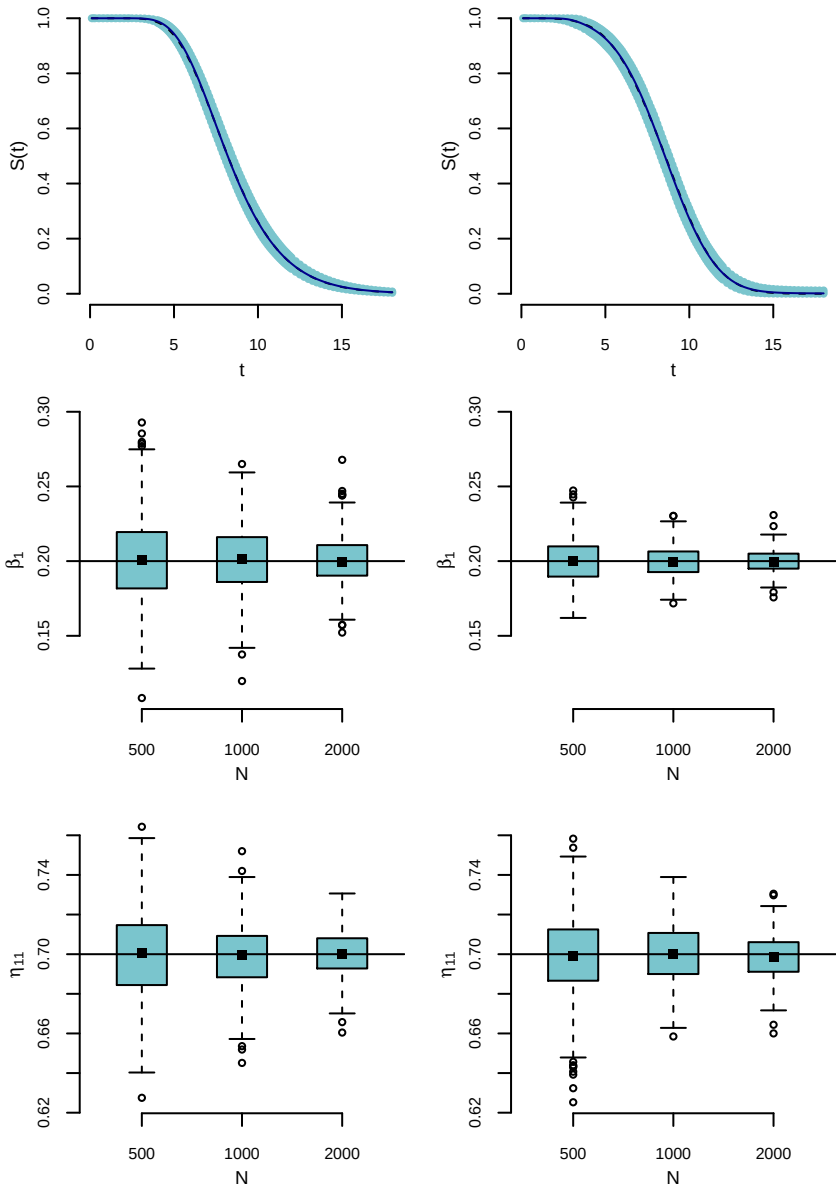


FIGURE 1. Simulation study, left panel: $\sigma_b/\sigma_\varepsilon = 0.5$, right panel: $\sigma_b/\sigma_\varepsilon = 5$. First row: true value (dashed line under the solid line), mean across simulations of the posterior mean (solid line) and the simulation based pointwise 95% confidence band for the marginal survival function with $x_{(i,j,1)} = 0.5$, $x_{(i,j,2)} = 0$ and $N = 500$. Second and the third row: boxplots across simulations of the posterior means of parameters $\beta_1 = 0.20$ and $\eta_{(1,1)} = 0.70$.

randomly chosen from an $\mathcal{N}(3.0,\ 0.2^2)$ distribution. The time between the consecutive visits was drawn from an $\mathcal{N}(1.0,\ 0.1^2)$ distribution. We assume that the assessment of the occurrence of the event was performed by $Q = 5$ examiners, allocated randomly to each subject and visit. Unit-specific sensitivity and specificity parameters were assumed ranging from 0.60 to 0.96.

For each scenario, 500 replicates were generated. Selected results illustrating usefulness of our approach are shown on Figure 1. It shows, for the lowest considered sample size of $N = 500$, the mean across simulations of the posterior mean of the marginal (random effects being integrated out) survival function for a particular covariate combinations. It is practically indistinguishable from the true survival function. Moreover, the simulation based pointwise 95% confidence band becomes even narrower when the sample size increases (not shown here) suggesting that despite the misclassification we are able to estimate the survival function not only unbiasedly but also consistently. The remaining part of Figure 1 suggests that also the model parameters from both the event times and the misclassification model are being estimated unbiasedly and consistently by the posterior means of the Bayesian model.

Acknowledgments: The second author was supported by Fondecyt grant 11110033. The third author was supported by Fondecyt grant 1141193. The work was partially performed during a visit of the first author to the Pontificia Universidad Católica de Chile, supported by Fondecyt grant 11110033.

References

García-Zattera, M. J., Jara, A., Lesaffre, E., and Marshall, G. (2012). Modelling of multivariate monotone disease processes in the presence of misclassification. *Journal of the American Statistical Association*, **107**, 976 – 989.

Komárek, A. and Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly-interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association*, **103**, 523 – 533.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
URL `http://www.R-project.org/`.

Vanobbergen, J., Martens, L., Lesaffre, E., and Declerck, D. (2000). The Signal-Tandmobiel® project – a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, **2**, 87 – 96.

Bayesian spatial disaggregating of shares: application to land use shares in EU

Matieyendou Lamboni[1], Renate Koeble[1], Adrian Leip[1]

[1] JRC-Institute for Environment and Sustainability, ISPRA - ITALY

E-mail for correspondence: matieyendou.lamboni@jrc.ec.europa.eu

Abstract: Shares of available resources, such as land resources, at a fine meshed grid of 1km*1km are required as *a priori* information for i) managing these resources at a local level by disaggregating the resource scenarios to a spatial explicit scale; ii) allowing the development of local environmental indicators, many of which depend on the local combination of land uses and environmental conditions. In this paper, we develop a Bayesian multinomial logit model for disaggregating the observations of shares, available only at a large scale (NUTS3 regions). The model runs at a fine scale, but, is aggregated at a large scale to integrate the observations of shares. Results of land use shares in EU countries are interesting for main crops.

Keywords: Bayesian modeling; disaggregating of shares; Prediction of land use.

1 Methods and developments

1.1 Resources share model at a fine scale

We consider a share model at a fine meshed grid (local level: h), which is characterized by the explanatory variables X_h and by its total resource A_h such as total land to be shared. The local shares model is a classical multinomial logit model. If $\mathbf{Y}$ is the vector of L possible and exclusive uses of the resource A_h and X_h is the vector of d independent variables; the share model $(S_{h,l})$ for each category of resource' use l is:

$$S_{h,l} = \frac{\exp^{(\beta_l^{\mathrm{T}} \mathbf{x}_h)}}{\sum_{l=1}^{L} \exp^{(\beta_l^{\mathrm{T}} \mathbf{x}_h)}} \times A_h \,, \tag{1}$$

with β_l the model parameters for the category of resource' use l.
The Multinomial Logit model (MNL) in equation (1), can be represented as the random utility model RUM (McFadden , 1974; Fruhwirth-Schmatter

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

and Fruhwirth, 2012). By choosing a baseline category, the difference random utility model dRUM is given as:

$$\mathbf{Z}_h = \mathcal{X}_h \beta + \epsilon_h \,, \tag{2}$$

where $\mathbf{Z}_h = [Z_{h,1}, \ldots, Z_{h,L-1}]^{\mathrm{T}}$ and $Z_{h,l} = log(\frac{S_{h,l}}{S_{h,L}})$ is the logarithm of the relative proportion of shares with respect to baseline category L. $\beta = \left[\beta_1^{\mathrm{T}}, \ldots, \beta_{L-1}^{\mathrm{T}}\right]^{\mathrm{T}}$ is a vector of model parameters. $\mathcal{X}_h = \mathbb{I} \otimes \mathbf{x}_h^{\mathrm{T}}$ is a matrix of input variables. ϵ_h is the error terms and each component of ϵ_h follows a standard logistic distribution.
In this paper, to avoid the expensive computational time in order to get the Bayesian estimates and as the logistic distribution is a special case of a normal distribution, we assume that the vector of transformed shares $\mathbf{Z}_h$ follows:

$$\mathbf{Z}_h \sim \mathcal{N}\left(\mathcal{X}_h \beta, R\right) \,, \tag{3}$$

with $R = (R_{i,i} = \pi^2/3,\ R_{i,j|i\neq j} = \pi^2/6)$

1.2 Likelihood of the model and posterior distributions

The relative proportions of resource shares ($\exp(\mathbf{Z}_h)$) follows a multivariate lognormal distribution and the geometric mean of unknown shares ($\exp(\mathbf{Z}_h)$) over all the fine scale (h) within a given large scale (NUTS2) follows a multivariate lognormal distribution if we assume independence between $\mathbf{Z}_h$. We use the geometric mean as we are working with relative proportions and as it allows for getting an analytic likelihood. Formally, if $\mathbf{Y}_n = [Y_{n,1}, \ldots, Y_{n,L-1}]^{\mathrm{T}}$ $(n = 1, 2, \ldots, N)$ are a vector of the logarithm of the known relative proportions at a large scale; $\mathcal{X}_n = \frac{1}{H_n} \sum_{h=1/h\in NUTS2}^{H_n} \mathcal{X}_h$ and $R_n = R/H_n$ with H_n the number of fine scale (h) within NUTS2, the general Bayesian linear model (Smith 1973) is defined as follows:

$$\pi\left(\mathbf{y}_n/\beta\right) \sim \mathcal{N}\left(\mathcal{X}_n \beta, R_n\right) \tag{4}$$

$$\pi\left(\beta\right) \sim \mathcal{N}\left(\beta_0, B_0\right) \,, \tag{5}$$

and the conjugate *a posteriori* distributions follow a multivariate normal distribution (Smith 1973; Fruhwirth-Schmatter and Fruhwirth, 2012),
$\pi\left(\beta/\mathbf{y}\right) \sim \mathcal{N}\left(\beta^*, B\right)$, with :

$$\beta^* = B\left(B_0^{-1}\beta_0 + \sum_{n=1}^{N} \mathcal{X}_n^{\mathrm{T}} R_n^{-1} \mathbf{y}_n\right) \tag{6}$$

$$B = \left(B_0^{-1} + \sum_{n=1}^{N} \mathcal{X}_n' R_n^{-1} \mathcal{X}_n\right)^{-1} \tag{7}$$

2 Application to Land use shares in EU countries

We apply the share model to disaggregate the crop shares from the administrative levels (NUTS2/3) to a local level or Homogenous Spatial Unit (HSU: h). The HSU is a spatial unit at a fine meshed grid, being constructed from 1km x 1km pixel.

2.1 Explanatory variables: (X)

To better discriminate all the land use within a HSU, we choose the biophysical variables used during the process of delineating the HSUs such as the soil texture, organic content, sand, and clay: the relief (slope and altitude); the meteorological parameters (annual rainfall, sum of temperature and vegetation period); land cover classes (CORINE). To these variables, we add the prices of main crops and meats such as wheat, barley, rape seeds, potatoes, milk, beef and pork as the prices of main crops and meats can impact the decision of farmers.

2.2 Available land use shares

The Farm Structure Survey (FSS) of 2010, conducted by EUROSTAT provides the observations of land use areas or shares at an administrative level (called NUTS3) for some European countries. The FSS data gives the distribution of agricultural land use shares within each NUTS2 region. Forest data comes from the European Forest cover maps and is available at 25m×25m grid.

2.3 Choice of prior distributions

To have a conjugate posterior distribution, we assume that the prior distributions of the model parameters follow a multivariate normal distribution. The mean and the covariance matrix are previously estimated by using the multinomial logit regression with the LUCAS survey data of land use/cover (Lamboni *et al.*, 2013). LUCAS survey is a point-based observation and it gives at any sample point, the land cover found.

2.4 Results

To assess the accuracy of the Land Share Model, used to predict the crop shares at HSU level Figure 1 shows the QQ-plots of aggregated predictions of land share versus observations at NUTS3 level. As the forest areas are already available at a HSU level, we have constrained the predictions of forest shares to match with the observations. Moreover, we have left out the land use with zero as share at the large scale or NUTS3.
The predictions of the most frequent crops with available observations are very interesting. For non-frequent crops such as flower (FLOW), nurseries (NURS), tobacco (TOBA), there is a need to add additional constraints so that the predictions will be consistent with the observation at NUST3 level.

3 Conclusion

In this paper, we investigate the predictions of crop shares over the new Homogeneous Spatial Unit (HSU) by using the Land share Model (LSM).

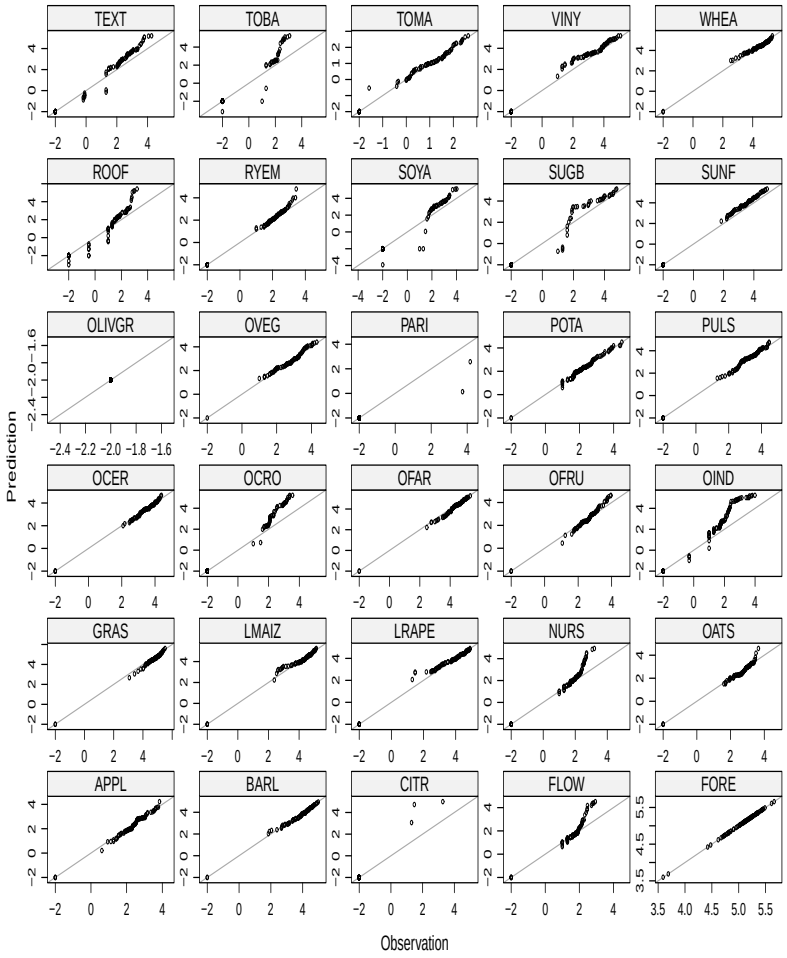


FIGURE 1. QQ-plots of aggregated predictions of shares at NUTS3 in France (in $\log_{10}$). On the figure, the value of $\log_{10}(0)$ (constrained prediction) is represented by -2 to avoid the concentration of points in one side.

Bayesian modelling helps for integrating the observations of land use available at a large scale (NUTS3) and the point-based observations (LUCAS survey data) via a local share model. The predicted crop shares at a fine scale are reasonable and interesting for frequent crops. The constrained, predicted shares will be integrated into CAPRI system (Common Agricultural Policy Regionalized Impact) for deriving environmental indicators and for linking different models.

References

Fruhwirth-S., S. and Fruhwirth, R. (2012). Bayesian inference in the multinomial logit model. *Austrian Journal of Statistics*, **41**, 27 – 43.

Lamboni M., Koeble R., Leip A. (2013). Prediction of crop land shares for environmental impact assessment over EU. In: *Proceedings of ICAS VI Conference*, Brazil.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviours. In: P. Zarembka (Ed), *Frontiers of Econometrics* New York: Academic, pp. 105 – 142.

Smith A. F. (1973). General Bayesian linear model. *Journal of the Royal Statistical Society*, **35**, 67 – 75.

Joint modeling of a multilevel factor analytic model and a multilevel covariance regression model

Emmanuel Lesaffre[1] [2], Baoyue Li[1], Luk Bruyneel[3]

[1] Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands
[2] L-Biostat, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium
[3] Centre for Health Services and Nursing Research, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium

E-mail for correspondence: E.Lesaffre@erasmusmc.nl

Abstract: We propose a novel modeling approach that could model both the mean structure and the covariance structure with a mixed effects model in a multivariate context. We call this multilevel covariance regression (MCR) model. When the dimension of the response is high, a joint model of a multilevel factor analytic (MFA) model and an MCR model is then proposed.

Keywords: multivariate; multilevel; covariance regression; factor analytic model.

1 Introduction

A traditional multilevel regression model assumes a constant residual variance (homoscedasticity) after adjusting for fixed and random effects. When homoscedasticity does not hold, the variance may depend on covariates and even random effects. Modeling this on top of a traditional multilevel regression model could be quite challenging, and even more challenging for a covariance matrix in the case of a multivariate response. We propose a novel modeling approach that extends the multivariate multilevel regression model in the following two ways: 1) the covariance matrix of the multivariate response is modeled with both fixed and random effects, which is called the multilevel covariance regression (MCR) model (Li et al., 2013), 2) for a high-dimensional multivariate response, we propose to combine a multilevel factor analytic (MFA) model with the MCR model by using the factor scores as the multiple responses in the MCR model, resulting in the multilevel higher-order factor (MHOF) model (Li et al., 2014).

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

2 Proposed models

A p-variate 2-level MCR model has the following form:

$$\boldsymbol{z}_{ij} = \boldsymbol{B}\boldsymbol{x}_{ij} + \boldsymbol{u}_j + \boldsymbol{\delta}_{ij},$$
$$\boldsymbol{\delta}_{ij} = \boldsymbol{\lambda}_{ij} F_{ij} + \boldsymbol{\varepsilon}_{ij}, \quad \boldsymbol{\lambda}_{ij} = \boldsymbol{B}^* \boldsymbol{x}^*_{ij} + \boldsymbol{u}^*_j,$$

with

$$\boldsymbol{u}_j \sim N(0, \Sigma_u), \quad \boldsymbol{u}^*_j \sim N(0, \Sigma^*_u),$$
$$F_{ij} \sim N(0,1), \quad \boldsymbol{\varepsilon}_{ij} \sim N(0, \Sigma_\varepsilon),$$

where the p-variate response $\boldsymbol{z}_{ij}$ is modeled using fixed effects $\boldsymbol{B}\boldsymbol{x}_{ij}$ and random effects $\boldsymbol{u}_j$, and the residual $\boldsymbol{\delta}_{ij}$ is further modelled with a factor analytic model having a single factor F_{ij} with a structured factor loading $\boldsymbol{\lambda}_{ij}$. The implied covariance matrix of the response, constructed by the factor model, therefore depends on both fixed and random effects.
Our proposed MHOF model further replaces the response $\boldsymbol{z}_{ij}$ in an MCR model with factor scores from an MFA model, and estimates both models simultaneously. A 2-level MFA model is:

$$\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{b}_j + \boldsymbol{L}\boldsymbol{z}_{ij} + \boldsymbol{\varepsilon}^{FA},$$

with

$$\boldsymbol{b}_j \sim N(0, \Sigma_{bu}), \quad \boldsymbol{\varepsilon}^{FA}_{ij} \sim N(0, diag(\sigma^2_1, \sigma^2_2, ..., \sigma^2_q)),$$

where the observed response $\boldsymbol{y}_{ij}$ has q dimensions with $q \gg p$. The p-dimensional common factor $\boldsymbol{z}_{ij}$ is further used as the response in an MCR model. In this way, the MHOF model can handle high-dimensional responses.

3 RN4CAST data set and research questions

We first applied the MCR model to data from the RN4CAST (Sermeus et al. 2011) FP7 project which involves 33,731 registered nurses in 2,169 nursing units in 486 hospitals in 12 European countries. The MHOF model was applied to the Belgium part of the project. As response we have taken in the first analysis the historically derived three burnout dimensions (Maslach and Jackson, 1981), while the MHOF model is based on the raw data, i.e. the responses to the 22-item questionnaire. The three burnout dimensions are emotional exhaustion (EE), depersonalization (DP) and personal accomplishment (PA). Applying the MHOF model to burnout could address the following questions simultaneously: 1) is the burnout structure the same as the commonly used structure by Maslach and Jackson? 2) how much variation of burnout could be explained by the level-specific fixed and random effects? 3) do the variances and correlations among burnout stay constant across level-specific characteristics and units at each level?

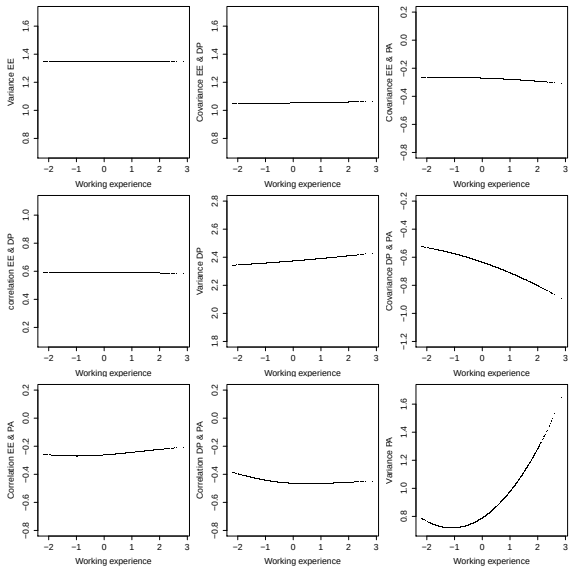


FIGURE 1. (co)Variances (upper triangle) and correlations (lower triangle) with *working experience* at the nurse level.

4 Computational aspects

We opted for the Bayesian approach as our estimating method for the MCR and MHOF models. The JAGS (just another Gibbs sampler) MCMC (Markov chain Monte Carlo) program was used through the R package rjags. Most parameters were assigned a non-informative prior except for the fixed and random effects in the factor loadings in the MCR part. These parameters were assigned a mixture prior respectively to overcome the "flipping states" issue in Bayesian context. Model comparison was done using the pseudo Bayes factor (PSBF).

5 Simulation study

A limited simulation study was performed that compared the parameters estimates from the MHOF model and the two-stage model, i.e. first run an MFA model and then model the factor scores with an MCR model.
Compared with the MHOF model, the two-stage model is less computationally intensive, but it estimates the regression coefficients biasedly in both the mean structure and the covariance structure, as well as some variance/covariance parameters of the random effects.

6 Main results

For the mean structure of the three-dimensional burnout response, after taking into account the multilevel structure, several covariates at each level

are found relevant. Fulltime nurses suffer more from burnout than part-time nurses, and more experienced nurses suffer less from burnout. At higher levels (nursing unit and hospital levels), better work environment and heavier work load within a nursing unit/hospital result in less burnout nurses.
The covariance structure provides additional insights into the burnout phenomenon. Findings suggest a significantly larger variation in personal accomplishment for experienced nurses (see Figure 1). The justification of including random effects implies that the covariances and the correlations among burnout are different across hospitals and nursing units. The fixed effects reflect the temporal or demographical measurement of the variation of burnout, while the random effects reflect the spatial or geographical measurement of the variation of burnout.

7 Conclusions

The proposed MHOF model provides a way to directly assess the heteroscedasticity of the multi-dimensional lower-order factor scores in a complex situation. This modeling can reveal some "hidden" information that may have never been obtained through modeling only the mean of the measurements, thus could provide valuable information for the administration of the hospitals and nursing units in nursing affairs. Our methods apply equally well to numerous research topics in psychology, sociology and political science, to name a few, which often deal with multilevel research designs, latent constructs, and an interest in covariance regression.

References

Li, B., Bruyneel, L., and Lesaffre, E. (2013). A multivariate multilevel Gaussian model with a mixed effects structure in the mean and covariance part. *Statistics in Medicine*, DOI: 10.1002/sim.6062.

Li, B., Bruyneel, L., and Lesaffre, E. (2014). Multilevel higher-order factor model: Joint modeling of a multilevel factor analytic model and a multilevel covariance regression model. *Structural Equation Modeling: A Multidisciplinary Journal*, (under review)

Maslach, C. and Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior*, **2**, 99 – 113.

Sermeus, W., Aiken, L., Van den Heede, K., et al. (2011). Nurse forecasting in Europe (RN4CAST): Rationale, design and methodology. *BMC Nursing*, **10**, 6.

Separate regression modelling of the Gaussian and Exponential components of an EMG response from respiratory physiology

Gianfranco Lovison[1], Christian Schindler[2]

[1] Department od Economics, Business and Statistics, University of Palermo, Italy
[2] Swiss Tropical and Public Health Institute, Basel, Switzerland; University of Basel, Basel, Switzerland

E-mail for correspondence: gianfranco.lovison@unipa.it

Abstract: If $Y_1 \sim \mathcal{N}(\mu, \sigma^2)$ and $Y_2 \sim Exp(\nu)$, with Y_1 independent of Y_2, then their sum $Y = Y_1 + Y_2$ follows an Exponentially Modified Gaussian (EMG) distribution. In many applications it is of interest to model the two components separately, in order to investigate their (possibly) different important predictors. We show how this can be done through a GAMLSS with EMG response, and apply this separate regression modelling strategy to a dataset on lung function variables from the SAPALDIA cohort study.

Keywords: Exponentially Modified Gaussian distribution; GAMLSS; Deconvolution.

1 Introduction

The sum of two independent r.v.'s, one Gaussian and one Exponential, follows an Exponentially Modified Gaussian (EMG) distribution. Such a distribution has found interesting applications in some specific areas: modelling inter-mitotic time in genetics (Golubev, 2009), response times in experimental psychology (Palmer et al., 2011), peaks in chromatography, but seems to have received very little attention in biostatistics. We show in this paper how to fit separate regression models to the two components of an EMG response through a GAMLSS, and apply this separate regression modelling strategy to one of the lung function variables which arise in spirometry.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

2 The Exponentially Modified Gaussian distribution

If $Y_1 \sim \mathcal{N}(\mu, \sigma^2)$ and $Y_2 \sim Exp(\nu)$, where $\nu = E(Y_2)$, with Y_1 independent of Y_2, then their sum $Y = Y_1 + Y_2$ follows an **Exponentially Modified Gaussian** (EMG) distribution, and one can then write $Y \sim \mathcal{EMG}(\mu, \sigma, \nu)$. By convolution, the p.d.f. of $Y \sim \mathcal{EMG}(\mu, \sigma, \nu)$ can be shown to be:

$$f_Y(y; \mu, \sigma, \nu) = \frac{1}{2\nu} \exp\left[\frac{1}{2\nu}(2\mu + \frac{\sigma^2}{\nu} - 2y)\right] \text{ erfc}\left(\frac{\mu + \frac{\sigma^2}{\nu} - y}{\sqrt{2}\sigma}\right) \quad (1)$$

where $\text{erfc}(z) = 1 - \text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty \exp(-t^2)dt$ is the complementary error function. Exploiting the known relation: $\text{erfc}(\frac{z}{\sqrt{2}}) = 2\Phi(-z)$, where $\Phi(\cdot)$ is the Standard Normal distribution function, (1) can be written in the following form, perhaps more familiar to statisticians:

$$f_Y(y; \mu, \sigma, \nu) = \frac{1}{\nu} \exp\left(\frac{\mu - y}{\nu} + \frac{\sigma^2}{2\nu^2}\right) \Phi\left(\frac{y - \mu}{\sigma} - \frac{\sigma}{\nu}\right) \quad (2)$$

This is the parameterisation used by the R library `gamlss` (Rigby and Stasinopoulos, 2007) and adopted in this paper. The following expressions for the first four moments can be easily derived:

$$E[Y] = \mu + \nu; \quad Var[Y] = \sigma^2 + \nu^2;$$

$$Sk[Y] = 2\left(1 + \frac{\sigma^2}{\nu^2}\right)^{-\frac{3}{2}}; \quad Ku[Y] = 6\left(1 + \frac{\sigma^2}{\nu^2}\right)^{-2}.$$

Our interest in the EMG distribution arose in the study of lung function variables, where it accommodates in a flexible way both the (positive) skewness and the "peakedness" which characterise such variables. This flexibility, along with the possibility of a mechanistic interpretation of its derivation as the convolution of a Gaussian and an Exponential distribution, have motivated our preference for this distribution over other well-fitting, but somewhat more complex and less interpretable, positively skewed distributions, in analysing the dataset presented in Sec. 4.

3 Regression models for the Gaussian and Exponential components of an EMG response

Suppose a response variable Y is known to be the sum of two unobservable components Y_1, Y_2, which are of substantive interest, and that two GLMs: $\mathcal{M}_1 : E[Y_1] = h_1(\boldsymbol{X\beta}); Var[Y_1] = \phi_1 V(E[Y_1])$ and $\mathcal{M}_2 : E[Y_2] = h_2(\boldsymbol{Z\gamma}); Var[Y_2] = \phi_2 V(E[Y_2])$ are set up for modelling the effects of explanatory variables $\boldsymbol{X}$ and $\boldsymbol{Z}$ on the expected values of the two components; the model matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ can be formed by the same, by partly different or by completely separated explanatory variables.

Clearly, in general, if only the "convoluted response variable" $Y = Y_1 + Y_2$ is available, there will be serious problems of identifiability and estimability of the parameters $(\boldsymbol{\beta}, \phi_1)$ and $(\boldsymbol{\gamma}, \phi_2)$ in the two separate GLMs, depending on the degree of separation and orthogonality of $\boldsymbol{X}$ and $\boldsymbol{Z}$. This difficulty parallels the complexity of deconvolving the distribution of the sum of two r.v.'s.

From this point of view, an EMG response Y is a fortunate exception. As outlined above, the location parameter of the Gaussian component enters only in the expression of $E[Y]$, while, for fixed σ, the higher moments depend only on the location parameter ν of the Exponential component. This makes it possible to specify two separate regression models for the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, assuming σ unknown but fixed:

$$\boldsymbol{y} \sim EMG(\boldsymbol{\mu}, \sigma, \boldsymbol{\nu}) \tag{3}$$

$$\boldsymbol{\mu} = h_\mu(\boldsymbol{X}\boldsymbol{\beta}) \tag{4}$$

$$\boldsymbol{\nu} = h_\nu(\boldsymbol{Z}\boldsymbol{\gamma}) \tag{5}$$

and to consider (3), (4) and (5) as a GAMLSS with EMG response distribution (Rigby, Stasinopoulos, 2005).

4 Application to respiratory physiology

SAPALDIA (Swiss Cohort Study on Air Pollution and Lung and Heart Diseases In Adults) is a large population-based cohort study, initiated in 1991 in eight areas of Switzerland. Participants were between 18 and 60 years old at recruitment. They were re-examined in 2002 and 2010/11. Besides responding to a computer-based interview with detailed questions on respiratory health and allergies, lifestyle, socio-demographic characteristics, home and workplace environment, study participants also underwent several examinations, including lung function testing. Methodological details are provided in Martin et al. (1997). SAPALDIA spirometry data have been used to derive sex-, age- and height- based reference equations for lung function variables in adults (Brändli et al., 1996 and 2000). Since the focus of these analyses was on modeling percentile functions, quantile regression methods were applied. Later, with the advent of GAMLSS modelling and related software, it became possible to fit models with skewness and kurtosis parameters. The Global Lung Function Initiative used this new methodological framework to develop a global set of spirometric reference equations for adults and children taking into account differences according to geography and race (Cole et al., 2009, Quanjer et al., 2012).

Two fundamental outcome variables of spirometry (i.e., lung function testing) are FVC, the Forced Vital Capacity of the lung, and FEV_1, the Forced Expiratory Volume in the 1st second. We focus in this paper on the difference $FEV_{a1} = FVC - FEV_1$, where FEV_{a1} stands for "Forced Expiratory Volume after the 1st second".

An extensive exploratory analysis on FEV_{a1} has shown a surprisingly good fit of the EMG distribution to the observed data. It is not yet clear whether

this reflects a precise causal mechanism, related to the physiology of respiration. In any case, it is of interest to try to find out the determinants of the two components, the Gaussian and the Exponential, through the approach outlined in Sec. 3.

For the purpose of illustration, we fitted the GAMLSS defined in (3), (4) and (5) to the sub-sample of male non-smokers in the first (1991) SAPALDIA survey; in keeping with the default options in `gamlss`, we chose $h_\mu = identity$ and $h_\nu = log$. The results of the final model, chosen through a stepwise procedure based on AIC, are reported in Table 1. Inspection of the table shows that the individual characteristics (Age, Height and BMI) combine in different ways to determine the Gaussian and Exponential components. In particular, BMI has a strong, both linear and quadratic, effect on the Gaussian component, along with an interaction with Age, but no significant effect on the Exponential component.

TABLE 1. Parameter estimates for the EMG model

	Estimate	Std. Error	t value	p-value
Regression model for the Gaussian component				
Intercept	-13.2116	3.14420	-4.20	0.00002
Age	0.0503	0.00586	8.58	0.00000
Height	0.1118	0.03596	3.10	0.00189
BMI	0.1288	0.02031	6.34	0.00000
Age^2	-0.0001	0.00005	-3.42	0.00063
$Height^2$	-0.0002	0.00010	-2.69	0.00718
BMI^2	-0.0013	0.00042	-3.27	0.00105
Age:BMI	-0.0010	0.00024	-4.29	0.00001
$\log(\sigma)$	-1.231	0.02141		
Regression model for the Exponential component				
Intercept	-5.8710	0.82903	-7.08	0.00000
Age	-0.0368	0.01581	-2.33	0.01978
Height	0.0290	0.00430	6.74	0.00000
Age^2	0.0005	0.00019	2.80	0.00500

An insightful way of presenting this model is to plot the two estimated component densities for a subject with a given combination of explanatory variables. As an example, in Figure 1 the plots on the same row have the same combination of Age and Height (top row: Age=20 yrs., Height=175 cm.; bottom row: Age=60 yrs., Height= 195 cm.), and therefore they have the same Exponential component. The left and right plot in each row differ only by BMI (left panel: BMI=24 kg/m^2, right panel=48 kg/m^2), and therefore their comparison helps to visualise the role of BMI, which affects only the Gaussian component. From inspection of these plots, it is evident how older and taller people have a "flatter" (i.e. with larger mean) Exponential component, and also a more marked effect of BMI on reducing the

mean of the Gaussian component.

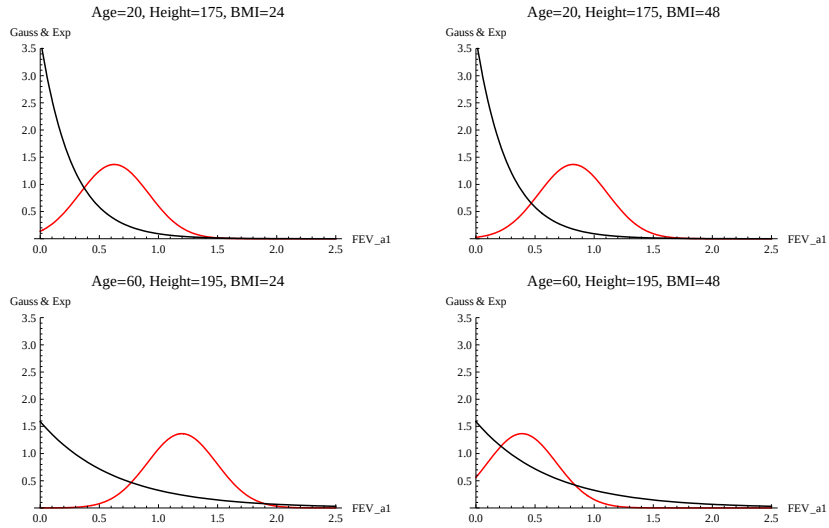


FIGURE 1. Estimated Gaussian and Exponential components for four exemplary individuals

The interplay of Age, Height and BMI in determining the two component distributions can be appreciated in Figure 2, where we report the estimated Gaussian and Exponential components for the two "extreme" individuals (i.e. having the two largest and smallest combinations of estimates $(\hat{\mu}, \hat{\nu})$) in our sample: in the left panel, a 51 years old man 197 cm. tall and with BMI = 27.1 kg/m^2; in the right panel, a 21 years old man, 164 cm. tall and with BMI = 19.3 kg/m^2.

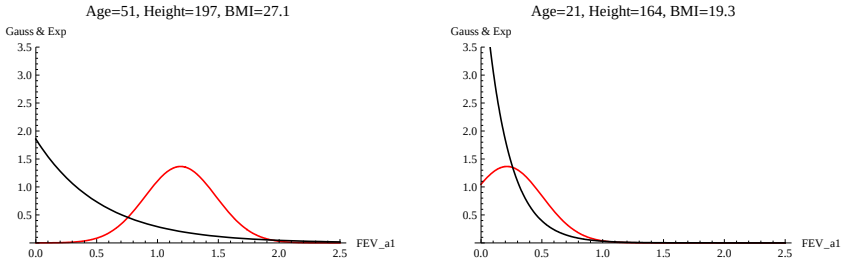


FIGURE 2. Estimated Gaussian and Exponential components for two "extreme" individuals

The combined effect of the three variables yields larger values of FEV_{a1} in the older, taller and overweight subject in the left panel compared to the younger, shorter and normal weight subject in the right panel: this is the consequence of both the Exponential and the Gaussian components being shifted to the right for the latter compared to the former. In interpreting

these findings, one should keep in mind that a large value of the $\frac{FEV_{a1}}{FEV_1}$ ratio is an indicator of obstructed expiration.

References

Brändli, O., Schindler, C., Künzli, N., et al. (1996). Lung function in healthy never smoking adults: reference values and lower limits of normal of a Swiss population. *Thorax*, **51**, 277 – 283.

Brändli, O., Schindler, C., Leuenberger, P., et al. (2000). Re-estimated equations for 5th percentiles of lung function variables. *Thorax*, **55**, 173 – 174.

Cole, T.J., Stanojevic, S., Stocks, J., et al. (2009) Age- and size-related reference ranges: A case study of spirometry through childhood and adulthood. *Statistics in Medicine*, **28**, 880 – 898.

Golubev, A. (2009). Exponentially Modified Gaussian (EMG) relevance to distributions related to cell proliferation and differentiation. *Journal of Theoretical Biology*, **6**, 15 – 51.

Martin, B.W., Ackermann-Liebrich U., Leuenberger, P., et al. (1997). SAPALDIA: Methods and participation in the cross-sectional part of the Swiss Study on Air Pollution and Lung Diseases in Adults. *Sozial- und Präventivmedizin*, **42**, 67 – 84.

Palmer, E.M., Horowitz Todd, S., Torralba, A. et al. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology*, **37**, 58 – 71.

Quanjer, P.H., Stanojevic, S., Cole, T.J., et al., (2012) Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *European Respiratory Journal*, **40**, 1324 – 1343.

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

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1 – 46.

New perspectives on rating data modelling: the Nonlinear CUB

Marica Manisera[1], Paola Zuccolotto[1]

[1] University of Brescia, Italy

E-mail for correspondence: marica.manisera@unibs.it

Abstract: In this contribution, rating data are modelled using the new class of Nonlinear CUB models, recently introduced to generalize the standard CUB. A case study is presented, with an application to Eurobarometer data, concerned with the European citizens' perceptions about economy.

Keywords: CUB models; Rating data; Likert-type scales; Latent variables; Transition probability.

1 Introduction

Rating data are very common in several fields where the individuals' perceptions and attitudes are often investigated by means of Likert-type scales or, more generally, questionnaires with several questions whose possible responses are ordered. Among the methods and techniques proposed in the literature to model rating data (see for example Agresti, 2013; Tutz, 2012), an interesting proposal is given by CUB models (Piccolo, 2003; D'Elia and Piccolo, 2005). Several papers concerning CUB inferential issues, identifiability problems, fitting measures, computational strategies and software routines have been published (see Iannario and Piccolo, 2012, 2014 and the references therein). In addition, CUB models have been extended in several directions (for example, Iannario, 2012a,b; Grilli *et al.*, 2013; Manisera and Zuccolotto, 2014b; Piccolo, 2014) and applied in different fields (for example, Iannario *et al.*, 2012). A generalization of CUB models is the so-called Nonlinear CUB (NLCUB; Manisera and Zuccolotto, 2014a), a new class of models recently proposed in order to take account of the categorical nature of the rating data. While CUB models imply that the response categories are equally spaced in the respondents' mind, NLCUB can address their (possible) unequal spacing, that is a situation where respondents, in their unconscious search for the "right" answer, find it easier to move, for example, from rating 1 to 2 than from rating 4 to 5. This

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

corresponds to the concept of "nonlinearity" introduced by NLCUB models, defined as the nonconstantness of the transition probabilities (that is the probabilities to move from one rating to the next one). Unlike CUB, NLCUB can be used to model rating data when the transition probabilities are not constant. Simulation studies and real data analyses (Manisera and Zuccolotto, 2014a) show promising results that encourage further research. The aim of this contribution is to present the functioning and some features of the NLCUB models by means of an application to real data coming from the Eurobarometer survey and concerned with the European citizens' perceptions about economy.
The paper is organized as follows: in Section 2 we describe the basic features of CUB models and the new class of NLCUB models. In Section 3 we show the results of the application and draw the main conclusions.

2 CUB and Nonlinear CUB models

CUB models have been introduced in the literature (Piccolo, 2003; D'Elia and Piccolo, 2005) to analyse ordinal data and fit into the latent variable framework. With CUB, rating or ranking data are modelled by a mixture of a Uniform and a Shifted Binomial random variables: the observed rating r ($r = 1, \cdots, m$) is a realization of the discrete random variable R with probability distribution

$$Pr\{R = r|\theta\} = \pi Pr\{V(m, \xi) = r\} + (1-\pi)P\{U(m) = r\} \quad r = 1, 2, \ldots, m$$

with $\theta = (\pi, \xi)'$, $\pi \in (0, 1]$, $\xi \in [0, 1]$. For a given m, $V(m, \xi)$ is a Shifted Binomial random variable, with trial parameter m and success probability $1 - \xi$, modelling the *feeling* component of a decision process, and $U(m)$ is a discrete Uniform random variable defined over the support $\{1, \ldots, m\}$, aimed to model the *uncertainty* component. CUB models are identifiable for $m > 3$. In terms of interpretability, $1 - \xi$ is the *feeling* parameter and measures the agreement with the object being evaluated, while $1 - \pi$ is the *uncertainty* parameter and measures the intrinsic uncertainty in choosing the ordinal response.
Nonlinear CUB models (NLCUB) are a generalization of CUB introduced by Manisera and Zuccolotto (2014a). With NLCUB, the discrete random variable R generating the observed rating r has a probability distribution that depends on a parameter T ($T \geq m - 1$) and is given by

$$Pr\{R = r|\theta\} = \pi \sum_{y \in l^{-1}(r)} Pr\{V(T + 1, \xi) = y\} + (1 - \pi)P\{U(m) = r\}$$

where l is a function mapping from $(1, \cdots, T+1)$ into $(1, \cdots, m)$. In detail, l is defined as

$$l(y) = \begin{cases} 1 & \text{if} \quad y \in [y_{11}, \ldots, y_{g_1 1}] \\ 2 & \text{if} \quad y \in [y_{12}, \ldots, y_{g_2 2}] \\ \vdots & \vdots \quad \vdots \\ m & \text{if} \quad y \in [y_{1m}, \ldots, y_{g_m m}] \end{cases}$$

where y_{hs} is the h-th element of $l^{-1}(s)$, and

$$(y_{11}, \ldots, y_{g_1 1}, y_{12}, \ldots, y_{g_2 2}, \ldots, y_{1m}, \ldots, y_{g_m m}) = (1, \cdots, T+1).$$

We denote with $g_s = |l^{-1}(s)|$, where $|\cdot|$ denotes the cardinality of a set, the number of "latent" values to which rating s corresponds according to l. The values $g_1, \cdots, g_m$ univocally determine the l function and can be considered as parameters of the model. We have $T = g_1 + \cdots + g_m - 1$.
When $T = m - 1$ and $g_s = 1$ for all $s = 1, \cdots, m$, then the proposed model collapses to the classical CUB. In Manisera and Zuccolotto (2014a) the NLCUB formulation is derived as a special case of a general framework describing the Decision Process (DP) that drives individuals' responses to questions with ordered response levels. In this general model, the DP is composed of two different approaches, which, borrowing the CUB terminology, are called feeling and uncertainty approach, respectively. The feeling approach consists of a step-by-step reasoning, which proceeds through T consecutive steps and is called feeling path. At each step, an elementary judgment is given. The last rating of the feeling path results from these elementary judgments that are summarized and transformed into a Likert-scaled rating. The uncertainty approach consists of a random judgment that can be given by an uncertain respondent that hesitates in choosing the ordinal response, due to a variety of reasons. In the end, the expressed rating can derive from the feeling or the uncertainty approach with given probabilities. Some existing statistical models can be viewed as special cases of this general framework.
Due to comparability issues, the feeling parameter in NLCUB is given by the expected number μ of one-rating-point increments during the feeling path, while, $1 - \pi$ still is the uncertainty parameter. An interesting feature of the NLCUB model is the possibility to express the so-called transition probabilities (i.e. the probability of moving to the next rating at the next step of the feeling path), which describe the state of mind of the respondents about the response scale used to express judgments in the feeling path. Transition probabilities account for the above mentioned unequally spacing between response categories, in the sense that when the probability of moving, say, from rating 1 to 2 is higher than that from rating 4 to 5, then ratings 1 and 2 can be interpreted as "closer" than ratings 4 and 5 in the respondents' minds. For ease of interpretation, the average transition probabilities, obtained averaging over the steps, are generally used. Transition probabilities can be transformed, by means of a proper function, into "perceived distances" between two consecutive ratings and used for constructing the so-called transition plot, a graphical representation of the spacing existing between rating categories. A linear transition plot suggests that the ratings are perceived as equally-spaced in the respondents' mind (all the transition probabilities are equal) while a nonlinear transition plot accounts for unequally-spaced perceived ratings. Details about the parameter estimation of NLCUB models, interesting insights about the behaviour of this new class of models, suggestions on the future theoretical developments and some applications are in Manisera and Zuccolotto

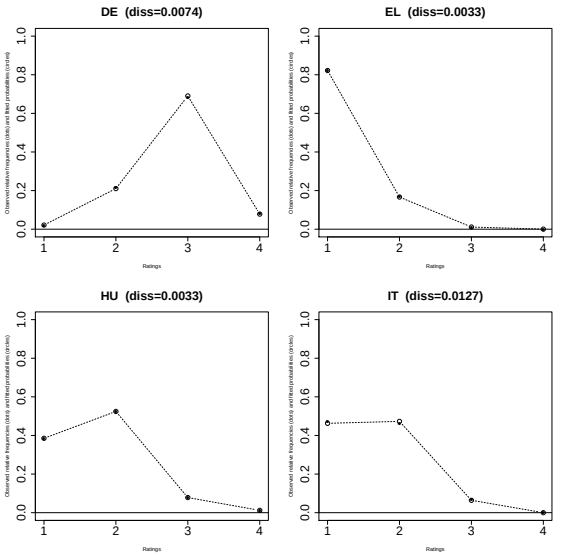


FIGURE 1. Observed relative frequencies vs NLCUB fitted probabilities, Standard Eurobarometer 78 (QA3.1), Germany (top-left) Greece (top-right), Hungary (bottom-left), Italy (bottom-right).

(2014a) and the references therein. R routines for estimation and graphical representations of NLCUB models are freely available upon request to the authors.

3 Case study and conclusions

This section describes a case study dealing with real data coming from the wave 78.1 of Eurobarometer, a sample survey of the European Commission carried out in the 27 EU Member States (European Commission, 2013 - http://ec.europa.eu/public_opinion/archives/eb/eb78/eb78_en.htm). Due to space constraints, the results here shown only focus on the responses given to one question (QA3.1: "How would you judge the current situation of your national economy?", with possible responses given by "very bad", "rather bad", "rather good" and "very good") for four selected countries (Germany, DE; Greece, EL; Hungary, HU; Italy, IT). Figure 1 shows the observed relative frequencies and the corresponding NLCUB fitted probabilities. In all the cases, the NLCUB model almost perfectly fits the observed frequencies.

On the whole, all the respondents from the four considered countries show a very low uncertainty ($1 - \pi$ is 0.06 for Germany and 0.01 for Greece, Hungary and Italy), while some differences can be observed for the feeling parameters μ (in this case $\mu \in [0, 3]$): Greeks are the most pessimistic about their national economy (μ=0.19), followed by Italians (μ=0.60) and Hungarians (μ=0.72). Germans, instead, are much more confident (μ=1.85). Figure 2 shows the transition plots for the four countries. Greece has a linear

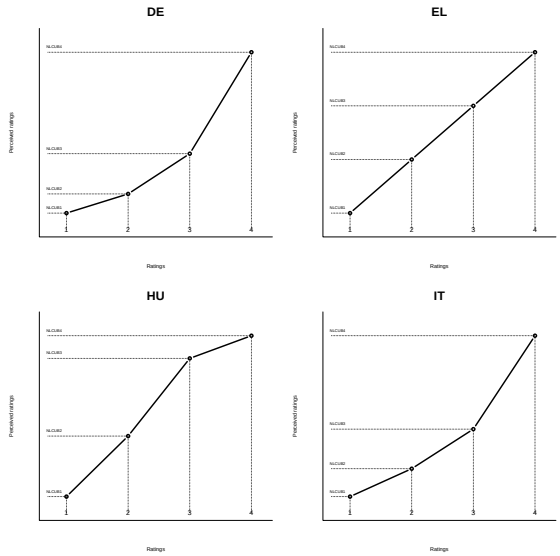


FIGURE 2. Transition plots of the NLCUB models, Standard Eurobarometer 78 (QA3.1), Germany (top-left) Greece (top-right), Hungary (bottom-left), Italy (bottom-right).

transition plot: in this case, the estimated NLCUB model exactly matches the linear CUB structure. The other three countries have nonlinear patterns: Germany and Italy are characterized by a decreasing probability of moving to higher ratings, while this probability is increasing for Hungary. This means that, in general, for Greek respondents moving, for instance, from rating 1 to 2 is as hard as from rating 3 to 4. Instead, Germans and Italians find it easier to move from rating 1 to 2 than from rating 3 to 4. On the contrary, Hungarians find it harder to move from rating 1 to 2 than from 3 to 4.

Concluding, this case study shows how NLCUB models allow us to model rating data resulting from cognitive mechanisms with non-constant transition probabilities, thus extending the possibilities of application of the well-known framework of CUB models.

Acknowledgments: This research was partially funded by STAR project (University of Naples Federico II - CUP: E68C13000020003) and partially by a grant from the European Union Seventh Framework Programme (FP7-SSH/2007-2013); *'SYstemic Risk TOmography: Signals, Measurements, Transmission Channels, and Policy Interventions'* - *SYRTO* - Project ID: 320270.

References

Agresti, A. (2013). *Categorical Data Analysis,* 3^{rd} *edition.* New York: J. Wiley & Sons.

D'Elia, A. and Piccolo, D. (2005). A mixture model for preference data analysis. *Comput Stat Data An*, **49**, 917 – 934.

European Commission (2013). *Eurobarometer 78.1 (2012). Technical Report*. Brussels: TNS Opinion & Social.

Grilli, L., Iannario, M., Piccolo, D., Rampichini, C. (2013). Latent class CUB models. *Adv Data Anal Classif*, **8**, 105 – 119.

Iannario, M. (2012a). Hierarchical CUB models for ordinal variables. *Commun Stat-Theor M*, **41**, 3110 – 3125.

Iannario, M. (2012b). Modelling shelter choices in a class of mixture models for ordinal responses. *Stat Method Appl*, **20**, 1 – 22.

Iannario, M., Manisera, M., Piccolo, D., Zuccolotto, P. (2012). Sensory analysis in the food industry as a tool for marketing decisions. *Adv Data Anal Classif*, **6**, 303 – 321.

Iannario, M. and Piccolo, D. (2012). CUB Models: Statistical Methods and Empirical Evidence. In: R.S. Kenett, S. Salini (eds.): *Modern Analysis of Customer Surveys*, pp. 231 – 258. New York: J. Wiley & Sons.

Iannario, M. and Piccolo, D. (2014). A Short Guide to CUB 3.0 Program. Available at https://www.researchgate.net/publication/260959050.

Manisera, M. and Zuccolotto, P. (2014a). Modeling rating data by Nonlinear CUB models. *Comput Stat Data An*, forthcoming.

Manisera, M. and Zuccolotto, P. (2014b). Modelling "don't know" responses in rating scales. *Pattern Recogn Lett*, forthcoming.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85 – 104.

Piccolo, D. (2014). Inferential issues on CUBE models with covariates. *Commun Stat-Theor M*, forthcoming.

Tutz, G. (2012). *Regression for categorical data*. UK: J. Cambridge University Press.

Network-based Source Detection: From Infectious Disease Spreading to Train Delay Propagation

Juliane Manitz[1], Jonas Harbering[2], Marie Schmidt[2], Thomas Kneib[1], Anita Schöbel[2]

[1] Department of Statistics and Econometrics, Georg-August University Göttingen, Göttingen, Germany
[2] Institute for Numerical and Applied Mathematics, Georg-August University Göttingen, Göttingen, Germany

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

Abstract: The correct identification of the source of a propagation process is crucial for research questions in a wide range of research fields: to determine the epicenter of infectious disease outbreaks, the onset of blackouts in power grids, the root of computer virus attacks in the Internet, the origin of misinformation in social networks, or the starting point of the invasion of non-endemic species in ecology. Here, we consider source determination of train delays in railway systems, which mimic many-faceted diffusion patterns. Delays can never be entirely avoided, but their impact has to be kept to a strict minimum. We enhance a fast and efficient approach for the source identification of propagation processes on networks, which is structurally quite general and only requires a minimum data basis. In extensive simulation studies, we investigate the performance in dependency of time and network node centrality. We examine the robustness of the approach by the application of different delay management strategies, which mimic various propagation mechanisms. Finally, we test for performance improvement due to the integration of additional knowledge in the network definition. Altogether, the source detection framework turns out to be robust for diverse spatio-temporally evolving processes, which promises the general applicability in many research fields.

Keywords: Source Detection; Complex Network; Public Transportation.

1 Introduction

Many spreading phenomena, e.g., the transmission of diseases and the propagation of delays in railway networks can be modeled as processes on networks. The aim of source detection is to find the starting point of such

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

a propagation process from data about the observed event counts at the network nodes. With the knowledge of the origin of a propagation process, one is able to truly combat further spreading. Additionally, the origin is the basis for the prediction of the propagation process. Therefore, source detection plays a crucial role in the problem assessment of spreading phenomena.

Thereby, modern propagation patterns are highly complex and irregular. They can be described best by processes on complex networks. Therefore, we enhance the network-based approach for source detection by Manitz et al. (2014), which has been originally developed to reconstruct the epicenter of food-borne disease outbreaks. As a many-faceted application, we chose delay propagation in railway networks. Based on a well-defined network, the application has the advantage that good models for delay propagation exist. Thus, the spreading of delays can be easily simulated and various complex diffusion patterns can be mimicked. Hence, delay propagation on railway networks is a good candidate example to test whether the network-based approach can be applied for source detection problems other than the spreading of food-borne diseases.

2 Methods and Data

2.1 Network-Based Source Detection

The approach requires an underlying network, which can be specified as a collection of nodes $k = 1, \ldots, K$, which are connected by direct links between them. When modeling food-borne infectious diseases, the underlying network represents the transportation routes of contaminated food. Here, the network is defined by a public transportation system, where nodes represent railway stations. Two nodes are connected by a link (also called edge) if there is a track between the corresponding stations, which is used by a scheduled train.

An appropriate distance $d(k, l)$ is defined for a specific path γ_{kl} between all pairs of nodes $k, l = 1, \ldots, K$ of the network. This will be specified in the next section. Furthermore, we assume a time-dependent stochastic process $\{X_k(t)\}$ on the network nodes characterizing a propagation mechanism in a time range $t = 1, \ldots, T$. Corresponding observations $x_k(t)$ in each node k are conducted at different time points $t = 1, \ldots, T$ to find sequential pictures of the distribution pattern.

The basic assumption is that propagation phenomena are spreading in a circular pattern from the correct origin k_0. The focal idea of source reconstruction is testing different source candidates and examine the concentricity of the observed pattern on a minimum shortest-path tree with the candidate k_0 as the root. Thus, given an appropriate distance definition d, the source can be reconstructed as the median of the observed pattern at time t, which is obtained by minimizing the expected distance $\mu_X(d; k_0, t)$ from the origin k_0 to all other network nodes, i.e.

$$\hat{k}_0(t) \in \arg\min_{k_0 \in \mathcal{K}_0} \mu_X(d; k_0, t), \tag{1}$$

where $\hat{k}_0(t)$ is from the set of nodes $k_0 \in \mathcal{K}_0$ for which the expected distance attains the smallest value, i.e. $\mu_X(d; \hat{k}_0, t) = \min_{k_0 \in \mathcal{K}_0} \mu_X(d; k_0, t)$.

Thereby, the expected distance $\mu_X(d; k_0, t)$ can be estimated by the average distance $d(k, k_0)$ from source k_0 to all destination nodes k weighted by the observed number of delays $x_k(t)$ in node k until time t. Thus,

$$\hat{\mu}_X(d; k_0, t) = \frac{1}{N_x(t)} \sum_{k=1}^{K} x_k(t) \cdot d(k, k_0), \tag{2}$$

where $N_x(t) = \sum_k x_k(t)$ is the total number of delays in the network at time t.

For the transformation of the irregular diffusion pattern into a typical concentric spreading circle, the replacement of the classic geographic distance by a network-based effective distance is necessary (see Brockmann and Helbing, 2013; Manitz et al., 2014).

2.2 Characteristics of the Railway Network

The public transportation network consists of $K = 319$ nodes connected by 446 links, which results in a very low link density of 0.009. Only about 1% of all possible links in a fully connected network are present (see Figure 1). The average link number to other stations is 2.8 and similar to other PTNs (e.g., von Ferber et al., 2009). The majority of the stations are stops on a line (median is 2). The degree distribution is left-skewed, so that there are a few stations of high importance with a large number of links in various directions.

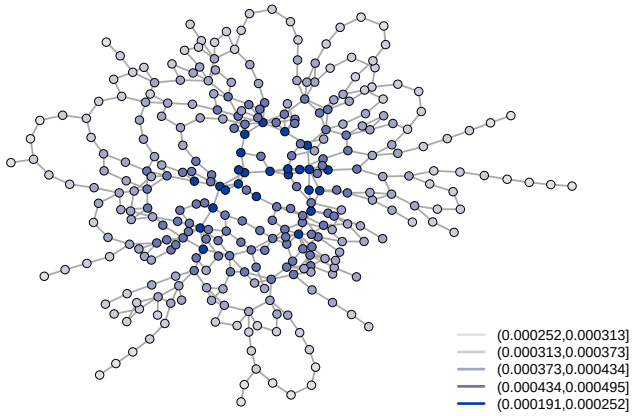


FIGURE 1. **Railway network.** Nodes are color-coded according to closeness centrality, which measures the inverse distance of a node to all other nodes in the network. Network data bases on Public Transportation Network, which is obtained from the optimization software LinTim (Goerigk et al., 2014), which is similar to the German high-speed railway network.

2.3 Train Delay Simulation

Based on a public transportation network (see Section 2.2) we compute a line concept and a timetable. We generate a set of initial delays, which model exterior influences such as weather conditions, strikes or construction work. Those initial delays are then propagated, because of dependencies between the trains due to passenger transfers or track occupation of subsequent trains. The decision which passenger transfers can be hold and the sequence of trains running along a track are made according to a prescribed delay management strategy. They allow to remove transfers from the delayed trains and to switch train orders in order to decrease the impact of delays. Using the LinTim software package (Goerigk et al., 2014) for executing delay management strategies we are able to generate diverse propagation mechanisms to mimic various interesting spreading patterns.

3 Results and Conclusions

The simulation results confirm the applicability of the source detection approach, that suggest that train delay spreading has similar underlying propagation mechanism as the transmission of infectious diseases. We observe decreasing source detection performance over time, while the influence of node centrality is moderate, if regular networks are considered. It can be also shown that our approach for source detection is extremely robust in regard to different propagation mechanisms. Furthermore, the incorporation of additional knowledge in the network definition improves the source detection performance. However, the unweighted network performs only a little worse, so that the approach can be recommended even without knowledge for link weighting. The simulation results illustrate the applicability of the method not only in the area of infectious diseases but also in the area of train delays.

Acknowledgments: This work was supported by the German Research Foundation, Research Training Group 1644 'Scaling Problems in Statistics' and the Simulation Science Center Clausthal/Göttingen.

References

Brockmann, D., and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science*, **342(6164)**, 1337–1342.

von Ferber, C., Holovatch, T., Holovatch, Yu., and Palchykov, V. (2009). Public transport networks: empirical analysis and modeling. *The European Physical Journal B*, **68(2)**, 261–275.

Goerigk, M., Harbering, J., and Schöbel, A. (2014). *LinTim - Integrated Optimization in Public Transportation.* Available online: http://lintim.math.uni-goettingen.de/; last access: 14 April 2014.

Manitz, J., Harbering, J., Schmidt, M., Kneib, T., and Schöbel, A. (2014). Network-based Source Detection for Train Delays on Railway System. *Technical Report*.

Manitz, J., Kneib, T., Schlather, M., and Brockmann, D. (2014). Origin Detection during food-borne Disease Outbreaks - A case study of the 2011 EHEC/HUS Outbreak in Germany. *PLOS Currents Outbreaks*, Edition 1.

Impact of misspecified random effect distributions on models for panel survey data

Louise Marquart[1,2], Michele Haynes[1], Peter Baker[3]

[1] Institute for Social Science Research, The University of Queensland, Australia
[2] Statistics Unit, QIMR Berghofer Medical Research Institute, Australia
[3] School of Population Health, The University of Queensland, Australia

E-mail for correspondence: l.marquart@uq.edu.au

Abstract: Mixed effects models are commonly used to analyse longitudinal data, typically under the assumption that the random effects are Gaussian distributed. However, this assumption may not always be valid. We consider the specific departure from normality characterized by multimodality. Such a scenario can arise if key categorical variables are omitted, or under the situation known as the mover-stayer scenario. In this case, a subset of the population never changes state over time, while another subset does. Research has indicated mixed evidence for the impact of misspecified random effects distribution on inference and prediction. This warrants further study, particularly in the panel survey setting, which is subject to more variability and non-response than in biomedical settings due to the collection of self-reported data. Through simulations we comprehensively investigate the effect of misspecifying random effects arising from a three component mixture of Gaussians in a logistic mixed model. We conclude with examples where misspecifying the random effects distribution has a substantial impact on interpretation.

Keywords: Generalised linear mixed models; random effects; misspecification; panel data; mixture distribution

1 Introduction

Generalised linear mixed models (GLMMs) are regularly used to analyse longitudinal data in health and social sciences. Parameters of interest are often estimated using maximum likelihood, typically under the assumption of Gaussian (or normal) distributed random effects with mean zero and covariance matrix D. However heterogeneity of the random effects may often occur in practice, particularly if a key categorical variable is omitted from the model, resulting in a random effects distribution following a finite mixture of Gaussian distributions (Verbeke and Molenberghs, 2009).

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

Another common situation leading to misspecification of the random effect distribution is the mover-stayer scenario. In the simplest case of examining transitions between two states, a subset of the population never changes states (stayers), while another subset does (movers). In this case, the random effects may not follow a Gaussian distribution as the stayers and movers will have considerably different probabilities of being in the state being modelled. To accommodate the movers and the two groups of stayers, a more flexible distribution such as a three-component mixture of Gaussian distributions could capture the inherent heterogeneity.

Evidence regarding the impact of misspecified random effect distributions on inference and prediction has been mixed (McCulloch and Neuhaus, 2011). Studies have considered assessing misspecification under a variety of true distributions including two- and three-component mixtures of normals. However, the impact of invalid assumptions about the random effects distribution has not been systematically investigated.

The confusion about the effect of misspecification has been further exacerbated by the lack of investigation of factors such as non-response and attrition prevalent in panel survey settings. This is an important area of study, as surveys are subject to more variability and non-response than in biomedical settings due to the collection of self-reported data.

We consider the specific departure from normality arising from a three component mixture of Gaussians in a logistic mixed model. By replicating the complexities inherent in panel survey data, including attrition, we comprehensively investigate the effect of misspecified random effects within the mover-stayer scenario.

2 Simulation Study

The simulated data were generated from a logistic random intercept model whereby the random intercepts arise from a three-component Gaussian mixture model representing the mover-stayer scenario.

The parameters and design matrix were motivated from the results of modelling employment outcomes in the Household and Income Labour Dynamic in Australia (HILDA) (Wooden and Watson, 2007). The generated data represents employment status (employed vs. not-employed) of a sample of 1000 women aged between 30 and 44 at baseline over a period of 11 years. The model includes explanatory variables to adjust for age (X_1), marital status (X_2, X_3), highest level of education achieved (X_4, X_5, X_6) and whether the woman has dependent children aged 14 or less (X_7). The corresponding parameter coefficients are denoted β_1 to β_7, respectively.

The random intercept b_i was simulated from a symmetric three component mixture of normal distributions with equal mixing proportions and component variances (σ_b^2),

$$b_i \sim \frac{1}{3}N(\mu_1, \sigma_b^2) + \frac{1}{3}N(\mu_2, \sigma_b^2) + \frac{1}{3}N(\mu_3, \sigma_b^2).$$

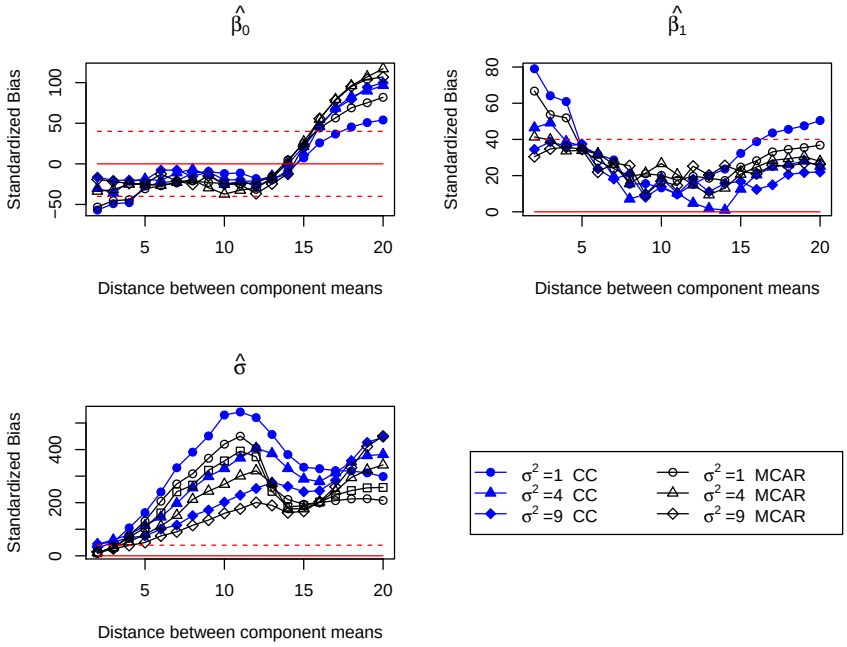


FIGURE 1. Standardized bias for selected parameter estimators of random intercept logistic model for complete case (CC) and MCAR missingness for increasing distances of random intercept component means ($\mu_3 - \mu_1$) under three component variance scenarios ($\sigma_b^2 = 1, 4, 9$). Red horizontal solid line at standardized bias=0 and red horizontal dashed lines at standardized bias ± 40 (Burton et al. (2006)).

Nineteen random effect distributions of increasing component mean distances, each with component variances $\sigma_b^2 = 1, 2, 4$ and 9, were considered. The different mean combinations for μ_1, μ_2 and μ_3 have a symmetric distribution with mean zero. We consider the case where $\mu_1 = -\mu_3$ and $\mu_2 = 0$, where μ_3 ranges from 1 to 10, increasing in increments of 0.5.
Simulations were performed under two missing data scenarios, complete data (CC) and incomplete data due to attrition. Attrition was assumed to be generated by the missing completely at random (MCAR) mechanism, with the same wave-to-wave attrition as observed in HILDA (Table 1).

TABLE 1. Wave-to-wave attrition (%) for main sample in HILDA

Wave										
1	2	3	4	5	6	7	8	9	10	11
–	13.2	9.6	8.4	5.6	5.1	5.3	4.8	3.7	3.7	3.5

For each random effect and missing data scenario (152 in total), 100 datasets were generated and a random intercept logistic model assuming Gaussian random effects was fitted to each dataset. Performance measures such as standardized bias, mean square error (MSE) and coverage were used to

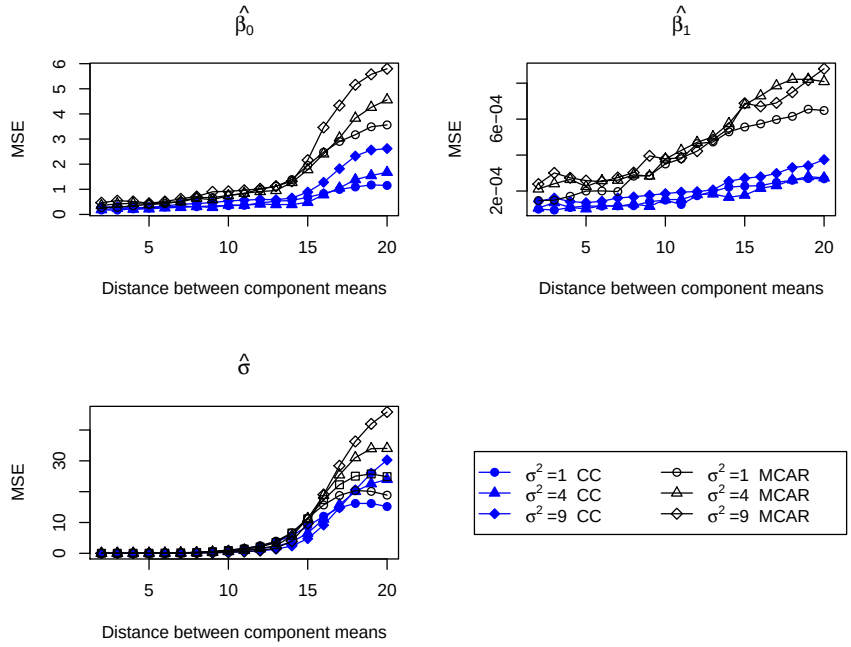


FIGURE 2. Mean square error (MSE) for selected parameter estimators of random intercept logistic model for complete case (CC) and MCAR missingness for increasing distances of random intercept component means ($\mu_3 - \mu_1$) under three component variance scenarios ($\sigma_b^2 = 1, 4, 9$).

assess the sensitivity of the normality assumption on estimating model parameters under random effects distribution misspecification. Criteria for acceptable performance were standardized bias within the $\pm$ 40 threshold and coverage rates within 91% and 99% (Burton et al. (2006)). For the performance measures relating to the random intercept distribution, the variance estimate of the random intercept was compared to the overall variance of the three component mixture distribution.

3 Results and Discussion

Figure 1 presents examples of standardized bias for the intercept (β_0), age coefficient (β_1) and the random intercept standard deviation estimate (σ) for component variances $\sigma_b^2 = 1, 4, 9$ and two missing data scenarios (CC and MCAR). With increasing distance of the random intercept component means, defined as $\mu_3 - \mu_1$, the standardized bias of σ exceed the threshold of ± 40 when component mean distance ≥ 4. For the CC scenario, β_0 exceeded 40 when component mean distance ≥ 17, and β_1 exceeded 40 for component mean distances ≤ 3 for $\sigma_b^2 \leq 4$. Results were similar for the MCAR missing data scenario, with larger component variances resulting in more bias for β_0. Results for the fixed effects coefficients $\beta_4, \beta_5, \beta_6$ had similar trends to β_0, whilst β_2, β_3 and β_7 showed no substantial bias.

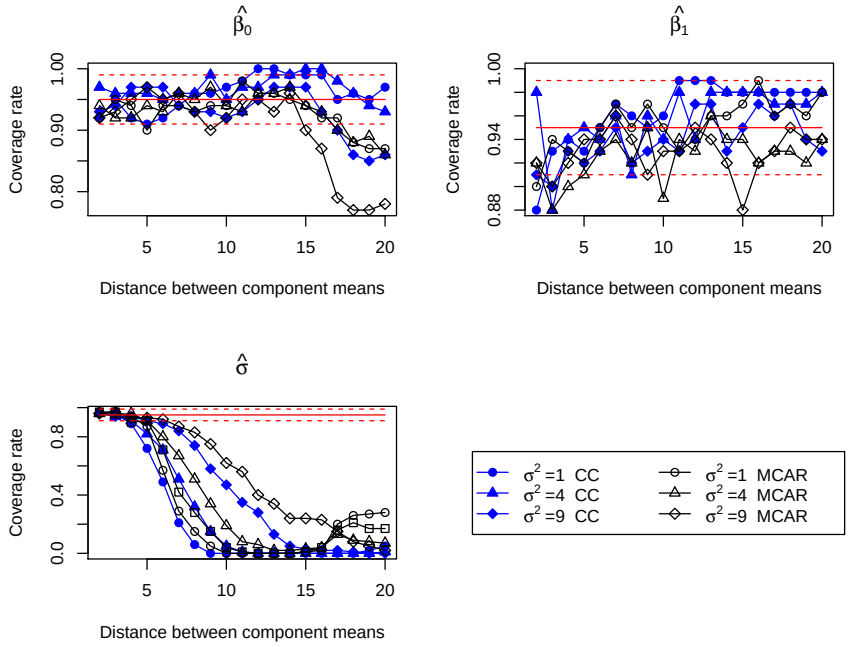


FIGURE 3. Coverage rates for model based 95% confidence intervals for selected parameter estimators of random intercept logistic model for complete case (CC) and MCAR missingness for increasing distances of random intercept component means ($\mu_3 - \mu_1$) under three component variance scenarios ($\sigma_b^2 = 1, 4, 9$). Red horizontal solid line at nominal coverage rate= 0.95 and red horizontal dashed lines at coverage rate=0.91 and 0.99 (Burton et al. (2006)).

The MSE increased as the distance between component means increased, with large increases occurring for β_0 and σ when component mean distances > 15 (Figure 2). The MSE for the MCAR missing data scenario was larger than the complete case for all coefficients and all variance scenarios. MSE was larger for increasing component variances. The MSE for the fixed effects coefficients β_2 to β_7 showed similar patterns to β_1.

For the complete case scenario coverage rates were close to the nominal rate of 95% for all parameter estimates with the exception σ^2 (Figure 3). For all component variances, the nominal coverage rates for σ were $\leq 91\%$ when component mean distance > 5. Nominal coverage rates for the MCAR data scenario were slightly smaller except for σ. The coverage rates for the fixed effects coefficients β_2 to β_7 showed similar patterns to β_1, with coverage rates generally within 91% and 99%. The standardized bias, MSE and coverage results for $\sigma_b^2 = 2$ were similar to $\sigma_b^2 = 1$.

The simulation study suggests that incorrectly assuming Gaussian random effects when the true distribution arises from a three component mixture of Gaussian distributions can impact on inference. Misspecification of the random effects distribution can result in seriously biased random effect standard deviation estimates and substantially low coverage rates. The impact of misspecification on fixed effect coefficients is minimal with bias experi-

enced for some coefficients including β_0 when random intercepts have large component means distances, or for small component mean distances such as β_1. The impact of attrition assuming MCAR missingness had little impact on bias and coverage, though it resulted in larger MSE than the complete case. Future work will need to consider attrition generated by other missing data mechanisms and random effects generated from asymmetric mixture distributions.

In conclusion, this study demonstrates that misspecification of the random effects distribution in logistic mixed models within the panel survey setting can seriously impact estimates of the random effect variance, with attrition having minimal additional impact.

Acknowledgments: The authors thank Dr. Emma Huang and Prof. Peter O'Rourke for providing valuable comments. This paper uses unit record data from the Household, Income and Labour Dynamics in Australia survey (HILDA). The HILDA project was initiated and is funded by the Australian Government Department of Families, Housing, Community, Services and Indigenous Affairs (FaHCSIA) and is managed by the Melbourne Institute of Applied Economic and Social Research (Melbourne Institute). The findings and views reported in this paper, however, are those of the authors and should not be attributed to either FaHCSIA or the Melbourne Institute.

References

Burton, A., Altman, D.G., Royston, P. and Holder, R.L. (2006). The design of simulation studies in medical statistics *Statistics in Medicine*, **25**, 4279 – 4292.

McCulloch, C.E. and Neuhaus, J.M. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter *Statistical Science*, **26**, 388 – 402.

Verbeke, G. and Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wooden, M. and Watson, N. (2007). The HILDA survey and its contribution to economic and social research (so far) *The Economic Record*, **83**, 208 – 231.

Flexible estimation of spatio-temporal interaction in a point process model for infectious disease spread

Sebastian Meyer[1], Leonhard Held[1]

[1] University of Zurich, Switzerland

E-mail for correspondence: Sebastian.Meyer@ifspm.uzh.ch

Abstract: Case reports from infectious disease surveillance with registered location and time of infection allow for spatio-temporal point process models of infectious disease spread. An endemic component describes the baseline risk of infection driven by population density as well as temporal and exogenous effects. A second, epidemic component captures interaction between cases and includes covariate effects on the force of infection. Here we investigate nonparametric estimation of spatial as well as temporal interaction using B-splines. Such flexible formulations disclose the distance-decay and time-course of infectivity in a more data-driven manner than previously used parametric models.

Keywords: infectious disease surveillance; self-exciting spatio-temporal point process with immigration; conditional intensity model; interaction function.

1 Introduction

Infectious disease surveillance aims at the timely detection of outbreaks as well as their prevention and control. Public health authorities routinely collect data on the occurrence of communicable diseases, registering location and date of infection, and the specific pathogen involved. The case reports often contain patient characteristics which are potentially associated with individual infectivity, e.g. the patient's age, and are ideally supplemented by lattice data on environmental and socio-demographic factors for spatial regression. Given such surveillance data, spatio-temporal point process models are a useful tool to estimate the role of individual characteristics and exogenous factors in shaping disease spread.

Following Held et al. (2005), disease risk is decomposed additively into two components: An *endemic* component describes the baseline risk driven by population density as well as temporal and exogenous effects (e.g., seasonality, population structure, prevalence of correlated diseases), whereas an

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

observation-driven *epidemic* component invokes explicit dependence between cases. Meyer et al. (2012) proposed a spatio-temporal point process model with such components, and applied it to 635 cases of invasive meningococcal disease (IMD) caused by the two most common meningococcal finetypes in Germany,2002–2008. They identified a time trend with seasonal pattern, and no evidence for an additional (lagged) association with local waves of influenza. The epidemic component revealed that the meningococcus of serogroup B was approximately twice as infectious as the C-type. However, spatial and temporal interaction of cases were modelled rather naively by assuming a *Gaussian* distance decay of *time-constant* infectivity. The former was improved by Meyer and Held (2014), who found power laws for spatial interaction to outperform previous formulations in two modelling frameworks for infectious disease surveillance data. In this paper, we investigate the use of B-splines to estimate interaction in a more flexible way.

2 Self-exciting spatio-temporal point process model

The spatio-temporal point process model proposed by Meyer et al. (2012) is designed for time-space-mark data $\{(t_i, \mathbf{s}_i, \mathbf{m}_i) : i = 1, \ldots, n\}$ of dependent events such as case reports of infectious diseases. It is defined through the conditional intensity function

$$\lambda(t, \mathbf{s}) = \nu_{[t][\mathbf{s}]}\, \rho_{[t][\mathbf{s}]} + \sum_{j:t_j<t} \eta_j \cdot g(t - t_j) \cdot f(\|\mathbf{s} - \mathbf{s}_j\|) \tag{1}$$

in a region $\mathbf{W} \ni \mathbf{s}$ during a period $(0, T] \ni t$. The first, endemic component consists of a log-linear predictor $\log(\nu_{[t][\mathbf{s}]}) = \beta_0 + \beta^{\mathrm{T}}\mathbf{z}_{[t][\mathbf{s}]}$ with a multiplicative offset $\rho_{[t][\mathbf{s}]}$, typically the population density. Both the offset and exogenous covariates in $\nu_{[t][\mathbf{s}]}$ are given on a spatio-temporal grid, hence the notation $[t][\mathbf{s}]$ for the period containing t in the region covering $\mathbf{s}$. Note that such a piecewise constant endemic model is equivalent to a Poisson regression model for the aggregated number of cases on the given grid.
However, with the second, epidemic component the intensity process depends on previously infected individuals and becomes "self-exciting". The epidemic force of infection at $(t, \mathbf{s})$ is the superposition of the infection pressures caused by previously infected individuals. The log-linear predictor $\log(\eta_j) = \gamma_0 + \gamma^{\mathrm{T}}\mathbf{m}_j$ weights infectivity by individual/infection-specific characteristics $\mathbf{m}_j$. Regional-level covariates from the endemic grid can also be included in η_j, e.g., to model ecological effects on infectivity.
Decreasing infection pressure over space is described by $f(x)$ as a function of the spatial distance from the infectious source. Meyer et al. (2012) originally used the Gaussian kernel $f_{\mathrm{G}}(x) = \exp\left\{-x^2/(2\sigma^2)\right\}$ as a standard choice. Subsequently, Meyer and Held (2014) showed that the power law

$$f_{\mathrm{PL}}(x) = (x + \sigma)^{-d}\,, \tag{2}$$

$\sigma, d > 0$, is more appropriate in describing distance decay of infectivity, which seems to translate from the power-law feature of human travel (Brockmann et al., 2006). For the time course of infectivity, $g(t)$, both works simply assumed a constant function over a fixed period.
However, the basic model framework (1) actually allows for arbitrary shapes with the only requirements that the interaction functions are differentiable with respect to their parameters, and that $g(t)$ and $f_2(\mathbf{s}) = f(\|\mathbf{s}\|)$ are integrable over $(0, T]$ and $\mathbf{W} - \mathbf{s}_j$, respectively. In what follows, we investigate flexible estimates of spatial and temporal interaction for the IMD data, retaining the endemic component and η_j from the previous analyses.

3 Spatial interaction

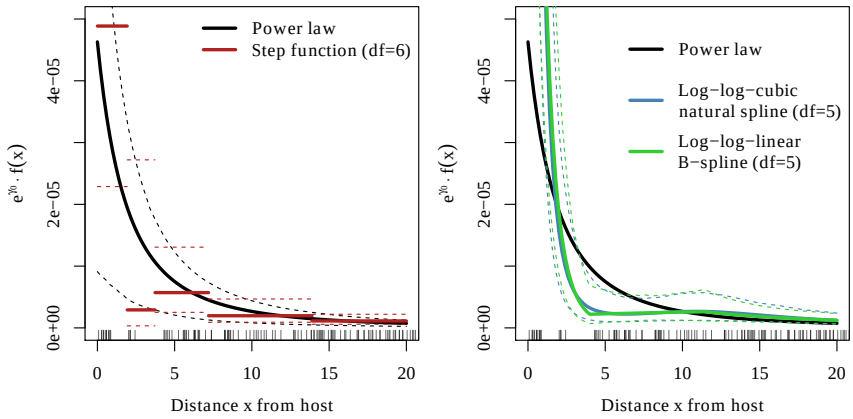


FIGURE 1. Flexible estimates of spatial interaction vs. the power law. Dashed lines represent 95% confidence intervals and the bottom "rug" shows the observed distances of events to their potential sources, i.e., to events of the past 30 days.

The left part of Figure 1 shows results from Meyer and Held (2014) with fixed $g(t) = \mathbb{I}_{(0,30]}(t)$. The estimated power law features a pronounced initial decay of infectivity as well as a heavy tail capturing occasional transmissions over large distances. The step function was estimated with six log-equidistant knots up to an upper-bound range of 100 kilometres. It suggests an even sharper initial drop but forfeits monotonicity.
A more sophisticated approach of flexible estimation is a log-log B-spline

$$f_{\mathrm{B}}(x) = \exp\left\{\sum_{k=1}^{K} \alpha_k B_k(\log(x+\sigma))\right\}, \tag{3}$$

where the B_k form a set of suitable basis functions (Fahrmeir et al., 2013, Section 8.1). The log-log formulation is motivated by the fact that the power law (2) turns into a simple linear relation on that scale:

$$\log(e^{\gamma_0} \cdot f_{\mathrm{PL}}(x)) = \gamma_0 - d \cdot \log(x+\sigma)\,.$$

This is also why linear basis functions might be sufficiently flexible, although resulting in non-differentiable joints. The right plot of Figure 1 shows estimates based on a linear and a natural cubic B-spline, respectively, each with 5 degrees of freedom. We fixed σ at the estimate from the power-law model ($\sigma = 4.60$), and the inner knots were again chosen to be equidistant on the log-scale. The spline fits are very similar and in accordance with the step function in suggesting an even steeper initial decay than the power law. With respect to AIC they perform substantially better: $\Delta\text{AIC} = -23.7$ for the cubic, and -22.2 for the linear variant, respectively. Note, however, that computational cost of the B-spline models is more than 10-fold compared to the power law. We have to evaluate basis functions and, most notably, we cannot simplify the spatial integrals $\int_{\mathbf{W}-\mathbf{s}_j} f(\|\mathbf{s}\|)\,\mathrm{d}\mathbf{s}$ in the likelihood to a one-dimensional quadrature problem (cf. Meyer and Held, 2014, Supplement B), but have to rely on product Gauss cubature.

4 Temporal interaction

Temporal interaction $g(t)$ has not been estimated in previous models for the IMD data, but assumed constant for 30 days from infection, vanishing to zero afterwards. This reluctance is mainly due to the sparseness of cases: on average, there are only 4 and 3.6 infections with types B and C, respectively, per month. For illustration, we estimate some alternative temporal interaction functions $g(t)$ while sticking to the upperbound length of 30 days for the infectious period and employing the spatial power law $f_{\text{PL}}(x)$.

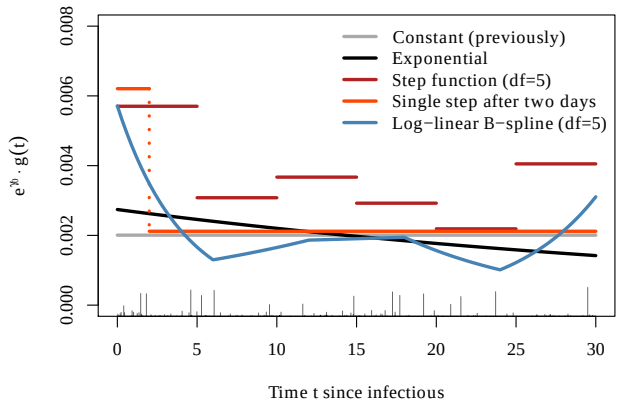


FIGURE 2. Point estimates of various models for temporal interaction. The bottom "rug" shows the observed time lags between events, where the size corresponds to the associated spatial interaction given by the estimated power law from Figure 1.

A simple parametric model for the time course of infectivity is exponential decay $g_{\text{E}}(t) = e^{-\alpha t}$, $\alpha > 0$. The B-spline formulation (3) is also applicable

for $g(t)$ but using plain rather than log-scale. Figure 2 shows estimates of temporal interaction assuming either exponential decay, or B-splines of degrees 0 (step function) or 1 with equidistant knots, or a simplified step function with a single knot after 2 days. Note that the overall level varies slightly between the alternatives since a change in $g(t)$ also affects the estimation of $f_{\mathrm{PL}}(x)$ and γ_0. One would expect the more flexible estimates to approach zero for larger time lags. However, considering late infections close to previously infective sites tends to improve the likelihood, since the endemic component is only constant within districts. It is thus necessary to determine a reasonable range of temporal interaction by other means, e.g., epidemiological considerations.
Concerning model performance, only the simplified one-step function improves upon the previous constant model ($\Delta\mathrm{AIC} = -6.4$). It suggests a partial drop of infectivity already after two days, which might correspond to quarantine actions taken after the appearance of symptoms.

5 Conclusion

We have shown that flexible B-spline formulations of interaction can be incorporated into a spatio-temporal point process model for infectious disease surveillance data. They may deliver additional insight into the spatial dependence structure and the time course of infectivity. It is crucial that the spatio-temporal resolution of the surveillance data is high enough to allow for flexible estimation of interaction. The IMD data set used for illustration is rather small and carries little information at small distances, which is why results should be regarded with caution.
As a common drawback, the regression splines depend on the chosen (boundary) knots and require much more computation time, especially in the spatial domain. A compromise are 0-degree B-splines, which don't require numerical integration and may serve as a quick initial benchmark for spatial and temporal interaction.
The application of this and two related model frameworks in R is described in detail in Meyer et al. (2014).
Software: All calculations have been carried out in the statistical software environment R 3.0.3. The point process model (1) is implemented in the R package **surveillance** as function `twinstim()`, and a simplified version of the IMD data is included as `data("imdepi")` (courtesy of the German Reference Centre for Meningococci). Spatial integrals in the likelihood have been evaluated using cubature methods from the R package **polyCub** (see Meyer and Held, 2014, Supplement B). The implementation also allows for other specifications of the interaction functions f and g, respectively.

Acknowledgments: The research is financially supported by the Swiss National Science Foundation (project 137919: *Statistical methods for spatio-temporal modelling and prediction of infectious diseases*).

References

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, **439**, 462 – 465.

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

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, **5**, 187 – 199.

Meyer, S., Elias, J., and Höhle, M. (2012). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, **68**, 607 – 616.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *Annals of Applied Statistics*. Accepted and available as `arXiv:1308.5115`.

Meyer, S., Held, L., and Höhle, M. (2014). Spatio-temporal modeling of epidemic phenomena using the R package **surveillance**. In preparation for the *Journal of Statistical Software*.

Random Forests for Functional Covariates

Annette Möller[1], Gerhard Tutz[2], Jan Gertheiss[13]

[1] Department of Animal Sciences, Georg-August-Universität Göttingen, Germany
[2] Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany
[3] Center for Statistics, Georg-August-Universität Göttingen, Germany

E-mail for correspondence: annette.moeller@agr.uni-goettingen.de

Abstract: We propose a random forest approach for functional covariates. The method is based on partitioning the functions' domain in intervals and to use the functions' mean values across those intervals as predictors in regression (or classification) trees. This approach appears to be more intuitive to applied researchers than usual methods for functional data, while also performing very well in terms of prediction accuracy. We apply our method to Raman spectra of boar meat.

Keywords: Functional Data; Functional Linear Model; Random Forests; Regression Trees; Spectroscopy.

1 Introduction

Functional data occurs frequently in various fields of applications. Many statistical methods have their functional counterparts tailored to the specific properties of such data, see for example Ramsay and Silverman (2005). Many applied researchers, however, are not familiar with the methods of functional data analysis and rely on more intuitive procedures. One of these more intuitive approaches is to discretize a signal $x(t)$, $t \in J \subset \mathbb{R}$, by partitioning its domain into several intervals and employ the mean values computed from each interval for example as covariates in a regression or classification model. Based on this idea of computing mean values over intervals, we propose a special form of random forests (Breiman, 2001) to analyze functional data. The covariates used for the single trees are the mean values over intervals partitioning the functional curves. The intervals are generated at random using exponentially distributed waiting times.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

2 Functional Random Forests

We consider data of the form $(y_i, x_i(t))$, $i = 1, \ldots, n$, with predictor curve $x_i(t)$, and y_i being the response value. In a first step, a training set is generated from the original data set by drawing, with replacement, a random sample $(y_j, x_j(t))$, $j = 1, \ldots, n$, of the same size n as the original data set. We then successively draw random numbers r_λ from an $\text{Exp}(\lambda)$ distribution, where we choose a fixed value for λ. A number r_λ represents the waiting time (on the "time" scale t of the curves $x_i(t)$) from the end of the previous interval to the beginning of the next interval. For a set of random numbers $r_\lambda^1, \ldots, r_\lambda^I$ (where I denotes the number of waiting times needed to cover the observed curve) independently drawn from an $\text{Exp}(\lambda)$ with a certain fixed λ, we obtain a pattern of intervals. The value of λ determines whether the functions' domain tends to be split into just a few (small λ) or many (large λ) intervals. Using this partition of the functions' domain, the mean values over the respective intervals are computed for each curve $x_i(t)$ form the training set and used as predictors in a regression (or classification) tree; for details on trees, see, e.g., Breiman et al. (1984). The partition pattern obtained is stored for future use when predicting new data. The procedure of data sampling, interval, predictor and tree construction is repeated m times and the resulting trees are used to construct a (random) forest. Given a regression problem, for instance, the respective functional random forest (FRF) predictions $\hat{y}_i^{FRF}$, $i = 1, \ldots, n$, are constructed by averaging the set of single tree predictions $\hat{y}_{i,1}, \ldots, \hat{y}_{i,m}$ by

$$\hat{y}_i^{FRF} = \frac{1}{m} \sum_{l=1}^{m} \hat{y}_{i,l}. \tag{1}$$

In a classification problem, the class of a new observation would be predicted as a majority vote among the trees forming the forest.

3 Predicting Skatole Concentration using Raman Spectra

We consider data from Raman spectroscopy on samples of boar meat. Raman spectroscopy is a technique based on the inelastic scattering of monochromatic light, e.g. laser light, on molecules (Raman scattering). The resulting spectrum gives information about vibrational modes in the analyzed system. This vibrational information provides a fingerprint by which molecules (and therefore substances present in the sample) can be identified.

Our Raman data is functional in nature, coming as curves $x_i(t)$ of intensity at wavelength t for meat sample i, with $t \in [332\text{nm}, 2105\text{nm}]$. The scalar response y_i of interest is the skatole content in parts per billion (ng/g) in the respective meat sample i, $i = 1, \ldots, n$. Skatole is a white crystalline organic compound present in the meat of boars that have not been castrated. If present in high concentration it has a strong off odor. The goal

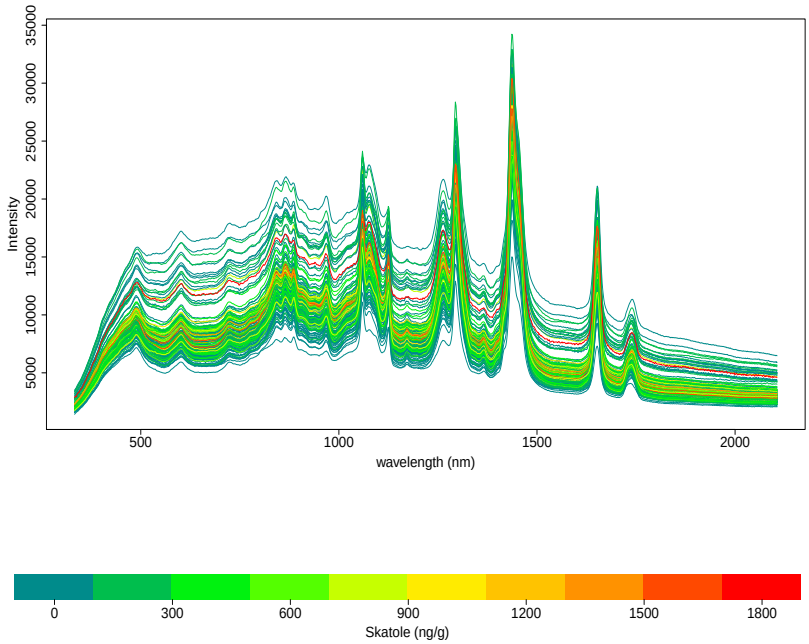


FIGURE 1. Mean curves of intensity measurements, with colors assigned according to the skatole content level.

of our analysis is predicting the skatole content from the spectra $x_i(t)$. If an accurate prediction of the skatole content is possible, the Raman spectroscopy can be utilized as an alternative to chemical analysis to determine the skatole content of meat.

The total number of meat samples is $n = 148$. For each of these samples, the Raman spectroscopy was repeated ten times on an inner layer of the meat sample. The response variable y_i, the skatole content, was determined for each meat sample i by chemical analysis. To obtain a single Raman spectra $x_i(t)$ for each sample i, we take mean values across the ten replicates. Figure 1 displays the spectra $x_i(t)$, with colors corresponding to the skatole content given by the response observations y_i, $i = 1, \ldots, 148$.

From the color distribution of the curves it becomes visible that there are a few curves corresponding to extremely high skatole content levels (more than 1000 parts per billion), while the majority of the curves correspond to lower levels. These extreme skatole content levels are hard to predict, especially when the training data mainly consists of moderate observations. Furthermore, by visual inspection, no direct relationship between the curves (measured intensity at the different wavelengths) and the skatole content is found. Therefore, prediction of the skatole content is expected to be a difficult task.

For our analysis we use $n_1 = 100$ meat samples as training set and the remaining $n_2 = 48$ as test set for prediction, with $n_1 + n_2 = n$. We per-

form the assignment of training and test cases by randomly drawing n_1 cases (without replacement), employ it as training set, and use the remaining n_2 cases as test set. Then the proposed random forest for functional data is used for predicting the test observations. As competing methods, we consider a functional linear model and a random forest using single measurements $x(t_j)$ as predictors, with t_j, $j = 1, \ldots, 1024$, being the discrete wavelengths at which intensity was actually measured. For fitting the linear model we use the method proposed by Goldsmith et al. (2012) and implemented in `R` package `refund` (Crainiceanu et al., 2013); for the non-functional random forest, we use the `R` package `randomForest` by Liaw and Wiener (2002). When constructing the functional random forest we consider different choices of λ (0.01, 0.03, 0.05, 0.2, 0.5) reflecting long, medium and short waiting times (where $\lambda = 0.5$ represents extremely short waiting times and results in a procedure very close to the non-functional random forests), and $m = 1000$ in each case. For the non-functional random forests, the number of grown trees is chosen as $m = 1000$ as well. The generation of training and test data, model fitting and test set prediction is repeated 20 times.

TABLE 1. Average MAE and RMSE values for the considered models, average taken over all values of the 20 repeated runs.

	MAE	RMSE
Functional Linear Model	253.210	361.356
Functional RF $\lambda = 0.01$	244.380	354.375
Functional RF $\lambda = 0.03$	242.495	352.205
Functional RF $\lambda = 0.05$	242.716	353.380
Functional RF $\lambda = 0.2$	242.181	354.942
Functional RF $\lambda = 0.5$	242.317	355.790
Non-functional RF	253.541	364.283

Table 1 shows the average over the 20 mean absolute error and root mean square error values of prediction for the different methods. We see that, both in terms of the MAE and the RMSE, the proposed functional random forest outperforms the functional linear model and the non-functional random forest which simply uses the measurements at $t_1, t_2, \ldots$ as predictors. The performance of the latter is even slightly inferior to the functional linear model. The predictive performance of the functional random forest, however, slightly depends on the chosen λ determining the average number and length of the employed intervals. For larger values of λ the functional random forests approximate to the non-functional version. For the largest considered value $\lambda = 0.5$ the predictive performance starts to deteriorate. This value corresponds to average interval lengths of 2 units on the scale of t, which is close to the procedure used for the non-functional random forest. On the contrary, too large interval lengths as generated by a small value such as $\lambda = 0.01$ yield a deterioration in predictive performance as well.

Obviously larger interval lengths are less able to account for the variability within each of the curves. Therefore, the optimal λ for a specific data set should be obtained in a data-driven way, for example, via a cross validation procedure. We will investigate such a procedure in additional case studies. Further analyzes may also consider spectra obtained from an outer layer of the meat, or a model utilizing both, the outer and the inner layer as covariates.

The functional and non-functional random forests employed for the current analysis are based on regression trees predicting the skatole content as a continuous response. However, our proposed functional random forests (and the non-functional random forests) can be applied with classification trees as well, as sketched in Section 2. As it is also of high interest to predict whether a sample of meat has a skatole content below or above a given threshold, we will analyze the predictive performance of our functional random forests with classification trees for a binary response (skatole content above/below threshold) in further case studies. As there are not many functional classification methods available for multi-categorical outcomes, further development of the proposed functional random forests in that direction is promising, too.

Acknowledgments: We thank Daniel Mörlein from the Department of Animal Sciences, Georg-August-University Göttingen, for providing the data used in Section 3 and related information.

References

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

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. London: Chapman & Hall.

Crainiceanu, C.M., Reiss, P., Goldsmith, J., Huang, L., Huo, L., and Scheipl, F. (2013). *refund: Regression with Functional Data*. R package version 0.1-9.

Goldsmith, J., Bobb, J., Crainiceanu, C.M., Caffo, B., and Reich, D. (2011). Penalized Functional Regression. *Journal of Computational and Graphical Statistics*, **20**, 830 – 851.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, **2**, 18 – 22.

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

Therneau, T., Atkinson, B., and Ripley, B. (2013). *rpart: Recursive Partitioning*. R package version 4.1-3.

Efficient Bayesian inference for Copula Gaussian graphical models

Abdolreza Mohammadi[1], Fentaw Abegaz[1], Ernst Wit[1]

[1] Dept. of Statistics and Probability, University of Groningen, Netherlands

E-mail for correspondence: a.mohammadi@rug.nl

Abstract: One of the main issue in science is to discover complicated interaction patterns between variables in multivariate data of various types. Copula Gaussian graphical models is one potential way to discover the underlying conditional independence of variables in such mixed data. In this paper, we proposed a comprehensive Bayesian approach in copula Gaussian graphical models that accomplished the diverse types of data, including binary, ordinal and continuous. We embed a graph selection procedure inside a semiparametric Gaussian copula. We carry out the posterior inference by using an efficient sampling scheme which is a trans-dimensional MCMC approach based on the continuous-time birth-death process. The proposed method is tested in real and simulated examples. We implemented the method as a general purpose in the R package `BDgraph` which is freely available online.

Keywords: Copula Gaussian graphical models; Bayesian model selection; Latent variable models; Birth-death process; Markov chain Monte Carlo.

1 Introduction

Graphical models provide an effective way to describe statistical patterns in data, specially for high-dimensional datasets such as gene expression data. In this context undirected Gaussian graphical models are commonly used, since inference in such models is often tractable. In undirected Gaussian graphical models, the graph structure is characterized by its precision matrix(the inverse of covariance matrix): the non-zero entries in the precision matrix show the edges in the graph.
In the real world, however data are often non-Gaussian or discrete. For non-Gaussian continuous data, variables can be transformed to Gaussian latent variables. Then a graph structure is inferred for the Gaussian variables. For discrete data, however, the situation is more convoluted; we can not transform them directly into latent Gaussian variables, since the mapping

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

is one-to-many. A common approach is to apply a Markov chain Monte Carlo method (MCMC) to simulate both the latent Gaussian variables and the posterior distributions (Hoff, 2007, Dobra and Lenkoski, 2011, Liu et al. 2009, and Pitt et al. 2006).
In this paper, we propose an efficient Bayesian framework in Copula Gaussian graphical models that can be contributed for binary, ordinal or continuous variables simultaneously. We embed graphical model selection inside a semiparametric Gaussian copula. For our copula framework we use the extended rank likelihood (Hoff, 2007). We carry out the posterior inference for the graph and the precision matrix by using an efficient sampling scheme which is a trans-dimensional MCMC approach based on the continuous-time birth-death process (Mohammadi and Wit, 2013).

2 Copula Gaussian graphical models

Let $X = (X_1, \ldots, X_p)$ be a p-dimensional random vector following a multivariate normal distribution $\mathcal{N}_p(0, K^{-1})$ with precision matrix K. A Gaussian graph model for the random vector X is represented by an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of p vertices and E is the edge set. Zero entries in the precision matrix correspond to the absence of edges on the graph and conditional independence between pairs of random variables given all other variables.
In practice, we encounter both discrete and continuous variables and copula Gaussian graphical modeling has been proposed to describe dependencies between such heterogeneous variables. Let X be a collection of continuous, binary, ordinal or count variables with F_j the marginal distribution of X_j and F_j^{-1} its pseudo inverse. Towards constructing a joint distribution of X, we introduce a multivariate normal latent variable $Z \sim \mathcal{N}(0, \Gamma(K))$, where $\Gamma(K)$ is the correlation matrix for a given precision matrix K. The joint distribution of X is given by

$$P\left(X_1 \leq x_1, \ldots, X_p \leq x_p\right) = C(F_1(x_1), \ldots, F_p(x_p) \mid \Gamma(K)), \tag{1}$$

where $C(\cdot)$ is the Gaussian copula given by

$$C(u_1, \ldots, u_p \mid \Gamma) = \Phi_p\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p) \mid \Gamma\right),$$

with $u_v = F_v(x_v)$, and $\Phi_p(\cdot)$ is the cumulative distribution of multivariate normal and $\Phi(\cdot)$ is the cumulative distribution of univariate normal distributions. It follows that $X_v = F_v^{-1}\left(\Phi(Z_v)\right)$.
In semiparametric copula estimation, the marginals are treated as nuisance parameters and estimated by the rescaled empirical distribution that results the joint distribution in (1) to be parametrized only by the correlation matrix of the Gaussian copula. Our aim is to infer the underlying graph structure G of the observed variables X implied by the continuous latent variables Z. Since Zs are unobservable we follow the idea of (Hoff, 2007) that relate them to the observed data as follows. Given the observed data

$\boldsymbol{x}$ from a sample of n observations, the latent samples $\boldsymbol{z}$ are constrained to belong to the set

$$A(\boldsymbol{x}) = \{\boldsymbol{z} \in \Re^{n \times p} : L_j^r(\boldsymbol{z}) < z_j^{(r)} < U_j^r(\boldsymbol{z}), r = 1, \ldots, n; j = 1, \ldots, p\}. \tag{2}$$

where

$$L_j^r(\boldsymbol{z}) = \max\left\{z_j^{(k)} : x_j^{(s)} < x_j^{(r)}\right\} \text{ and } U_j^r(\boldsymbol{z}) = \min\left\{z_j^{(s)} : x_j^{(r)} < x_j^{(s)}\right\}$$

. Further (Hoff, 2007) suggested that inference on the latent space can be performed by substituting the observed data $\boldsymbol{x}$ with the event $\mathcal{D} = \{\boldsymbol{z} \in A(\boldsymbol{x})\}$ and defined the likelihood as:

$$P(\boldsymbol{x} \mid K, F_v : v \in V) = P(\mathcal{D} \mid K)\ P(\boldsymbol{x} \mid \mathcal{D}, K, F_v : v \in V). \tag{3}$$

The only part of the observed data likelihood relevant for inference on K is $P(\mathcal{D} \mid K)$. Thus, the likelihood function is given by

$$P(\mathcal{D} \mid K) = P(\boldsymbol{z} \in A(\boldsymbol{x}) \mid K) = \int_{A(\boldsymbol{x})} P(\boldsymbol{z} \mid K) d\boldsymbol{z} \tag{4}$$

where

$$P(\boldsymbol{z} \mid K) \propto (\det(K))^{n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(KU)\right\}, \tag{5}$$

with $U = \boldsymbol{z}'\boldsymbol{z}$.

3 Bayesian Copula Gaussian graphical modelling

Consider the joint posterior distribution of $K \in P_G$ and the graph G given by

$$P(K, G \mid \mathcal{D}) \propto P(\mathcal{D} \mid K)\ P(K \mid G)\ P(G). \tag{6}$$

Sampling from this joint posterior distribution can be done by a computationally efficient birth-death MCMC sampler proposed in Mohammadi and Wit (2013) for Gaussian graphical models.

For the prior distribution of the graph, as a non-informative prior, we propose a discrete uniform distribution over the graph space, $P(G) \propto 1$. For other choice of priors see Mohammadi and Wit (2013).

For the prior distribution of the precision matrix, we use the G-Wishart which is attractive since it is conjugate for normally distributed data and places no probability mass on zero entries of the precision matrix. A zero-constrained random matrix $K \in \mathbb{P}_G$ has the G-Wishart distribution $W_G(b, D)$, if

$$P(K|G) = \frac{1}{I_G(b, D)}|K|^{(b-2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(DK)\right\},$$

where $b > 2$ is the degree of freedom, D is a symmetric positive definite matrix, and $I_G(b, D)$ is the normalizing constant,

$$I_G(b, D) = \int_{\mathbb{P}_G} |K|^{(b-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(DK)\right\} dK.$$

The G-Wishart prior is conjugate to the likelihood (5), hence, the posterior distribution of K is

$$P(K|\mathcal{D}, G) = \frac{1}{I_G(b^*, D^*)} |K|^{(b^*-2)/2} \exp\left\{-\frac{1}{2}\text{tr}(D^*K)\right\},$$

where $b^* = b + n$ and $D^* = D + S$, that is, $W_G(b^*, D^*)$.

3.1 The proposed birth-death MCMC algorithm

We describe here a MCMC sampler scheme for the joint posterior distribution which is proposed by Mohammadi and Wit (2013). Here we extend their algorithm for more general case of Copula Gaussian graphical models. Our algorithm is based on a continuous time birth-death Markov process in which the algorithm explores over the graph space by adding or removing an edge in a birth or death event. The birth and death rates of edges occur in continuous time with the rates determined by the stationary distribution of the process. The algorithm is considered in such a way that the stationary distribution equals the target joint posterior distribution of the graph and the precision matrix (6).

The birth and death processes are independent Poisson processes. Thus, the time between two successive events is exponentially distributed, with mean $1/(\beta(K)+\delta(K))$. Therefore, the probability of birth and death events are proportional to their rates.

Suppose that we consider the birth and death rates as

$$\beta_e(K) = \frac{P(G^{+e}, K^{+e} \setminus (k_{ij}, k_{jj})|\mathcal{D})}{P(G, K \setminus k_{jj}|\mathcal{D})}, \quad \text{for each} \quad e \in \overline{E}, \tag{7}$$

$$\delta_e(K) = \frac{P(G^{-e}, K^{-e} \setminus k_{jj}|\mathcal{D})}{P(G, K \setminus (k_{ij}, k_{jj})|\mathcal{D})}, \qquad \text{for each} \quad e \in E. \tag{8}$$

It can be shown that based on above birth and death rates, the algorithm converge to the target joint posterior distribution of the graph and the precision matrix (6). The extended birth-death MCMC algorithm for Copula Gaussian graphical models are summarized as follows.

Algorithm 3.1. Given a graph $G = (V, E)$ with a precision matrix K, iterate the following steps:

Step 1. Sample the latent data. For each $r \in V$ and $j \in \{1, 2, ..., n\}$, we update the latent value $z_r^{(j)}$ from its full conditional distribution

$$Z_r | Z_{V\setminus\{r\}} = z_{V\setminus\{r\}}^{(j)} \sim N(-\sum_{r'} K_{rr'} z_{r'}^{(j)} / K_{rr}, 1/K_{rr}),$$

truncated to the interval $\left[L_r^j, U_r^j\right]$ in (2).

Step 2. Sample the graph based on birth and death process.

2.1. Calculate the birth rates by equation 7 and $\beta(K) = \sum_{e\in\overline{E}} \beta_e(K)$,
2.2. Calculate the death rates by equation 8 and $\delta(K) = \sum_{e\in E} \delta_e(K)$,
2.3. Calculate the waiting time by $w(K) = 1/(\beta(K) + \delta(K))$,
2.4. Simulate the type of jump (birth or death),

Step 3. Sample the new precision matrix, according to the type of jump.

In step 3, we use the direct sampling algorithm from the precision matrix K, which is proposed by Lenkoski (2013).

4 Results

Hoff (2007) considered the analysis of multivariate dependencies among income, eduction and family background. The data concerns 1002 males in the U.S labor force. The data is available in R package `BDgraph` at http://CRAN.R-project.org/package=BDgraph. The seven observed variables which have been measured on various scales are as follow: the income (inc), degree (deg), number of children (child), parents income (pinc), parents degree (pdeg), number of parents children (pchild), and age (age).
We run our algorithm for 100,000 iterations and 60,000 as a burn-in. Figure 1 (in the left) shows convergency of our algorithm. Figure 1 (in the right) shows the graph with the highest posterior probability. Table 1 reports the posterior edge inclusion probabilities for all edges based on our proposed method. To compare the performance of our method (in the same conditions) with the existing Bayesian approaches, we check our result with the method proposed by Dobra and Linkeski (2011). Both approaches select the same graph as the best graph. The main difference is that our algorithm converges faster than their algorithm.

5 Conclusion

We proposed an efficient Bayesian framework to extend Gaussian graphical models which applicable to binary, ordinal or continuous data. Our results show that our proposed method leads to similar results as the existing methods, while being more computationally efficient and more generally applicable.

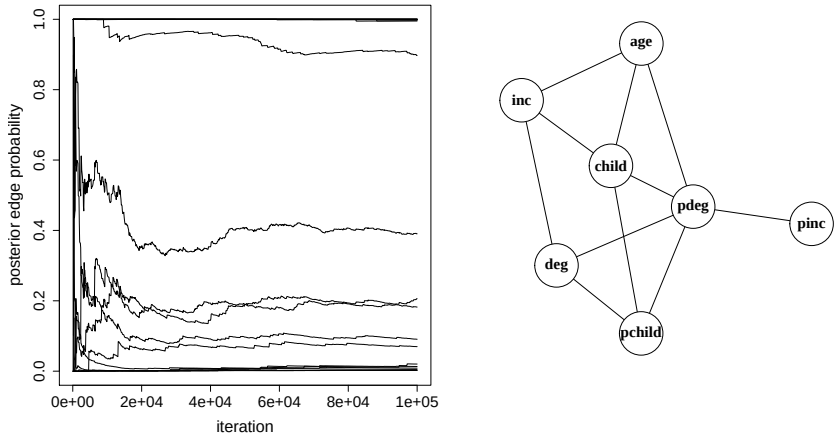


FIGURE 1. (Left) Plot of the cumulative occupancy fractions of all possible edges to check convergence of our algorithm. (Right) Most probable graph with 11 edges for the data based on the output of our algorithm.

TABLE 1. The posterior edge inclusion probabilities for all edges.

	inc	deg	child	pinc	pdeg	pchild	age
inc		1.00	1.00	0.01	0.01	0.00	1.00
deg			0.27	0.00	1.00	0.68	0.02
child				0.01	0.86	1.00	1.00
pinc					1.00	0.04	0.00
pdeg						0.99	1.00
pchild							0.00
age							

References

Dobra, A. and Lenkoski, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, **5**, 969 – 993.

Hoff, P.D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, **1**, 265 – 283.

Lenkoski, A. (2013). A direct sampler for G-Wishart variates. *Stat*, **2**, 119 – 128.

Liu, H., Lafferty, J. and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, **10**, 2295 – 2328.

Mohammadi, A. and Wit E. (2013). Bayesian model selection in sparse Gaussian graphical models. *arXiv*, 1210.5371v5.

Pitt, M., Chan, D., and Kohn, R. (2006). Efficient Bayesian inference for Gaussian copula regression models. *Biometrika Trust*, **93**, 537 – 554.

Wavelet Estimation of Functional-coefficient Regression Models

Pedro A. Morettin[1], Chang Chiann[1], Michel H. Montoril[2]

[1] University of São Paulo, Brazil
[2] University of Campinas, Brazil

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

Abstract: In this work we study the estimation of functional coefficient regression (FCR) models using wavelets. We will present the convergence rates of the proposed estimator and carry out a simulation study to evaluate which selection criterion is helpful to provide better resolution levels in order to find the more adequate model. Moreover, we will use a real data set to make forecasts and to compare our method with others known in the literature.

Keywords: nonlinear time series,wavelets,functional regression.

1 Introduction

In this work we study functional coefficient regression (FCR) models using expansion in father wavelets to estimate the coefficient functions. Let $\{Y_t, U_t, \mathbf{X}_t\}$ be a jointly strictly stationary process, where U_t is a real random variable and $\mathbf{X}_t$ a random vector in $\mathbb{R}^d$. Let $E(Y_t^2) < \infty$. Considering the multivariate regression function $m(\mathbf{x}, u) = E(Y_t|\mathbf{X}_t = \mathbf{x}, U_t = u)$, the FCR model has the form

$$m(\mathbf{x}, u) = \sum_{i=1}^{d} f_j(u)x_i, \tag{1}$$

where the $f_j(\cdot)$s are measurable functions from $\mathbb{R}$ to $\mathbb{R}$ and $\mathbf{x} = (x_1, \ldots, x_d)^{\intercal}$, with $\intercal$ denoting the transpose of a matrix or vector.
Differently from the usual, in our study it is not necessary the assumption of independence for the errors.

2 Estimation

Any wavelet basis has an associated multiresolution analysis, which is a sequence of nested and closed spaces $\{V_j\}_{j\in\mathbb{Z}}$ of $L_2(\mathbb{R})$ satisfying certain

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

properties. One of them states that there exists a function $\varphi \in V_0$ such that $\{\varphi(\cdot - k)\}_{k\in\mathbb{Z}}$ is a Riesz basis for V_0.
Usually φ is called father wavelet (or scaling function) and it is well-known that it generates a basis $\{\varphi_{Jk}\}_k$, where $\varphi_{Jk}(\cdot) = 2^{J/2}\varphi(2^J \cdot -k)$, of the space V_J, where J is called resolution level.
Now, let J_i be a resolution level associated to the coefficient function f_i, and, for sake of simplicity, denote $\phi_{ik}(\cdot) = 2^{J_i/2}\varphi_{(i)}(2^{J_i}\cdot -k)$. Thus, following the idea of Huang and Shen (2004), it is possible to approximate each coefficient function by an orthogonal projection in a multiresolution space V_{J_i} and, then, approximate (1) by

$$m(\mathbf{x}, u) \approx \sum_{i=1}^{d}\sum_{k} \alpha_{ik}\phi_{ik}(u)x_i. \tag{2}$$

If the coefficient functions and the father wavelet have compact support, there are just a finite number r_i, $i = 1, \ldots, d$, of wavelet coefficients different from zero. Thus, it is possible to estimate the wavelet coefficients of wavelets and then estimate the functions f_j of the model (1) by

$$\hat{f}_j(u) = \sum_{k=1}^{r_i} \hat{\alpha}_{ik}\phi_{ik}(u),$$

where $\hat{\alpha}_i = (\hat{\alpha}_{i1}, \ldots, \hat{\alpha}_{ir_i})^{\mathrm{T}}$, $i = 1, 2, \ldots, d$, is the estimator of α_i.
Thus, denoting the covariance matrix of the errors by $\mathbf{\Sigma}$, and supposing initially that it is known, one can estimate the wavelet coefficients vector minimizing the least squares function

$$\ell(\alpha) = (\mathbf{Y} - \mathbf{X}\alpha)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\alpha), \tag{3}$$

where $\alpha = (\alpha_1^{\mathrm{T}}, \ldots, \alpha_d^{\mathrm{T}})^{\mathrm{T}}$, $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ir_i})^{\mathrm{T}}$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and the t-th row of $\mathbf{X}$ corresponds to the vector $\phi_{ik}(U_t)X_{tj}$, $k = 1, 2, \ldots, r_i$, $i = 1, \ldots, d$. Hence, the coefficient vector estimator is given by

$$\hat{\alpha} = (\mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Y}. \tag{4}$$

Note that when the errors are independent, $\mathbf{\Sigma}$ is a identity matrix. With assumptions similar to those used by Huang and Shen (2004), we derive rates of convergence for distances between the estimators and the real functions, which are presented bellow as a theorem. Since $f_i^{J_i}$ is the orthogonal projection of f_i in V_{J_i}, denote $\rho_i = \|f_i^{J_i} - f_i\|$.

Theorem 1. *Under appropriate assumptions, we have*

$$\sum_{i=1}^{d} E\|\hat{f}_i - f_i\|_2^2 \le C\sum_{i=1}^{d}\left(\frac{2^{J_i}}{n} + \rho_i^2\right),$$

for some $C > 0$. In particular, if $\rho_i = o(1)$, then $E\|\hat{f}_i - f_i\|_2^2 = o(1)$, $i = 1, \ldots, d$.

As in practical situations the covariance matrix $\mathbf{\Sigma}$ is unknown, it has to be estimated (e.g., $\hat{\mathbf{\Sigma}}$), and with such estimator, the wavelet coefficients can be computed as

$$\tilde{\alpha} = (\mathbf{X}^{\mathrm{T}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{Y}. \tag{5}$$

If the estimator of the covariance matrix is consistent in probability, in the sense that all eigenvalues of $\hat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma} - \mathbf{I}$ are $o_p(1)$, with $\mathbf{I}$ being a identity matrix, it is possible to find that

$$|\tilde{\alpha} - \hat{\alpha}|^2 = o_p(1). \tag{6}$$

Thus, denoting $\tilde{f}_i(u) = \sum_{k=1}^{r_i} \tilde{\alpha}_{ik}\phi_{ik}(u)$, $i = 1, \ldots, d$, based on (6), we can derive the following result.

Proposition 1. *Under the same assumptions of Theorem 1, with $\hat{\mathbf{\Sigma}}$ consistent in probability in estimating $\mathbf{\Sigma}$, then*

$$\sum_{i=1}^{d} \|\tilde{f}_i - f_i\|_2^2 = O_p\left(\sum_{i=1}^{d}\left(\frac{2^{J_i}}{n} + \rho_i^2\right)\right).$$

In particular, if $\rho_i = o(1)$, then $\tilde{f}_i$ is consistent in probability in estimating f_i, i.e., $\|\tilde{f}_i - f_i\|_2 = o_p(1)$, $j = 1, \ldots, d$.

Now, it is possible to find a consistent estimator $\hat{\mathbf{\Sigma}}$ supposing that $\mathbf{\Sigma} = \mathbf{\Sigma}(\theta)$, i.e., the covariance matrix is a function of a parameter vector $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$, where p is a fixed number. In general, one can suppose that the errors of the model are represented by autoregressive processes AR(p). Then it is possible to find a consistent estimator to θ, say $\hat{\theta}$, and hence obtain a consistent estimator for the covariance matrix.

Borrowing ideas of Cochrane and Orcutt (1949), we can proceed the estimation iteratively. Firstly, the wavelet coefficients vector can be estimated acting as if the errors were independent ($\mathbf{\Sigma} = \mathbf{I}$) and then computing the residuals. Next, one can fit an autoregressive model to the residuals and by using the estimate of the autoregressive coefficients, the covariance matrix can be estimated. In the following, the wavelet coefficients vector $\tilde{\alpha}$ could be computed by (5), with the estimate of covariance matrix. This double stage procedure (computation of $\hat{\mathbf{\Sigma}}$ and $\tilde{\alpha}$) can be repeated until, for example, the convergence of the residual mean square is achieved.

Another procedure, that we will use in this work, is the following. Denoting by η the vector $(\alpha^{\mathrm{T}}, \theta^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{x}_t$ as the t-th row of $\mathbf{X}$, we estimate jointly the coefficients of the FCR model α and the autoregressive coefficients θ minimizing numerically

$$\ell(\eta) = \sum_{t=1}^{n} \left\{\theta_p(L)\left(Y_t - \mathbf{x}_t^{\mathrm{T}}\alpha\right)\right\}^2, \tag{7}$$

where $\theta_p(L) = 1 - \theta_1 L - \ldots - \theta_p L^p$ and the backshift satisfying $L^k V_t = V_{t-k}$, $k > 0$.

Selection of the resolution level

In this estimation procedure it is important to choose an adequate resolution level J. In this work, we proceed similarly to Huang and Shen (2004). We will use the information criteria AIC, AICc and BIC. Denoting the sample size by n, the number of parameters to be estimated by p and the residual mean square by RMS, these criteria are can be defined as

$$AIC = \log(RMS) + \frac{2p}{n}, \quad AICc = AIC + \frac{2(p+1)(p+2)}{n(n-p-2)}$$

and

$$BIC = \log(RMS) + \frac{p}{2}\log(n).$$

3 Simulation and application

A simulation study was carried over to assess the theoretical results and an application to the industrial production index of USA was done and will be given at the presentation.

Acknowledgments: The authors are thankful to Fapesp for the financial support (grant 2013/00506-1)

References

Huang, J.Z. and Shen, H. (2004). Functional coefficient regression models for non-linear time series: A plynomial spline approach. *Scandinavian Journal of Statistics*, **31**, 515 – 534.

Cochrane, D. and Orcutt, G.H. (1949). Application of least squares regression to relationships containing autocorrelated errors. *Journal of the American Statistical Association*, **44**, 32 – 61.

A restricted composite likelihood approach to modelling Gaussian spatial data

C.K. Mutambanengwe[1], C. Faes[1], M. Aerts[1]

[1] Hasselt University, Diepenbeek, Belgium

E-mail for correspondence: chenjerai.mutambanengwe@uhasselt.be

Abstract: We introduce restricted pairwise composite likelihood methods for estimation of mean and covariance parameters in a Gaussian random field, without resorting back to the full likelihood. A simulation study is carried out to investigate how this method works in settings of increasing domain as well as in-fill asymptotics, whilst varying the strength of correlation. Preliminary results showed that simple marginal pairwise likelihoods tend to underestimate the variance parameters, especially when there is high correlation. Using RECL together with the pairwise method improved the estimates, more so when weighting was done using effective sample size. However, the choice of how the effective sample size is calculated also affects the estimates, and some sub analyses may need to be done in order to get correct estimates for final model.

Keywords: composite likelihood; spatial dependence; weighting.

1 Introduction

Likelihood based methods are often used to estimate parameters of interest in spatial analysis. Unfortunately, even with simple likelihoods such as those obtained from Gaussian data, maximum likelihood involves inversion of matrices for each likelihood function calculated, which quickly increases the computational effort as the number of observations increases. To reduce the computational burden, composite likelihood methods have become popular in spatial statistics. A recent review of composite likelihood methods is given by Varin (2008) and Varin et al. (2011). The idea of composite likelihoods (CL) is to replace the likelihood by a simpler function, constructed from summing over the contributions of the likelihoods on subsets of the data, as such leading to a simpler function to be evaluated, but at the cost of efficiency loss. This idea was proposed by Besag (1974) in the context of spatial data, and called pseudo-likelihood. Later, it was called composite likelihood by Lindsay (1988). We will focus on the specification

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

of the composite likelihood for spatial geostatistical data based on pairwise differences, as done by Curriero and Lele (1999), and on pairwise likelihood contributions as defined by Varin (2008). However, when variance parameters are of interest, for example when interest is in the variogram, maximum likelihood (ML) estimation is known to be biased as a result of the loss in degrees of freedom. The same applies for the composite likelihood estimation of the covariance parameters. This bias can be reduced substantially by using restricted maximum likelihood (REML). In this paper, it will be investigated how the composite likelihood method can be penalised in a similar way as in REML, in order to reduce bias in the variance parameter. The proposed method will be called the restricted composite likelihood method (RECL).

2 Methods

Let $\boldsymbol{Z} = (Z(s_1), ..., Z(s_n))$ be a univariate random variable from a Gaussian random field with observations $Z(s_i)$ recorded at sites s_i $(i = 1, ..., n)$ such that

$$\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{C}(\sigma^2, \rho))$$

where $\boldsymbol{\mu} = X\beta$ is the mean of the variable and $\boldsymbol{C}$ is the covariance function, capturing the spatial dependence. The latter is a function of the variance of the spatial process σ^2, and the correlation between sites ρ determined by the distance between them. This second-order stationary process has semi-variogram

$$\gamma(s_i, s_j) = \frac{1}{2}\text{var}(Z(s_i) - Z(s_j)).$$

The most popular semivariogram is the Matérn class, with as special case the exponential semivariogram. The latter can be parametrized as $\gamma(|s_i - s_j|; \boldsymbol{\phi}) = c_0 + \sigma^2(1 - \rho^{|s_i - s_j|})$ where $= \boldsymbol{\phi} = (c_0, \sigma^2, \rho)$. The parameters c_0 and σ^2 are called the nugget and the sill, respectively, and $c_0 + \sigma^2$ represents the process variance, ρ is the spatial dependence.

Two types of composite likelihood methods are considered: (1) pairwise differences (CL_1), (2) marginal pairwise method (CL_2). When the mean parameter $\boldsymbol{\mu}$ needs to be estimated, the latter approach will result in biased estimates for the covariance function. Similar as with REML, a penalisation is added to the (log) composite likelihood function to come up with the RECL formulated as

$$RECL = \sum_{i=1}^{n-1}\sum_{j>i}^{n}(w_{ij} \ln f\,(Z(s_i), Z(s_j); \boldsymbol{\mu}_{ij}, \boldsymbol{C}_{ij})) - \frac{1}{2}\ln\left|\sum_{i=1}^{n}\sum_{j>i}^{n}\boldsymbol{X}'_{\boldsymbol{ij}}\boldsymbol{C}_{ij}^{-1}\boldsymbol{X}_{\boldsymbol{ij}}\right|$$

with weights $w_{ij} = \frac{n'}{n(n-1)}$, and with n the number of locations and n' denoting the effective sample size (ESS) given by

$$n' = \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}\rho^{|s_i - s_j|}}$$

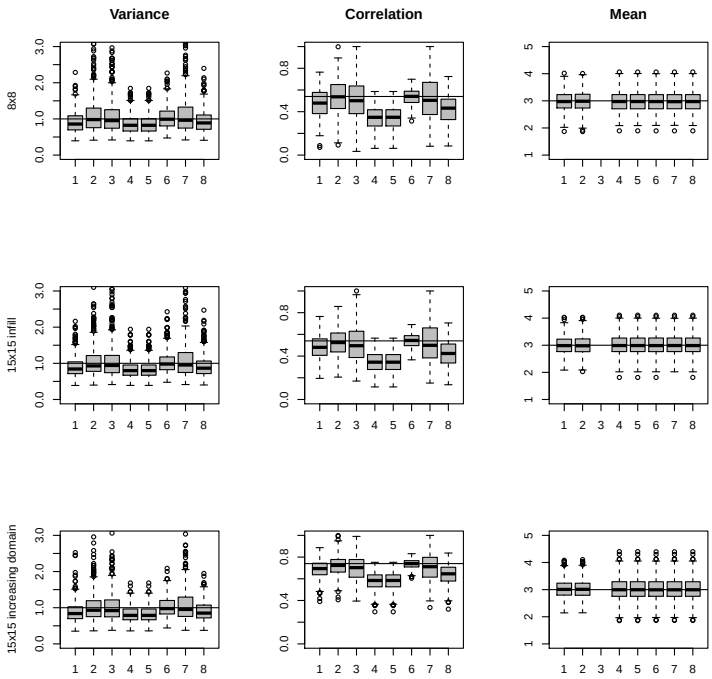


FIGURE 1. Box plot of variance parameter (left) and correlation parameter (right) for moderate dependence settings. Estimation methods: 1: ML; 2: REML; 3: CL_1; 4:CL_2; 5: $RECL_1$; 6: $RECL_2$; 7: $RECL_3$; and 8: $RECL_4$.

Fortin and Dale (2005). The weights w_{ij} are (1) set equal to 1 ($RECL_1$), or estimated by setting ρ equal to (2) known ρ ($RECL_2$), (3) $\widehat{\rho}$ from CL_1 ($RECL_3$) and (4) $\widehat{\rho}$ from CL_2 ($RECL_4$).

3 Simulation study

A simulation study is carried out to explore the properties of our estimators in a similar fashion to Curriero and Lele (1999). Data are simulated on an 8 x 8 regular grid with 1 unit interval spacing, and on two 15 x 15 grids obtained by halving the grid spacings to 0.5 (infill asymptotics) and doubling the grid spacing to 2 (increasing domain asymptotics). ρ was varied to represent relatively weak, moderate, and strong levels of spatial dependence by setting the distance (effective range) at which values become approximately uncorrelated to be 0.2, 0.5, and 0.8 times the maximum distance over the domain S. The variance parameter σ^2 was set at 1, and the mean μ was set at 3. Part of the results are summarised in Figure 1 (results for moderate dependence). The figure shows box plots of the σ^2 and ρ parameters for eight estimation methods. The horizontal line corresponds with the true underlying value.

4 Conclusions

The ML and REML estimates that use the full likelihood have better estimates than all other methods, but have the drawback that it is computationally intensive. CL_1 also works very good, but treats the mean parameters as nuisance, while they could be of interest in practice. CL_2 and $RECL_1$ perform similar to each other, with bias in mainly the parameter ρ. Inclusion of weights greatly improved the point estimates. While fixing the value of ρ to determine the weights w_{ij}, it is not useful in practice and estimation of $\hat{\rho}$ from the differences method came up with the best alternative to be used in practice. Note that all models perform relatively well when the correlation is weak. However, larger differences are observed as the correlation gets stronger, especially for the ρ parameter, with the proposed method providing a good correction.
In conclusion, penalization seems important also in composite likelihood methods, and the choice of weights is key in obtaining good results. Weighting with inverse distance and using only nearby neighbours does give better parameter estimates than the unweighted model fits. However, using estimated effective sample size shows better improvement. There is continuing work to explore variance estimation for the proposed methods, as well as inclusion of covariates and application to real data.

References

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 192 – 236.

Curriero, F.C., and Lele, S.R. (1999). A Composite Likelihood Approach to Semivariogram Estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, **4(1)**, 9 – 28.

Fortin, M.-J., and Dale, M.R.T. (2005). *Spatial Analysis: A Guide for Ecologists.* Cambridge University Press.

Lindsay, B.G. (1988). Composite likelihood methods. *Contemporary Mathematics*, **80**, 221 – 240.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, **92(1)**, 1 – 28.

Varin, C., Reid., N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5 – 42.

Estimating conditional extreme quantiles under random censoring

Pathé Ndao[1], Aliou Diop[1], Jean-François Dupuy[2]

[1] LERSTAD, University Gaston Berger, Saint Louis, Senegal
[2] IRMAR-INSA of Rennes, France

E-mail for correspondence: Jean-Francois.Dupuy@insa-rennes.fr

Abstract: We address the estimation of the extreme quantiles of a conditional distribution when the data are randomly right-censored. This issue is of particular interest in the analysis of failure time data. We propose a new estimator of the conditional extreme quantiles and we investigate its properties via simulations. The proposed estimator outperforms alternative existing methods.

Keywords: Conditional tail index; Moving window; Simulations.

1 Introduction

Consider a set of n independent observations $(Y_1, X_1), \ldots, (Y_n, X_n)$ of the couple (Y, X) where Y is the duration until some event of interest (recovery of a patient, ruin of a company...) and X is a p-vector of covariates (biological markers recorded on a patient, economic characteristics of a company...). In practice, it is often of interest to estimate extreme quantiles of the conditional distribution $F(\cdot|x)$ of Y given $X = x$ *i.e.* quantities of the form:

$$F^{\leftarrow}(1-\alpha|x) = \inf\{y : F(y|x) \geq 1-\alpha\}$$

where α is so small that this quantile falls beyond the range of the observations $Y_1, \ldots, Y_n$. Several papers address this issue, *e.g.* Gardes and Girard (2008, 2010).

In this work, we consider this problem in the more complex setting where $Y_1, \ldots, Y_n$ are randomly right-censored. Under censoring, the observations consist of triplets $(Z_i, \delta_i, X_i), i = 1, \ldots, n$ where $Z_i = \min(Y_i, C_i)$, $\delta_i = 1_{\{Y_i \leq C_i\}}$, $1_{\{\cdot\}}$ is the indicator function and C_i is a random censoring time. The estimation of extreme quantiles under censoring but without covariates is addressed by Matthys *et al.* (2004) and Einmahl *et al.* (2008).

Here, we consider the estimation of extreme quantiles when both censoring and covariates are present. We construct a new estimator of $F^{\leftarrow}(1-\alpha|x)$

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

by combining a moving-window approach with the inverse probability-of-censoring weighting principle (Section 2). The proposed estimator is asymptotically normal. We examine its finite-sample performance via simulations and we illustrate our methodology on a set of AIDS survival data (Section 3).

2 The proposed estimator

For all x, we denote by $q(\alpha, x)$ the conditional quantile of order $1-\alpha$ ($\alpha \in (0,1)$) of $F(\cdot|x)$. We assume that $F(\cdot|x)$ belongs to the domain of attraction of a Fréchet distribution with shape $\gamma_1(x)$. That is, $F(\cdot|x)$ can be written as:

$$F(u|x) = 1 - u^{-1/\gamma_1(x)} L_1(u,x)$$

where $\gamma_1(\cdot)$ is an unknown positive function of x (referred to as the conditional tail index function) and $L_1(\cdot,x)$ is a slowly varying function at infinity. A preliminary step in the estimation of $q(\alpha,x)$ is to estimate $\gamma_1(x)$. Some further notations are needed first.

Let $B(x,r) = \{t \in \mathbb{R}^p, d(x,t) \leq r\}$ and $h_{n,x}$ be a positive sequence tending to 0 as n tends to infinity. Let $m_{n,x} = \sum_{i=1}^{n} 1_{\{x_i \in B(x,h_{n,x})\}}$ be the number of observations (Z_i, X_i) lying in $[0,\infty) \times B(x,h_{n,x})$. Let $Z^x_{(1)} \leq \ldots \leq Z^x_{(m_{n,x})}$ be the ordered values of Z for these observations and $\delta^x_{(1)}, \ldots, \delta^x_{(m_{n,x})}$ be the corresponding δ's (that is, $\delta^x_{(i)} = \delta_j$ if $Z^x_{(i)} = Z_j$). The usual conditional Hill estimator of $\gamma_1(x)$ is of the form

$$\widehat{\gamma}^{(H)}_{k_x, m_{n,x}}(x) = \frac{1}{k_x} \sum_{i=1}^{k_x} i \log(Z^x_{(m_{n,x}-i+1)} / Z^x_{(m_{n,x}-i)})$$

where k_x is an integer such that $1 \leq k_x \leq m_{n,x}$. First, we adapt this estimator to censoring by dividing it by the proportion

$$\widehat{p}_x = \frac{1}{k_x} \sum_{i=1}^{k_x} \delta^x_{(m_{n,x}-i+1)}$$

of uncensored observations among the k_x largest Z_i in a neighbourhood of x. Our estimator of $\gamma_1(x)$ is thus:

$$\widehat{\gamma}^{(c,H)}_{k_x, m_{n,x}}(x) = \frac{\widehat{\gamma}^{(H)}_{k_x, m_{n,x}}(x)}{\widehat{p}_x} \tag{1}$$

Based on this, we now address the estimation of conditional extreme quantiles $q(\alpha_{m_{n,x}}, x)$ of order $1-\alpha_{m_{n,x}}$ of $F(\cdot|x)$. Such quantiles verify $1 - F(q(\alpha_{m_{n,x}}, x)|x) = \alpha_{m_{n,x}}$ where $\alpha_{m_{n,x}} \to 0$ as $m_{n,x} \to +\infty$.

We define a conditional Kaplan-Meier-type estimator based on the moving-window approach:

$$1 - \widehat{F}_{m_{n,x}}(y|x) = \prod_{i=1}^{m_{n,x}} \left(\frac{m_{n,x} - i}{m_{n,x} - i + 1} \right)^{\delta^x_{(i)} 1_{\{Z^x_{(i)} \leq y\}}} \tag{2}$$

Based on (1) and (2), we finally propose the following Weissman-type estimator of $q(\alpha_{m_{n,x}}, x)$:

$$\widehat{q}^{(c,H)}(\alpha_{m_{n,x}}, x) = Z^x_{(m_{n,x}-k_x)} \left(\frac{1 - \widehat{F}_{m_{n,x}}(Z^x_{(m_{n,x}-k_x)}|x)}{\alpha_{m_{n,x}}} \right)^{\widehat{\gamma}^{(c,H)}_{k_x,m_{n,x}}(x)} \tag{3}$$

The estimator (3) is consistent and asymptotically normal (see Ndao *et al.*, 2013). The next section reports a small part of the results of a comprehensive simulation study conducted by Ndao *et al.* (2013) and a case-study.

3 Simulation study and a real-data example

3.1 A simulation study

We first simulate 1000 samples of size n ($n = 500, 1000, 1500$) of independent replicates (Z_i, δ_i, x_i). The conditional distribution of Y_i given $X = x_i$ is Pareto with parameter $\gamma_1(x) = .5(.1+\sin(\pi x))(1.1-.5\exp(-64(x-.5)^2))$ (with $x \in [0, 1]$) and the distribution of C_i is chosen to yield various censoring percentages c ($c = 10\%, 25\%, 40\%$). For each sample, we obtain the estimate (3) of the conditional extreme quantile $q(1/5000, 0.5) \approx 19.70786$ of order $1 - 1/5000$ of $F(\cdot|0.5)$. Then, we obtain the averaged value of the 1000 estimates along with their RMSE and MAE. We also obtain asymptotic 95%-level confidence intervals for $q(1/5000, 0.5)$, along with the empirical coverage probabilities over the 1000 intervals. Table 1 reports the results.

TABLE 1. Conditional extreme quantile estimation: simulation results.

n		$c = 10\%$	$c = 25\%$	$c = 40\%$
500	average est.	19.777	20.225	20.072
	RMSE	(.258)	(.265)	(.310)
	MAE	[.326]	[.333]	[.383]
	conf. interval	[16.00,25.88]	[15.90,27.77]	[15.56,28.25]
	coverage prob.	0.594	0.936	0.970
1000	average est.	19.381	19.960	20.086
	RMSE	(.182)	(.206)	(.222)
	MAE	[.226]	[.259]	[.280]
	conf. interval	[16.54,23.39]	[16.71,24.77]	[16.55,25.53]
	coverage prob.	0.708	0.971	0.989
1500	average est.	19.841	19.981	19.905
	RMSE	(.142)	(.161)	(.179)
	MAE	[0.177]	[0.199]	[0.223]
	conf. interval	[17.75,22.47]	[17.21,23.59]	[17.26,23.70]
	coverage prob.	0.910	0.990	0.992

From these results (and the additional results in Ndao *et al.*, 2013), the proposed estimator (3) provides a satisfactory approximation of the true

extreme quantile in a wide range of simulation scenarios. Moreover, this estimator outperforms competing alternatives such as the complete-case estimator (see Ndao *et al.*, 2013).

3.2 A real-data example

We now illustrate our methodology on a set of AIDS survival data. The dataset contains $n = 2754$ male patients diagnosed with AIDS in Australia before 1 July 1991 (see Venables and Ripley, 2002). The information on each patient includes the age x at diagnosis, the date of death or end of observation and an indicator which equals 1 if the patient died and 0 otherwise. 1708 patients died, the other survival times are right-censored. We estimate the quantile $q(1/1000, x)$ of order $1-1/1000$ of the conditional distribution of the survival time given x for $x = 27, 37, 47$. The estimated conditional extreme quantiles obtained from (3) are 10.04 years (when $x = 27$), 12.77 years ($x = 37$) and 11.29 years ($x = 47$). The "naive" complete-case method provides substantially smaller (and certainly biased) quantile estimates, which is consistent with the findings of our simulation study.

Acknowledgments: Pathé Ndao acknowledges financial support from the Agence Universitaire de la Francophonie.

References

Einmahl, J.H.J., Fils-Villetard, A., and Guillou, A. (2008). Statistics of extremes under random censoring. *Bernoulli*, **14**, 207 – 227.

Gardes, L. and Girard, S. (2008). A moving window approach for nonparametric estimation of the conditional tail index. *Journal of Multivariate Analysis*, **99**, 2368 – 2388.

Gardes, L. and Girard, S. (2010). Conditional extremes from heavy-tailed distributions: an application to the estimation of extreme rainfall return levels. *Extremes*, **13**, 177 – 204.

Matthys, G., Delafosse, E., Guillou, A., and Beirlant, J. (2004). Estimating catastrophic quantile levels for heavy-tailed distributions. *Insurance: Mathematics & Economics*, **34**, 517 – 537.

Ndao, P., Diop, A., and Dupuy, J.-F. (2013). Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. Accepted for publication in *Computational Statistics & Data Analysis*.

Venables, W.N., and Ripley, B.D. (2002). *Modern Applied Statistics with S*. New York: Springer.

On (Semiparametric) Mode Regression

Margret-Ruth Oelker[1], Fabian Sobotka[2], Thomas Kneib[2]

[1] Ludwig-Maximilians Universität München, Germany
[2] Georg-August-Universität Göttingen, Germany

E-mail for correspondence: margret.oelker@stat.uni-muenchen.de

Abstract: We propose an iteratively reweighted least squares algorithm (IRLS) to perform a regression that approximates the conditional mode of a response. The adaption of the tuning parameters of the approximation on the fly results in a stable estimate. For linear predictors, a close link to kernel methods allows to derive asymptotic properties easily. Furthermore, the quadratic approximation allows for the inclusion of semiparametric models and quadratic penalties.

Keywords: Mode regression; Kernel Methods; Semiparametric Models; IRLS.

1 Introduction

Recent years have seen a tremendous increase in interest related to regression beyond the mean of the conditional distribution of a response given covariates. Surprisingly, regression models for the conditional mode of the response distribution given covariates have received very little attention. However, estimating conditional modes would be of high interest since (i) the mode is by far the visually most prominent feature of a density; (ii) the mode is extremely robust with respect to outliers; (iii) the mode provides a location measure that is easily communicated to practitioners; (iv) there may be situations where the dependence of the mode on covariates may be quite different from the dependence of the median and/or the mean.
Consider the regression specification $y = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where y is the response variable of interest, $\boldsymbol{x} \in \mathbb{R}^q$ is a vector of covariates supplemented with regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^q$ and $\boldsymbol{\varepsilon}$ is the error term. Unlike in mean regression, we do not assume $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$, but

$$\arg\max_{\xi} f_{\varepsilon|\boldsymbol{x}}(\xi|\boldsymbol{x}) = 0, \tag{1}$$

i.e. the conditional density of the error terms $f_{\varepsilon}(\cdot|\boldsymbol{x})$ is assumed to have a global mode at zero. This implies that the regression predictor $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$ is

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

the conditional mode of the response distribution $f_y(\cdot|\boldsymbol{x})$ and the mode regression coefficient can be defined as

$$\boldsymbol{\beta} = \arg\max_{\boldsymbol{b}} f_\varepsilon(y - \boldsymbol{x}^\mathrm{T}\boldsymbol{b}|\boldsymbol{x}). \tag{2}$$

An equivalent formulation is obtained based on the loss function $L(\xi) = 1(\xi \neq 0)$, i.e. the indicator function for arguments different from zero. In this case, the mode regression coefficient is given by

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}} \mathbb{E}\left[L(y - \boldsymbol{x}^\mathrm{T}\boldsymbol{b})|\boldsymbol{x}\right]. \tag{3}$$

Unfortunately, an estimate for the mode regression coefficient cannot be determined by an empirical analogue to (2) unless specific assumptions are made for the error density $f_\varepsilon(\cdot|\boldsymbol{x})$. Criterion (3) is anyway not useful for modal regression based on data with continuous error distribution since in this case, there will in general be no unique solution even if the density of the errors ε has a global mode. As a consequence, earlier attempts to mode regression usually either rely on nonparametric kernel regression from which the mode is then derived in a second step (Collomb et al., 1987; Einbeck and Tutz, 2006) or on different types of approximations to the loss function $L(\xi)$ (Lee, 1989; Kemp and Santos Silva, 2012). We build upon Kemp and Santos Silva (2012) and (i) provide a differentiable approximation to the loss function defining mode regression such that an iteratively re-weighted least squares (IRLS) algorithm can be used to estimate the mode regression coefficients, (ii) study the properties of this estimate and show its consistency and asymptotic distribution, (iii) extend the purely linear mode regression model to additive models combining nonparametric effects of several covariates in a penalized IRLS framework. The main advantage of the IRLS framework we apply, is that it allows to easily include extended regression functionality from (generalized) additive models which also rely on IRLS estimation. In fact, we can further exploit this connection by determining smoothing parameter estimates within the IRLS framework such that the semiparametric mode regression estimate is fully data-driven.

2 Iteratively Reweighted Least Squares Estimation

The basic idea to proof the equivalence of the two mode coefficient conditions in (2) for the density of the error term and in (3) for the zero-one-loss is based on an approximation that, in the limit, represents the zero-one-loss. Consider

$$\mathcal{L}_\epsilon(\xi) = 1(-\epsilon \leq \xi \leq \epsilon), \tag{4}$$

as an approximation to $L(\xi)$, where ϵ defines a local environment around zero. Based on this approximate loss function, any value within the $\pm\epsilon$ interval around the mode is a solution to minimizing the expected loss. In the limiting case $\epsilon \to 0$, we re-obtain the original loss function and the estimate approaches the true mode.

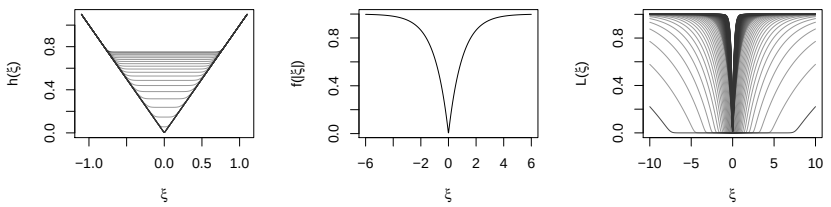


FIGURE 1. Illustration of the loss function. Left panel: function $h(\xi)$. Middle panel: function $f(|\xi|)$. Right panel: $\mathcal{L}(\xi)$. In all figures, the parameter $c = 10^{-5}$ is fixed, parameters k and g vary: $g = 20, \ldots, 1$; $k = 0.1, \ldots, 6$ in 99 steps.

Our approach to mode regression follows a similar reasoning: the loss function is approximated such that it is zero not only for $\xi = 0$, but in a surrounding of $\xi = 0$. The approximation – denoted by $\mathcal{L}(\xi)$ – will replicate the idea of (4), i.e. $\mathcal{L}(\xi)$ will have a very broad minimum in the early iterations and it will be very close to $L(\xi)$ for the final iteration of the proposed algorithm. However, $L(\xi)$ is approximated by a continuously differentiable function. This has two important advantages: (i) the approximated loss $\mathcal{L}(\xi)$ can be linked to iteratively re-weighted least squares estimation, and (ii) the smooth approximation allows to determine asymptotic properties such as consistency and asymptotic normality. To imitate the idea of the approximation (4), the tuning parameters of $L(\xi)$ approximate their limiting values while iterating such that $\mathcal{L}(\xi)$ is close to $L(\xi)$ when the algorithm converges. To allow for a smooth transition, the algorithm will have a small step length and thus relatively many iterations until convergence. Concretely, we employ the function

$$\mathcal{L}(\xi) = 1 - \exp(c^{\frac{1}{2g}} - ((k\xi)^{2g} + c)^{\frac{1}{2g}}),$$

depending on the set of tuning parameters $\mathcal{T} = \{g, k, c\}$ and $\lim_{\mathcal{T}\to T} \mathcal{L}(\xi) = L(\xi)$ for some set of limiting values T. $\mathcal{L}(\xi)$ is constructed by the scaled composition $f(h(k\cdot\xi))$ of the two functions $f(\xi)$ and $h(\xi)$: $h(\xi) = (\xi^{2g}+c)^{\frac{1}{2g}}$, where g is as a positive integer and c is a small, positive constant. As illustrated in Figure 1, $h(\xi)$ accounts for the broad minimum that is needed to imitate approximation (4) of the zero-one-loss. For the limiting value $g = 1$, $h(\xi)$ simply approximates the absolute value function. Due to the constant c, it is continuously differentiable. $f(\xi) = 1 - \exp(-\xi)$; let k be a positive number, then $f(k \cdot \xi)$ actually approximates the indicator $L(\xi)$; the approximation is the closer to $L(\xi)$, the larger k is. Hence, the tuning parameters have to be chosen such that g is relatively large in the early iterations of the IRLS algorithm; it should equal one for the final iteration. k is relatively small in the beginning of the algorithm and as large as possible for the final iteration. The constant c is as small as possible. As the value of k affects the width of the minimum of $\mathcal{L}(\xi)$ for $g > 1$, it is possible to choose a fixed sequence for g and to address all issues of tuning by a properly chosen sequence of k. We propose to choose the initial value of k driven by the data: the minimum of the loss function in the initial

iteration of the IRLS algorithm should capture all observed error terms; whereat $\boldsymbol{\varepsilon}$ is estimated by a preceding median regression.

3 Asymptotic Properties

We consider the properties of the estimate $\hat{\boldsymbol{\beta}}_n$ for $n \to \infty$. For the algorithm's final iteration and with vector notation, minimizing $\mathcal{L}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ is equivalent to the minimization of $1 - K(u)$ where

$$u \quad \mapsto \quad K(u) = \frac{1}{2} \exp\left\{-\sqrt{u^2 + c}\right\}, \; 0 < c \leq 1, \tag{5}$$

where $u = k \cdot (y - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})$ and where k is a scaling parameter with $k \to \infty$ at a proper rate. $K(u)$ in turn is an approximation of $\frac{1}{2}\exp(-|u|)$ which is the density of a Laplace distributed random variable U with mean $\mathbb{E}(U) = 0$ and variance $\mathbb{V}(U) = 2$. That is, for the final iteration, the proposed approximation can be interpreted as one minus a rounded (and thus, differentiable) Laplace kernel. Kemp and Santos Silva (2012) derive asymptotic properties for mode regression for a general kernel $K(u)$. One can show that function (5) structurally fits in this framework as the tuning parameter k relates inversely to the bandwidth δ of Kemp and Santos Silva (2012). (A scaled version of) function (5) meets all requirements needed to prove consistent and asymptotically normal estimates. There is a consistent estimate of the asymptotic covariance matrix of $\boldsymbol{\beta}$. However, the speed of convergence is at most $n^{2/7}$. Even though it is possible to link the choice of the final value of k to the required assumptions, the results depend crucially on the observed sample when the number of observations is in a realistic range. In practice, we advise to apply bootstrap methods to assess the estimate's variance.

4 Semiparametric Mode Regression

Semiparametric regression models allow for unspecified predictive functions such as

$$\boldsymbol{y} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{r} f(\boldsymbol{z}_j) + \varepsilon,$$

where as before, $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$ represents linear effects. Functions $f(\cdot)$ can represent nonlinear smooth effects of continuous covariates $\boldsymbol{z}_j$, $j = 1, \ldots, r$, for example, modeled by penalized B-splines (Eilers and Marx, 1996); but they can also depict flexible spatial effects or other additive components. Approximating mode regression by replicating the idea of the theoretical approximation (4) continuously is a stable procedure. It allows for reliable estimation even when there are relatively many covariates. Moreover, approximating mode regression with an IRLS algorithm provides a very versatile computational framework. Estimation can be easily amended by quadratic penalties of form $P_\lambda(\boldsymbol{\beta}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}_\lambda\boldsymbol{\beta}$, $\boldsymbol{K}_\lambda \in \mathbb{R}^{q \times q}$; the index denotes the

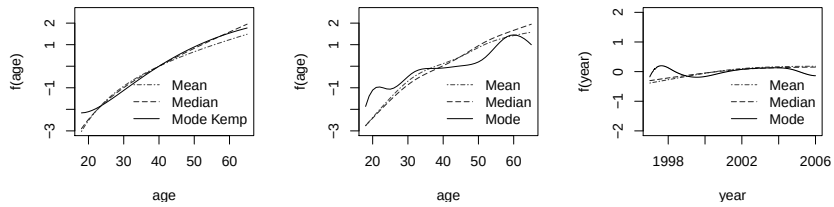


FIGURE 2. Left panel: estimated effect of the age in the model of Kemp and Santos Silva (2010). Middle panel: estimated effect of the age in the semiparametric model. Right panel: estimated effect of the year in the semiparametric model.

dependency of $\boldsymbol{K}_\lambda$ on one or several smoothing parameter(s) λ. Hence, the stable computational approach and its compatibility with quadratic penalties enable semiparametric mode regression. Like a modular system and with none but the usual restrictions, mode regression can be combined with any quadratic penalty and/or smooth component. In the IRLS framework, the proposed method is incorporated by weighting the algorithm with the derivatives of the approximation $\mathcal{L}(\xi)$. Hence, it is possible to combine mode regression with existing software by alternating the updates of the approximation and of the IRLS algorithm. Concretely, we employ the package `mgcv` (Wood, 2011; R Core Team, 2013) to implement penalized cubic B-splines, Markov random fields and different methods for the estimation of the smoothing parameter(s) λ. The results of conducted numerical experiments are promising – especially when the smoothing parameters are chosen by the negative log restricted likelihood (REML criterion).

5 Application and Outlook

To illustrate the value of semiparametric mode regression, we reanalyze the exemplary data set employed in Kemp and Santos Silva (2010). The aim of the analysis is to explain the development of the body mass index (BMI) based on the composition of the population in England. We focus on non-pregnant women between the ages of 18 and 65 observed in the period between 1997 and 2006. The available covariates are the age, the year of the observation and a binary factor indicating non-white women. In Kemp and Santos Silva (2010), the effect of log(age) is modeled by a third order polynomial while the other covariates are considered with one coefficient only. The left panel of Figure 2 shows the re-transformed effect of log(age). In mean, median and mode regression, the effect of the age differs only slightly. The scalar effect of the year (not shown) stands out as it is positive in mean and median regression but negative in the mode regression estimated with the methods of Kemp and Santos Silva (2010). However, the methodology of Sections 2 and 4 allows to model the effects of age and year smoothly: the middle and the right panel of Figure 2 illustrate that the effects of age and year estimated with the proposed method differ substantially from those in the parametric model

of Kemp and Santos Silva (2010). The effect of the age seems to follow the aging process of women – the effect changes e.g. in the span of the menopause. In contrast, the effect of the year meanders around zero. Amongst others, future work has to address the impact of the error distribution – especially cases where the mode is at the margin of the domain of the error distributions are interesting as they are not yet covered by existing theory. Moreover, it would be eligible to find approaches for multi-modal error distributions.

References

Collomb, G., Härdle, W. and Hassani, S. (1987). A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference*, **15**, 227 – 236.

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

Einbeck, J. and Tutz, G. (2006). Modelling beyond regression functions: An application of multimodal regression to speed-flow data. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, **55**, 461 – 475.

Kemp, G.C. and Santos Silva, J. (2010). Regression towards the mode. *Economics Discussion Papers, Department of Economics, University of Essex*, **686**.

Kemp, G.C. and Santos Silva, J. (2012). Regression towards the mode. *Journal of Econometrics*, **170**, 92 – 101.

Lee, M. (1989). Mode regression. *Journal of Econometrics*, **42**, 337 – 349.

R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*,, http://www.R-project.org/.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, **73**, 3 – 36.

Inference for generalized linear mixed models with sparse structure

Helen Ogden [1]

[1] University of Warwick, Coventry, U.K.

E-mail for correspondence: H.Ogden@warwick.ac.uk

Abstract: Generalized linear mixed models are a natural and widely used class of models, but one in which the likelihood often involves an integral of very high dimension. Because of this apparent intractability, many alternative methods have been developed for inference in these models, but all can fail when the model is sparse, in that there is only a small amount of information available on each random effect. A new approximation method is introduced, which exploits the structure of the integrand of the likelihood to reduce the cost of finding a good approximation to the likelihood in models with sparse structure. The method is demonstrated for models for tournaments between pairs of players, and for models with nested random-effect structure.

Keywords: Graphical model, Laplace approximation, Nested model, Pairwise competition model

1 A motivating example: pairwise competition models

Consider a tournament among n players, consisting of m contests between pairs of players. Each contest has a binary outcome y_{ij}, which takes value 1 if i beats j, and 0 if j beats i. We suppose that each player i has some ability λ_i, and that conditional on all the abilities, the outcomes of the contests are independent, with distribution depending on the difference in abilities of the players i and j. In particular, we suppose that $Pr(Y_{ij} = 1|\lambda) = h(\lambda_i - \lambda_j)$ for some known function $h(.)$. For example, if $h(x) = \text{logit}^{-1}(x)$, then this describes a Bradley-Terry model (Bradley and Terry, 1952).
If covariate information $\mathbf{x}_i$ is available for each player, then interest may lie in the effect of the observed covariates on ability, rather than the individual abilities λ_i themselves. To model how the ability of a players depends on the observed covariates, we might initially suppose that $\lambda_i = \beta^T \mathbf{x}_i$ for some unknown parameter β. However, in such a model any two players with the

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

same covariate values must have the same ability. This is unlikely to be true in practice, so we add an extra error term to our model for λ_i, so that $\lambda_i = \beta^T \mathbf{x}_i + \sigma u_i$, where u_i are independent $N(0,1)$ samples.
In order to do inference on the parameters $\theta = (\beta, \sigma)$ of the model, we attempt to compute the likelihood, given by

$$L(\theta; \mathbf{y}) = \int_{\mathbb{R}^n} \prod_{i,j} h\left[\left(\beta^T(\mathbf{x}_i - \mathbf{x}_j) + \sigma(u_i - u_j)\right)(-1)^{y_{ij}+1}\right] \prod_{j=1}^{n} \phi(u_j) d\mathbf{u}.$$

Unless n is very small, it will not be possible to approximate the likelihood well by direct computation of this n-dimensional integral. It is therefore common to use some approximation to the likelihood in place of the true likelihood for inference. If a poor-quality approximation is used, the resulting inference can have some very bad statistical properties.
This problem is not unique to the example of pairwise competition models, and the likelihood of the parameters of any generalized linear mixed model may be similarly written as an n-dimensional integral over the random effects. In the special case of a nested model with only one layer of random effects, the likelihood factorizes into a product of one-dimensional integrals, but in other cases the likelihood does not factorize in this way.

2 Existing methods for approximating the likelihood

Pinheiro and Bates (1995) suggest using a Laplace approximation to the likelihood. Write $g(u_1, \ldots, u_n | \mathbf{y}, \theta)$ for the integrand of the likelihood. This may be thought of as a non-normalized version of the posterior density for $\mathbf{u}$, given $\mathbf{y}$, and θ. For each fixed θ, the Laplace approximation relies on a normal approximation $g^{\mathrm{na}}(.|\mathbf{y}, \theta)$ to $g(.|\mathbf{y}, \theta)$, so that

$$g^{\mathrm{na}}(\mathbf{u}|\mathbf{y}, \theta) = \frac{g(\mu_\theta|\mathbf{y}, \theta)}{\phi_n(\mu_\theta; \mu_\theta, \Sigma_\theta)} \phi_n(\mathbf{u}; \mu_\theta, \Sigma_\theta)$$

for some μ_θ and Σ_θ, where we write $\phi_n(.; \mu, \Sigma)$ for the $N_n(\mu, \Sigma)$ density. When we integrate over $\mathbf{u}$, only the normalizing constant remains, so that

$$L^{\mathrm{Laplace}}(\theta|\mathbf{y}) = \frac{g(\mu_\theta|\mathbf{y}, \theta)}{\phi_n(\mu_\theta; \mu_\theta, \Sigma_\theta)} = (2\pi)^{-\frac{n}{2}} (\det \Sigma_\theta)^{-\frac{1}{2}} g(\mu_\theta|\mathbf{y}, \theta).$$

In the case of a linear mixed model, the approximating normal density is precise, and there is no error in the Laplace approximation to the likelihood. In other cases, and particularly when the response is discrete and may only take a few values, the error in the Laplace approximation may be large, especially when the model is sparse, so that there is only a small amount of information available on each random effect.
In cases where the Laplace approximation fails, Pinheiro and Bates (1995) suggest constructing an importance sampling approximation to the likelihood, based on samples from the normal distribution $N_n(\mu_\theta, \Sigma_\theta)$. Unfortunately, there is no guarantee that the variance of the importance weights

will be finite. In such a situation, the importance sampling approximation will still converge to the true likelihood, but the convergence may be slow and erratic.

3 A new method for approximating the likelihood

The approximations to the likelihood introduced so far have ignored the structure of the integrand of the likelihood, $g(.|\mathbf{y}, \theta)$. We define a graph, $\mathcal{G}_1$, which we call the posterior dependence graph, so that the posterior distribution of $\mathbf{u}$ given $\mathbf{y}$ is a graphical model with conditional independence structure represented by $\mathcal{G}_1$. To do this, we define $\mathcal{G}_1$ to have a node for each random effect, with an edge between two nodes if both random effects are involved in a single observation. In a pairwise competition model, $\mathcal{G}_1$ is the graph with a node for each player and an edge between two nodes if there is at least one contest between the corresponding pair of players.
To see how the graphical model structure can help to simplify computation of the likelihood, we first require some definitions. A complete graph is one in which there is an edge from each node to every other node. A clique of a graph $\mathcal{G}$ is a complete subgraph of $\mathcal{G}$, and a clique is said to be maximal if it is not itself contained within a larger clique. We write M_1 for the set of maximal cliques of $\mathcal{G}_1$. By construction of $\mathcal{G}_1$, we are able to factorize the the integrand of the likelihood over these maximal cliques, as $g(\mathbf{u}|\mathbf{y}, \theta) = \prod_{C \in M_1} g_C(\mathbf{u}_C)$.
We now describe a new method, which we call the 'sequential reduction' method, which exploits this factorization to reduce the cost of finding a good approximation to the likelihood. The new method is similar to the variable elimination algorithm for finding the marginal distributions of an undirected graphical model (see, for example, Jordan (2004)), although the posterior distribution of the random effects is continuous, so existing methods which work for discrete distributions may not be applied directly.

1. The u_i may be integrated out in any order. Later, we discuss how to choose a good order, with the aim of minimizing the cost of approximating the likelihood. Reorder the random effects so that we integrate out $u_1, \ldots, u_n$ in that order.

2. Factorize $g(\mathbf{u}|\mathbf{y}, \theta)$ over the maximal cliques M_1 of the posterior dependence graph, as $g(\mathbf{u}|\mathbf{y}, \theta) = \prod_{C \in M_1} g^1_C(\mathbf{u}_C)$.

3. Once $u_1, \ldots u_{i-1}$ have been integrated out (using some approximate method), we have the factorization $\tilde{g}(u_i, \ldots, u_n|\mathbf{y}, \theta) = \prod_{C \in M_i} g^i_C(\mathbf{u}_C)$, of the (approximated) non-normalized posterior for $u_i, \ldots, u_n$. Write

$$g^i_{N_i}(\mathbf{u}_{N_i}) = \prod_{C \in M_i : C \subset N_i} g^i_C(\mathbf{u}_C).$$

   We first store an approximate representation $\tilde{g}^i_{N_i}(.)$ of $g^i_{N_i}(.)$, then integrate over u_i, to give an approximate representation $\tilde{g}^i_{N_i \setminus i}(.)$ of

$g^i_{N_i \setminus i}(.)$. An approximate representation of $g^i_{N_i}(.)$ could be found by evaluating $g^i_{N_i}(\mathbf{u}_{N_i})$ at a fixed set of points for $\mathbf{u}_{N_i}$, then interpolating between those fixed points.

4. Write

$$\tilde{g}(u_{i+1}, \ldots, u_n | \mathbf{y}, \theta) = \tilde{g}^i_{N_i \setminus i}(\mathbf{u}_{N_i \setminus i}) \prod_{C \in M_i : C \not\subset N_i} g^i_C(\mathbf{u}_C),$$

defining a factorization of the (approximated) non-normalized posterior density of $\{u_{i+1}, \ldots, u_n\}$ over the maximal cliques M_{i+1} of the new posterior dependence graph $\mathcal{G}_{i+1}$.

5. Repeat steps (3) and (4) for $i = 1, \ldots, n-1$, then integrate $\tilde{g}(u_n | \mathbf{y}, \theta)$ over u_n to give the approximation to the likelihood.

If we store an approximate representation of $g^i_{N_i}(.)$ by using a $|N_i|$ - dimensional grid of points, with K points in each direction, the cost of integrating out u_i is $O(K^{|N_i|})$, The random effects may be removed in any order, so it makes sense to use an ordering that allows approximation of the likelihood at minimal cost. This problem may be reduced to a problem in graph theory: to find an ordering of the vertices of a graph, such that when these nodes are removed in order, joining together all neighbors of the vertex to be removed at each stage, the largest clique obtained at any stage is as small as possible. This is known as the triangulation problem, and the smallest possible value, over all possible orderings, of the largest clique obtained at some stage is known as the treewidth of the graph. If the posterior dependence graph has treewidth T, the likelihood may be approximated at cost at most $O(nK^T)$. If the treewidth is small, this will be much less than the $O(K^n)$ cost of direct numerical integration.

4 Examples

We now demonstrate the sequential reduction method in two examples, using code written by the author in R (R Core Team, 2013).

4.1 A pairwise competition model

We simulate a tournament from a pairwise competition model with $n = 63$ players, which has 'tree' structure, so that the posterior dependence graph is a tree (a graph with no cycles). Any tree has treewidth 2, so in this case the sequential reduction method may be used to approximate the likelihood at cost $O(nK^2)$. Suppose that there is a single observed covariate x_i for each player, where $\lambda_i = \beta x_i + \sigma u_i$ and $u_i \sim N(0, 1)$. We simulate from the model with the moderately large parameter values $\beta = 1.5$ and $\sigma = 1.5$. The covariates x_i are independent draws from a Bernoulli($\frac{1}{2}$) distribution.

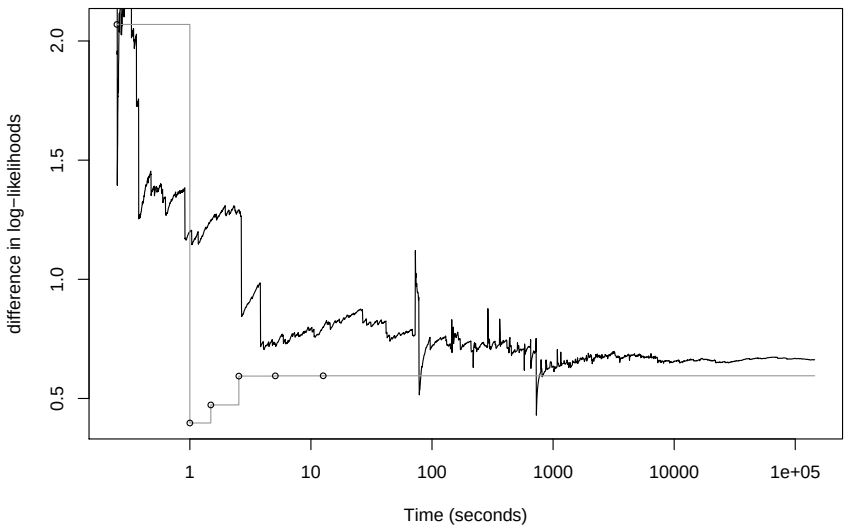


FIGURE 1. Importance sampling and sequential reduction approximations to $\ell^p(1.20, 1.06) - \ell^p(2.00, 1.50)$, plotted against the time taken to find the approximation, on a log scale.

The performance of the new method is compared with that of an importance sampling approximation. We consider approximations to the difference between the log-likelihood at two points for $\theta = (\beta, \sigma)$, $(1.20, 1.06)$ and $(2.00, 1.50)$, and consider the quality of each approximation relative to the time taken to compute it. Figure 1 shows the trace plots of importance sampling and sequential reduction approximations to this difference in log-likelihoods, plotted against the length of time taken to find each approximation, on a log scale. The improvement offered by the new method is dramatic: the sequential reduction method provides a more accurate likelihood approximation in a few seconds than an importance sampling approximation which takes more than a day to compute.

4.2 A nested model

Although the sequential reduction method has been described for pairwise competition models, the same method may be applied for other generalized linear mixed models. We demonstrate this using a nested model, with two layers of random effects. Suppose that binary observations are made on items, each contained within some level-1 grouping, which themselves are nested within some level-2 grouping. Writing $g_1(i)$ and $g_2(i)$ for the level-1 and level-2 groups of item i, we model

$$\text{logit}Pr(Y_i = 1|\mathbf{u}, \mathbf{v}) = \alpha + \beta x_i + \sigma_1 u_{g_1(i)} + \sigma_2 v_{g_2(i)}$$

where each $u_j, v_j \sim N(0, 1)$. In our example, there are 2 items in each level-1 group, 5 level-1 groups in each level-2 group, and 10 level-2 groups in total.

TABLE 1. Sequential reduction estimates for the three-level model

	Laplace	$k = 2$	$k = 3$	$k = 4$	$k = 5$
(Intercept)	-0.043	-0.044	-0.045	-0.045	-0.045
x	0.933	1.070	1.154	1.153	1.154
σ_1	0.908	1.240	1.441	1.437	1.439
σ_2	0.900	0.986	1.014	1.013	1.014

The treewidth of the posterior dependence graph is 2, so the cost of computing the sequential reduction approximation to the likelihood is $O(nK^2)$ Using a grid with $K = 2^k - 1$ points in each direction for storage gives the parameter estimates shown in Table 1. The estimates quickly stabilize as we increase k.

5 Discussion

Many common approaches to inference in generalized linear mixed models rely on approximations to the likelihood, which may be of poor quality if there is little information available on each random effect. There are many situations in which it is unclear how good an approximation to the likelihood will be, and how much impact the error in the approximation will have on the statistical properties of the resulting estimator. It is therefore very useful to be able to obtain an accurate approximation to the likelihood at reasonable cost.

The sequential reduction method allows a good approximation to the likelihood to be found in many models with sparse structure — precisely the situation where currently-used approximation methods perform worst. Some cases do remain in which the model is sparse and yet the treewidth is large, and further work is required to construct an accurate approximation to the likelihood in these cases.

References

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, **39** 324 – 345.

Jordan, M.I. (2004). Graphical models. *Statistical Science*, **19** 140 – 155.

Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the log - likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4** 12 – 35.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Joint modeling of between-subject and within-subject covariance matrices for financial data

Yi Pan[1], Jianxin Pan[1]

[1] School of Mathematics, The University of Manchester, U.K.

E-mail for correspondence: yi.pan@manchester.ac.uk

Abstract:
In this paper, we propose parsimonious models for joint modeling between-subject covariance matrix Σ_1 and within-subject covariance matrix Σ_2 for financial data such as panel data or longitudinally balanced data. In our approach, the modeling of Σ_1 is based on the alternative Cholesky decomposition (A.CD) in the form of $\Sigma_1 = DLL^TD$ (Chen and Dunson, 2003), where D is a diagonal matrix and L is a lower triangular matrix with 1's as the main diagonal entries, while the modeling of Σ_2 is based on the modified Cholesky decomposition (M.CD) in the form of $\Sigma_2 = LD^2L^T$ (Pourahmadi,1999). We also consider other covariance structures including compound symmetry ($\Sigma_1 = DRD$) specified to Σ_1 and $GARCH(1,1)$ structure to Σ_2. Simulation studies show that the proposed approach works quite well.

Keywords: Covariance matrix modelling; Cholesky Decomposition; GARCH.

1 Introduction

Nowadays, modern technologies make big panel data available in many areas like genomic, biomedical study and finance, where both the number of subjects P and the number of repeated measurements T are large. In finance, the statistical analysis of high-dimensional data usually involves the estimation of between-subject covariance matrix Σ_1 and its inverse Σ_1^{-1} (also referred to as precision matrix), because they are very useful in portfolio management and risk analysis. Many methods for modeling of between-subject covariance matrix Σ_1 are available in the literature, for example, Chang and Tsay (2010), Dellaportas and Pourahmadi (2012), Fan, Liao and Mincheva (2013), but their Σ_1 is assumed to have a specific structure, which may not be true in practice.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

We propose here a novel method for joint modeling of between-subject covariance matrix Σ_1 and within-subject covariance matrix Σ_2, where the panel data are assumed to follow a matrix normal distribution $N_{P,T}(0, \Sigma_1 \otimes \Sigma_2)$. Here the mean is assumed to be zero for simplicity, e.g., the expected return of the stocks is zero, and $\Sigma_1 \otimes \Sigma_2$ is an $PT \times PT$ matrix, implying the spatial-temporal correlations are separable. The model has a nice interpretation since the big $PT \times PT$ covariance matrix can be used to describe the covariances of subjects at various observation time points. In fact, the between-subject covariance matrix Σ_1 and within-subject covariance matrix Σ_2 explain associations/correlations with respect to subject and time space, respectively. Our proposed approach is more flexible since no structure is specified to either Σ_1 or Σ_2, and actually they both are modeled through data-driven methods as outlined in Section 2.

2 Estimation methods

Let Y be a $P \times T$ random matrix where y_{ij} is the jth of T measurements of log-return on the ith of P stocks and let t_{ij} be the time at which y_{ij} is observed. It is assumed that $Y \sim N_{P,T}(0, \Sigma)$, where $\Sigma = \Sigma_1 \otimes \Sigma_2$, implying Σ is completely determined by the between-subject covariance matrix Σ_1 and the within-subject covariance matrix Σ_2 through a Kronecker product. We exploit data-driven techniques for the modeling of the two covariance matrices Σ_1 and Σ_2. For the within-subject covariance matrix Σ_2, modified Cholesky decomposition (M.CD) is applied. While for the between-subject covariance matrix Σ_1, it is modeled by using alternative Cholesky decomposition (A.CD). This strategy is reasonable because when the correlation matrix R is of interest, estimation of the correlation matrix R is robust against the misspecification of models for the diagonal matrix D (Maadooliat and Pourahmadi, 2013).
We propose to model Σ_1 and Σ_2, respectively, due to the following facts (Pan and Fang, 2000)

$$\Sigma_1^{-1/2} Y \sim N_{P,T}(0, I_P, \Sigma_2),$$

$$Y \Sigma_2^{-1/2} \sim N_{P,T}(0, \Sigma_1, I_T).$$

Then an estimation of Σ can be achieved by updating the estimates of Σ_1 and Σ_2 iteratively until convergence.

2.1 A.CD for Σ_1 and M.CD for Σ_2

Given the estimate of Σ_2, we have $Y_1 = Y\Sigma_2^{-1/2} \sim N_{P,T}(0, \Sigma_1, I_T)$, or equivalently, $X_i \sim N_P(0, \Sigma_1)$ where X_i is the ith column of Y_1. Now the between-subject covariance matrix Σ_1 is modeled via alternative Cholesky decomposition (A.CD) as outlined below. Since it is positive definite, there exists a unique lower triangular matrix L with 1's as main diagonal entries and a unique diagonal matrix D with positive diagonal entries such

that $\Sigma_1 = DLL^TD$. The diagonal entries of D^2, innovation variances σ_i^2, can be modeled through $\log \sigma_i^2 = \omega_i^T \lambda$, where ω_i are the covariates related to the ith subject. While the below-diagonal entries of L are the moving average coefficients $\theta_{i,j}$ (Maadooliat and Pourahmadi, 2013). They are unconstrained and can be modeled through $\theta_{i,j} = z_{i,j}^T \gamma$ where $z_{i,j}$ are the covariates including the information of the subjects i and j.
The minus twice log-likelihood, up to a constant, is given by

$$-2l = 2\sum_{t=1}^{T} log|D| + \sum_{t=1}^{T} X_t^T \Sigma_1^{-1} X_t.$$

The following maximum likelihood estimating equations for each elements in γ and λ can be obtained by direct calculations:

$$U(\gamma_r) = tr(TD^{-1}(\sum_{t=1}^{T} X_t X_t^T)(TD^{-1})^T T L_{\lambda_r}) = 0$$

$$U(\lambda_s) = tr((\sum_{t=1}^{T} X_t X_t^T \Sigma_1^{-1} - TI_P)D^{-1} D_{\lambda_s}) = 0$$

where $T = L^{-1}$. Since the solutions satisfy the above equations, the parameters γ and λ can be sequentially solved with one parameter kept fixed in the optimization. Given λ, the estimate $\hat{\gamma}$ can be obtained by solving the equation of $U(\gamma)$., e.g., using Newton-Raphson algorithm. Similarly, when γ is given, the estimate of $\hat{\lambda}$ can be obtained by solving the equation of $U(\lambda)$. Our algorithm starts by initializing $\Sigma_1 = I_P$ and repeats until a pre-specified convergence criteria is met.
Given the estimate of Σ_1, we have $Y_2 = \Sigma_1^{-1/2} Y \sim N_{P,T}(0, I_P, \Sigma_2)$, or equivalently, $\tilde{X}_i \sim N_T(0, \Sigma_2)$ where $\tilde{X}_i$ is the ith row of Y_2. The within-subject covariance matrix Σ_2 can then be modeled via the modified Cholesky decomposition (M.CD) outlined as below. Since it is positive definite, there exists a unique lower triangular matrix T with 1's as main diagonal entries and a unique diagonal matrix D with positive diagonal entries such that $T\Sigma_2 T^T = D^2$ which can also be written as $\Sigma_2 = LD^2L^T$ where $L = T^{-1}$. The below-diagonal entries of T are the negatives of the autoregressive coefficients, $\phi_{i,j}$, which is unconstrained and can be modeled by $\phi_{i,j} = z_{i,j}^T \gamma$. While the diagonal entries of D^2, innovation variance σ_i^2, can be modeled by $\log \sigma_i^2 = h_i^T \lambda$, where h_i are the covariates of the ith subject. The MLEs of λ and γ can be archived in a similar way as before and has been well explained in Pan and Mackenzie (2003).

2.2 Other specifications to Σ_1 and Σ_2

A simple structure for the between-subject covariance matrix Σ_1 is compound symmetry in the form of $\Sigma_1 = DRD$, where the correlation between any two subjects/stocks, i.e., the off-diagonal elements in correlation matrix R, is assumed to be same and denoted by ρ. The standard deviation for

each subject/stock, i.e., diagonal element in the diagonal matrix D, may vary from subject to subject, which can be modeled through $\log \sigma_i^2 = g_i^T \omega$, where g_i are the covariates of the ith subject.
The other specification to the within-covariance matrix Σ_2 is the so-called GARCH(1,1) model, which is one of the commonly used models in modeling volatilities for financial data. Here we apply GARCH(1,1) models to each subject/stock and then take the average in order to obtain the estimates of the parameters ω, α and β involved in GARCH(1,1) model.

3 Data Analysis

TABLE 1. Average of the parameter estimates with simulated standard deviations in parentheses for 500 random samples generated from the matrix normal distribution, where A.CD decomposition is applied to Σ_1 and M.CD decomposition is applied to Σ_2

	True	P=50			P=100			P=200		
	value	T=5	T=10	T=20	T=5	T=10	T=20	T=5	T=10	T=20
λ_1^{acd}	-1.60	-1.621	-1.595	-1.596	-1.611	-1.585	-1.593	-1.612	-1.602	-1.594
		(0.301)	(0.168)	(0.128)	(0.292)	(0.165)	(0.111)	(0.252)	(0.135)	(0.088)
λ_2^{acd}	-4.50	-4.473	-4.494	-4.498	-4.508	-4.492	-4.484	-4.523	-4.503	-4.501
		(0.316)	(0.199)	(0.135)	(0.262)	(0.149)	(0.102)	(0.336)	(0.222)	(0.104)
λ_3^{acd}	-2.30	-2.336	-2.301	-2.288	-2.303	-2.295	-2.302	-2.304	-2.299	-2.291
		(0.265)	(0.167)	(0.118)	(0.282)	(0.175)	(0.129)	(0.269)	(0.199)	(0.094)
γ_1^{acd}	0.51	0.514	0.510	0.510	0.513	0.509	0.511	0.511	0.511	0.507
		(0.075)	(0.044)	(0.053)	(0.065)	(0.036)	(0.024)	(0.056)	(0.027)	(0.027)
γ_2^{acd}	0.47	0.475	0.468	0.471	0.473	0.471	0.469	0.478	0.472	0.473
		(0.095)	(0.059)	(0.040)	(0.063)	(0.040)	(0.028)	(0.069)	(0.040)	(0.038)
γ_3^{acd}	0.33	0.338	0.329	0.329	0.333	0.334	0.333	0.332	0.331	0.331
		(0.078)	(0.049)	(0.036)	(0.064)	(0.045)	(0.040)	(0.046)	(0.029)	(0.022)
λ_1^{mcd}	6.00	5.961	5.978	5.971	5.984	5.965	5.966	6.041	6.007	5.988
		(0.631)	(0.295)	(0.197)	(0.612)	(0.228)	(0.150)	(0.760)	(0.525)	(0.128)
λ_2^{mcd}	-0.50	-0.467	-0.498	-0.501	-0.474	-0.500	-0.496	-0.486	-0.498	-0.500
		(0.336)	(0.097)	(0.032)	(0.237)	(0.073)	(0.023)	(0.185)	(0.053)	(0.016)
λ_3^{mcd}	0.03	0.023	0.029	0.030	0.025	0.029	0.029	0.027	0.029	0.030
		(0.055)	(0.008)	(0.001)	(0.039)	(0.006)	(0.001)	(0.029)	(0.004)	(0.001)
γ_1^{mcd}	0.80	0.798	0.796	0.799	0.785	0.797	0.800	0.799	0.799	0.800
		(0.225)	(0.051)	(0.008)	(0.162)	(0.034)	(0.005)	(0.121)	(0.025)	(0.004)
γ_2^{mcd}	-0.3	-0.300	-0.298	-0.299	-0.286	-0.298	-0.299	-0.299	-0.299	-0.300
		(0.215)	(0.028)	(0.002)	(0.156)	(0.019)	(0.001)	(0.119)	(0.013)	(0.001)
γ_3^{mcd}	0.02	0.020	0.019	0.020	0.017	0.019	0.020	0.019	0.019	0.020
		(0.043)	(0.003)	(0.0002)	(0.031)	(0.002)	(0.0001)	(0.024)	(0.001)	(0.0001)

The results of our simulation studies are presented in Tables 1-3 below, where each result is the average of parameter estimates for 500 simulation runs. Table 1 reports the parameter estimates for the panel data generated from a matrix normal distribution, and the simulated standard errors (in parentheses). In total, nine combinations of the number of subjects

TABLE 2. Average of the parameter estimates with simulated standard deviations (Compound symmetry for Σ_1 and M.CD for Σ_2)

	True value	P=50 T=5	P=50 T=10	P=50 T=20	P=100 T=5	P=100 T=10	P=100 T=20	P=200 T=5	P=200 T=10	P=200 T=20
λ_1^{cs}	-3.20	-3.199 (0.056)	-3.200 (0.040)	-3.201 (0.026)	-3.196 (0.043)	-3.199 (0.031)	-3.200 (0.022)	-3.201 (0.031)	-3.201 (0.027)	-3.195 (0.018)
λ_2^{cs}	-4.70	-4.706 (0.074)	-4.701 (0.052)	-4.701 (0.035)	-4.702 (0.050)	-4.703 (0.034)	-4.703 (0.023)	-4.702 (0.032)	-4.703 (0.022)	-4.701 (0.014)
λ_3^{cs}	-1.70	-1.699 (0.065)	-1.697 (0.047)	-1.699 (0.031)	-1.702 (0.047)	-1.702 (0.032)	-1.701 (0.021)	-1.701 (0.033)	-1.698 (0.025)	-1.701 (0.019)
ρ^{cs}	0.7	0.634 (0.146)	0.672 (0.102)	0.705 (0.067)	0.673 (0.133)	0.714 (0.096)	0.739 (0.070)	0.766 (0.081)	0.779 (0.072)	0.787 (0.062)
λ_1^{mcd}	6.00	5.865 (0.610)	5.952 (0.366)	6.033 (0.263)	5.967 (0.489)	6.092 (0.347)	6.173 (0.293)	6.307 (0.394)	6.207 (0.225)	6.188 (0.128)
λ_2^{mcd}	-0.50	-0.494 (0.355)	-0.500 (0.096)	-0.498 (0.033)	-0.496 (0.226)	-0.496 (0.069)	-0.498 (0.023)	-0.497 (0.169)	-0.498 (0.054)	-0.500 (0.018)
λ_3^{mcd}	0.03	0.027 (0.058)	0.029 (0.008)	0.029 (0.001)	0.029 (0.037)	0.029 (0.006)	0.030 (0.001)	0.028 (0.027)	0.029 (0.004)	0.030 (0.001)
γ_1^{mcd}	0.80	0.794 (0.210)	0.795 (0.049)	0.799 (0.008)	0.800 (0.158)	0.798 (0.032)	0.800 (0.005)	0.795 (0.107)	0.799 (0.026)	0.798 (0.005)
γ_2^{mcd}	-0.3	-0.296 (0.200)	-0.297 (0.027)	-0.299 (0.002)	-0.301 (0.150)	-0.299 (0.017)	-0.299 (0.002)	-0.296 (0.102)	-0.298 (0.015)	-0.299 (0.001)
γ_3^{mcd}	0.02	0.019 (0.040)	0.019 (0.003)	0.020 (0.0002)	0.020 (0.030)	0.019 (0.002)	0.020 (0.0001)	0.019 (0.021)	0.019 (0.001)	0.020 (0.0001)

TABLE 3. Average of the parameter estimates with simulated standard deviations (Compound symmetry and A.CD for Σ_1 and GARCH(1,1) for Σ_2)

	True value	T=1000 P=50	T=1000 P=100	T=1000 P=200		True value	T=1000 P=50	T=1000 P=100	T=1000 P=200
λ_1^{cs}	-3.20	-3.201 (0.024)	-3.200 (0.018)	-3.201 (0.011)	λ_1^{acd}	-1.6	-3.196 (0.092)	-3.199 (0.065)	-3.201 (0.031)
λ_2^{cs}	-4.70	-4.703 (0.031)	-4.701 (0.019)	-4.702 (0.013)	λ_2^{acd}	-4.5	-4.508 (0.062)	-4.492 (0.049)	-4.504 (0.032)
λ_3^{cs}	-1.70	-1.699 (0.026)	-1.701 (0.024)	-1.698 (0.018)	λ_3^{acd}	-2.3	-2.303 (0.082)	-2.296 (0.075)	-2.301 (0.029)
ρ^{cs}	0.7	0.706 (0.075)	0.713 (0.064)	0.717 (0.063)	γ_1^{acd}	0.51	0.510 (0.065)	0.511 (0.064)	0.508 (0.034)
					γ_2^{acd}	0.47	0.478 (0.061)	0.472 (0.022)	0.473 (0.015)
					γ_3^{acd}	0.33	0.329 (0.038)	0.331 (0.025)	0.333 (0.013)
ω^{garch} ($\times 10^{-5}$)	0.80	0.798 (0.065)	0.796 (0.051)	0.799 (0.038)	ω^{garch} ($\times 10^{-5}$)	0.80	0.785 (0.062)	0.797 (0.034)	0.800 (0.025)
α^{garch}	0.121	0.116 (0.015)	0.124 (0.013)	0.125 (0.009)	α^{garch}	0.121	0.115 (0.015)	0.131 (0.019)	0.117 (0.011)
β^{garch}	0.852	0.820 (0.043)	0.819 (0.055)	0.870 (0.036)	β^{garch}	0.852	0.817 (0.052)	0.879 (0.042)	0.820 (0.061)

P and number of observations T are considered ($P = 50, 100, 200$ and $T = 5, 10, 20$). From Table 1, it is clear that the resulting parameter estimators are quite close to their true values, indicating the proposed method works very well. It is also clear that the simulated standard deviation is smaller for the cases with large T and fixed P, or large P and fixed T. Similar conclusions can also be drawn from Tables 2-3. Note that the estimation results are improved when increasing either the number of subjects P or the number of observations T. The proposed approach performs well as demonstrated by our simulation studies.

References

Chen, Dunson (2003). Random effects selection in linear mixed models. *Biometrics*, **59**, 762 – 769.

Chang, Tsay (2010). Estimation of covariance matrix via the sparse Cholesky factor with lasso. *Journal of Statistical Planning and Inference*, **140**, 3858 – 3873.

Dellaportas, Pourahmadi (2012). Cholesky-GARCH models with applications to finance. *Statistics and Computing*, **22**, 849 – 855.

Fan, Liao, Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 603 – 680.

Maadooliat, Pourahmadi (2013). Robust estimation of the correlation matrix of longitudinal data. *Statistics and Computing*, **23**, 17 – 28.

Pan, Fang (2000). *Growth Curve Models and Statistical Diagnostics*. Spring.

Pan, Mackenzie (2003). On modelling meancovariance structures in longitudinal studies. *Biometrika*, **90**, 239 – 244.

Pourahmadi (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, **86**, 667 – 690.

Bayesian Effect Fusion for Categorial and Ordinal Predictors

Daniela Pauger[1], Helga Wagner[1]

[1] Johannes Kepler University Linz, Austria

E-mail for correspondence: daniela.pauger@jku.at

Abstract: We propose sparse Bayesian modelling of the effects of categorial, i.e. nominal and ordinal covariates in regression type models. Sparsity is achieved by specifying spike and slab prior distributions and posterior inference relies on MCMC methods. For illustration we analyse Austrian data from EU-SILC 2010, where we model the effects of social and demographic characteristics on annual (personal) income.

Keywords: Spike and slab prior distribution; sparse modelling; dummy coding; EU-SILC; income

1 Introduction

In regression type models often many of the collected variables are categorial, measured on an ordinal or nominal scale. The usual modelling strategy is to use one of the levels as baseline and to define dummy variables for the other levels. This can easily lead to a high-dimensional vector of regression effects. A sparse representation of the model can be achieved by fusing category levels with essentially the same effect into one category and by removing variables where none of the levels has a non-zero effect. As an example we analyse the effects of social and demographic characteristics, such as age or educational level, on annual personal income. For this analysis we use Austrian data from the Survey on Income and Living Conditions (SILC) in 2010.

2 Model Specification

Let y denote the normal response in a standard linear regression model with $j = 1, \ldots, p$ categorial covariates c_j. We assume that the j-th covariate has

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

$K_j + 1$ categories $0, \ldots, K_j$ and the first category 0 defines the reference category. We specify the linear regression model as

$$y = \mu + \sum_{j=1}^{p} \sum_{k=1}^{K_j} X_{jk} \theta_{j,k0} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

with regressors X_{jk} defined as in Gertheiss and Tutz (2009): We use split coding for ordinal covariates, i.e.

$$X_{jk} = \begin{cases} 1, & \text{for } c_j \geq k \\ 0, & \text{otherwise,} \end{cases} \qquad k = 1, \ldots K_j$$

and usual dummy coding for nominal covariates.
For the nominal covariates the regression effects corresponding to the K_j dummy variables can be interpreted as the effect contrast of category k and the reference category. To allow for fusion of level effects we define for all $k > l$ by $\theta_{j,kl}$ the effect contrast of categories k and l of covariate c_j, which leads to the restriction

$$\theta_{j,k0} - \theta_{j,l0} - \theta_{j,kl} = 0, \qquad \text{for all } 0 < l < k \leq K_j.$$

We subsume all parameters $\theta_{j,kl}$ with $0 \leq l < k \leq K_j$ in the vector $\boldsymbol{\theta}_j$.

3 Priors and Bayesian inference

We assign a flat proper prior to the intercept, $\mu \sim \mathcal{N}(0, M_0\sigma^2)$ and the improper prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$ to the error variance.
To encourage sparsity in the coefficient vectors we specify for each element of $\boldsymbol{\theta}_j$ a spike and slab prior distribution (George and McCulloch, 1993) hierarchically as

$$p(\theta_{j,kl}|\delta_{j,kl}, \tau^2) \sim \delta_{j,kl}\mathcal{N}(0, \tau^2) + (1 - \delta_{j,kl})\mathcal{N}(0, r\tau^2),$$

where r is a small value and $\delta_{j,kl}$ is an indicator for the slab component with prior distribution $p(\delta_{j,kl} = 1) = w_j$. w_j corresponds to the weight of variable j and is assigned a Beta hyperprior, $w_j \sim \mathcal{B}(a_{0j}, b_{0j})$. Effect fusion is accomplished by the spike component: If $\delta_{j,kl} = 0$, the effect $\theta_{j,kl}$ is assigned to the spike component and hence shrunk to zero. Thus the corresponding level effects $\theta_{j,k0}$ and $\theta_{j,l0}$ are fused.
Bayesian inference is accomplished by sampling from the posterior distribution using MCMC methods. To guarantee the restriction on the parameters $\boldsymbol{\theta}_j$ we use the kriging algorithm described in Rue and Held (2005). Posterior means of the indicators $\delta_{j,kl}$ suggest which levels of variable c_j can be fused and which variables can be completely removed from the model.

4 Modelling Income in Austria

4.1 Data

We use data from EU-SILC (SILC = Survey on Income and Living Conditions) in 2010 to model personal income of full-time employees in Austria. The data set provides a wide range of variables on financial and living aspects of households as well as demographic characteristics of individuals. In our analysis we model the logarithm of the annual income and use sex, age (grouped), Austrian federal state of residence, citizenship and highest education achieved as potential regressors. As we restrict the analysis to full-time employees with no missing values in the covariates, the final data set comprises 4,029 subjects.

4.2 Results

Table 1 compares the results of a Bayesian analysis of the unrestricted model, where a flat normal prior is assigned to all regression coefficients, to the model averaged results using spike and slab priors. The variance parameter τ^2 is set to 1 and the parameters for the hyperprior on the weights w_j are $a_{0j} = b_{0j} = 1$.
In general, it can be seen that the income of female employees in Austria is lower than the income of their male colleagues. Age has a positive effect on the income, but the level effects are very similar for age categories above 51 years in the unrestricted model. Under the spike and slab prior distribution these categories are fused. Income varies between the federal states in Austria, but the differences are very small. Using sparse modelling, all category effects, and hence the whole variable, are excluded from the model. Employees with Austrian citizenship have a higher income than others. The negative effects of a citizenship from a state of former Yugoslavia (without Slovenia) or the 'New EU10' (Estonia, Latvia, Lithuania, Malta, Poland, Slovakia, Slovenia, Czech Republic, Hungary, Cyprus) are very similar and fused to one group under the spike and slab prior.
An educational level higher than secondary school generally has a positive effect on personal income. The effect increases with educational level and posterior means are almost identical for categories (1) and (2) as well as categories (3), (5), (6) and (7) under the spike and slab prior. Posterior means and 90% HPD intervals of the level effects are compared in Figure 1 for a flat prior and the spike and slab prior. Obviously, also HPD intervals of the effects of categories (1) and (2) on the one hand and (3), (5), (6) and (7) on the other hand are almost identical under the spike and slab prior. As fused categories will have the same posterior distribution, this suggests that the effect of education on income could be represented by only four levels. The corresponding HPD intervals are shorter than those under the flat prior. However, this is not generally the case, as 90% HPD intervals of the effects of categories (4) and (9) are even longer than under the flat prior.

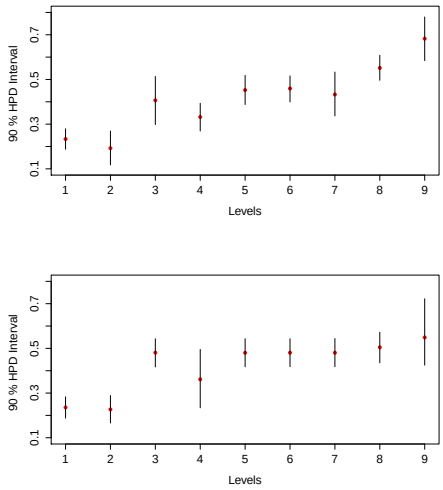


FIGURE 1. Posterior mean and 90% HPD intervals for level effects of highest education. Left plot: flat prior. Right plot: spike and slab prior.

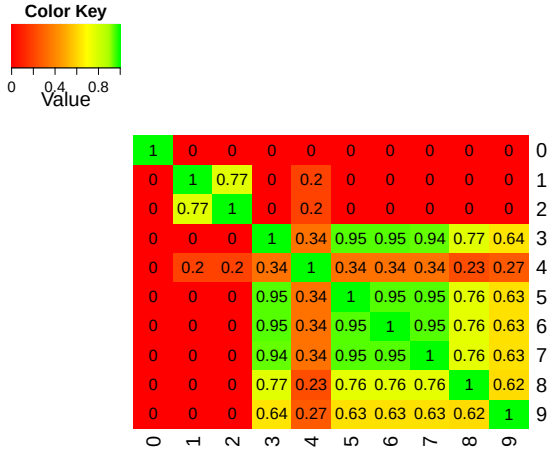


FIGURE 2. Heat map of fusion probabilities for categories of highest education

More insight into effect fusion is given in Figure 2 which shows a heat map of fusion frequencies in the MCMC iterations under the spike and slab prior. In 77% of the iterations the effects of categories (1) and (2), and in almost 95% those of the four categories (3), (5), (6) and (7) are fused. Results are not so clear for the remaining categories, since effects of categories (8) and (9) are fused with those of the four categories (3) and (5)–(7) in roughly 76% and 63%, and with that of category (4) in 23% and 27% of the iterations, respectively. Figure 3 shows the trace plots of

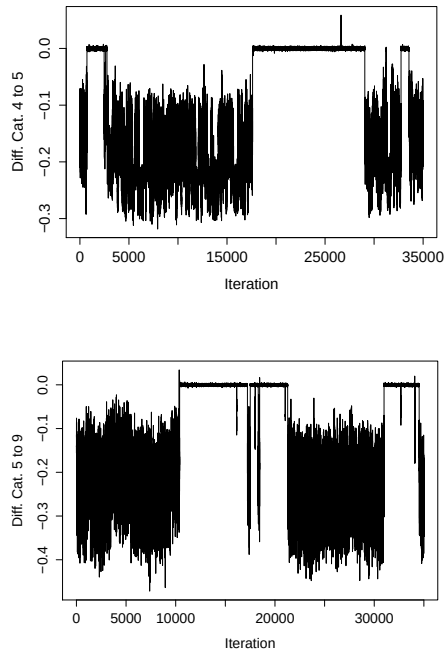


FIGURE 3. Trace plot of effect contrasts. Left plot: category (5) to (4). Right plot: category (5) to (9).

the effect contrasts of categories (5) and (4) (left plot) and categories (5) and (9) (right plot) which further illustrate that both differences are set to zero, and hence the corresponding effects are fused, in some, but not all iterations.

In addition to the category fusion probabilities shown in the heat map, more detailed information on models with high posterior probability might be of interest. Posterior model probabilities can be estimated by the number of visits to each model during MCMC. The two top models (each with estimated marginal posterior probability 0.12) both fuse the effect of category (9) with those of categories (3) and (5)–(8). In one of these models the effect of category (4) is also fused to this effect, in the other it is not fused with any category effect at all. The model with third highest posterior probability (0.08) fuses the effect of category (4) with that of categories (1) and (2). While in the three top models, which together account for roughly one third posterior model probability, the effect of category (9) is fused with other level effects, it is not fused at all in several models visited less often. This uncertainty on fusion is reflected in the larger posterior variance of the effects of categories (4) and (9).

5 Conclusion

We propose a method for sparse modelling of ordered and unordered categorial covariates by effect fusion and removing of (almost) non-zero effects. Bayesian inference is done using MCMC methods, a spike and slab prior distributions on the regression effects encourage sparsity.

Acknowledgments: We acknowledge greatfully financial support by the Austrian Science Fund FWF, projekt number P25850 'Sparse Bayesian modelling for categorial predictors'.

TABLE 1. Estimation results

variable	flat prior	spike and slab prior	incl. prob.
intercept	8.94	8.90	–
women	-0.22	-0.22	1.00
age (base:16 to 20 years)			
21 to 30 years	0.67	0.65	1.00
31 to 40 years	0.87	0.86	1.00
41 to 50 years	0.96	0.97	1.00
51 to 60 years	1.01	0.98	1.00
over 60 years	1.00	0.98	1.00
federal state (base: Upper Austria)			
Carinthia	-0.05	0.00	0
Lower Austria	-0.05	0.00	0
Burgenland	-0.09	0.00	0
Salzburg	-0.02	0.00	0
Styria	-0.11	0.00	0
Tyrol	-0.04	0.00	0
Vorarlberg	0.12	0.00	0
Vienna	-0.03	0.00	0
citizenship (base: Austria)			
EU15/EFTA	-0.07	-0.02	0.13
New EU10	-0.25	-0.19	1.00
Rest of Yugoslavia without Slovenia	-0.19	-0.19	1.00
Turkey	-0.19	-0.03	0.16
Others	-0.19	-0.07	0.38
Highest education achieved			
(base: max. secondary school degree)			
(1) apprenticeship, trainee	0.23	0.24	1.00
(2) master craftman's diploma	0.19	0.23	1.00
(3) nurse's training school	0.41	0.48	1.00
(4) other vocational school (medium level)	0.33	0.36	1.00
(5) academic secondary school (upper level)	0.45	0.48	1.00
(6) college for higher vocational education	0.46	0.48	1.00
(7) vocational school for apprentices	0.43	0.48	1.00
(8) university, academy, FH: first degree	0.55	0.50	1.00
(9) university: doctoral studies	0.68	0.55	1.00

References

George, I. and McCulloch, R.E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, Vol.88, No.423, 881 – 889.

Gertheiss, J. and Tutz, G. (2009). Penalized regression with ordinal predictors. *International Statistical Review*, 345 – 365.

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

Partitioned conditional generalized linear models for categorical data

Jean Peyhardi[1,2], Catherine Trottier[1], Yann Guédon[2]

[1] Université Montpellier 2, I3M, Montpellier, France
[2] CIRAD, UMR AGAP and Inria, Virtual Plants, Montpellier, France

E-mail for correspondence: jean.peyhardi@math.univ-montp2.fr

Abstract: In categorical data analysis, several regression models have been proposed for hierarchically-structured response variables, such as the nested logit model. But they have been formally defined for only two or three levels in the hierarchy. Here, we introduce the class of partitioned conditional generalized linear models (PCGLMs) defined for an arbitrary number of levels. The hierarchical structure of these models is fully specified by a partition tree of categories. Using the genericity of the (r, F, Z) specification of GLMs for categorical data, PCGLMs can handle nominal, ordinal but also partially-ordered response variables.

Keywords: hierarchically-structured categorical variable; partition tree; partially-ordered variable; GLM specification.

1 *(r,F,Z)* specification of GLM for categorical data

The triplet (r, F, Z) will play a key role in the following since each GLM for categorical data can be specified using this triplet; see Peyhardi et al. (2013) for more details. The definition of a GLM includes the specification of a link function g which is a diffeomorphism from $\mathcal{M} = \{\pi \in]0,1[^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\}$ to an open subset $\mathcal{S}$ of $\mathbb{R}^{J-1}$. This function links the expectation $\pi = E[Y|X{=}x]$ and the linear predictor $\eta = (\eta_1, ..., \eta_{J-1})^t$. It also includes the parametrization of the linear predictor η, which can be written as the product of the design matrix Z (as a function of x) and the vector of parameters β. All the classical link functions $g = (g_1, \ldots, g_{J-1})$, rely on the same structure which we propose to write as

$$g_j = F^{-1} \circ r_j, \quad j = 1, \ldots, J-1. \tag{1}$$

where F is a continuous and strictly increasing cumulative distribution function (cdf) and $r = (r_1, \ldots, r_{J-1})^t$ is a diffeomorphism from $\mathcal{M}$ to an

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

open subset $\mathcal{P}$ of $]0,1[^{J-1}$. Finally, given x, we propose to summarize a GLM for a categorical response variable by the $J-1$ equations

$$r(\pi) = \mathcal{F}(Z\beta),$$

where $\mathcal{F}(\eta) = (F(\eta_1), \ldots, F(\eta_{J-1}))^T$. In the following we will consider four particular ratios. The *adjacent*, *sequential* and *cumulative* ratios respectively defined by $\pi_j/(\pi_j + \pi_{j+1})$, $\pi_j/(\pi_j + \ldots + \pi_J)$ and $\pi_1 + \ldots + \pi_j$ for $j = 1, \ldots, J-1$, assume order among categories but with different interpretations. We introduce the *reference* ratio, defined by $\pi_j/(\pi_j + \pi_J)$ for $j = 1, \ldots, J-1$, useful for nominal response variables.
Finally, a single estimation procedure based on Fisher's scoring algorithm can be applied to all the GLMs specified by an (r, F, Z) triplet. The score function can be decomposed into two parts, where only the first one depends on the (r, F, Z) triplet.

$$\frac{\partial l}{\partial \beta} = \underbrace{Z^T \frac{\partial \mathcal{F}}{\partial \eta} \frac{\partial \pi}{\partial r}}_{(r,F,Z) \text{ dependent part}} \underbrace{\mathrm{Cov}(Y|X=x)^{-1}\,[y-\pi]}_{(r,F,Z) \text{ independent part}}. \tag{2}$$

We need only to evaluate the density function $\{f(\eta_j)\}_{j=1,\ldots,J-1}$ to compute the corresponding diagonal Jacobian matrix $\partial\mathcal{F}/\partial\eta$. For details on computation of the Jacobian matrix $\partial\pi/\partial r$ according to each ratio, see Peyhardi (2013).

2 Partitioned conditional GLMs

The main idea consists in recursively partitioning the J categories then specifying a conditional GLM at each step. This type of model is therefore referred to as partitioned conditional GLM. Such models have already been proposed, such as the nested logit model (McFadden, 1978), the two-step model (Tutz, 1989) and the partitioned conditional model for partially-ordered set (POS-PCM) (Zhang and Ip, 2012). Our proposal can be seen as a generalization of these three models that benefits from the genericity of the (r, F, Z) specification. In particular, our objective is not only to propose GLMs for partially-ordered response variables but also to differentiate the role of explanatory variables for each partitioning step using specific explanatory variables and design matrices. We are also seeking to formally define partitioned conditional GLMs for an arbitrary number of levels in the hierarchy.

PCGLM definition: Let $J \geq 2$ and $1 \leq k \leq J-1$. A ***k*-partitioned conditional GLM** for categories $1, \ldots, J$ is defined by:

- A **partition tree $\boldsymbol{\mathcal{T}}$** of $\{1, \ldots, J\}$ with $\mathcal{V}^*$, the set of non-terminal vertices of cardinal $\boldsymbol{k}$. Let Ω_j^v be the children of vertex $v \in \mathcal{V}^*$.
- A **collection of models $\boldsymbol{\mathfrak{C} = \{(r^v, F^v, Z^v(x^v)) \mid v \in \mathcal{V}^*\}}$** for each conditional probability vector $\pi^v = (\pi_1^v, \ldots, \pi_{J_v-1}^v)$, where $\pi_j^v =$

$P(Y \in \Omega_j^v | Y \in v; x^v)$ and x^v is a sub-vector of x associated with vertex v.

PCGLM estimation: Using the partitioned conditional structure of model, the log-likelihood can be decomposed as $l = \sum_{v \in \mathcal{V}^*} l^v$, where l^v represents the log-likelihood of GLM $(r^v, F^v, Z^v(x^v))$. Each component l^v can be maximised individually, using (2), if all parameters $\{\beta^v\}_{v \in \mathcal{V}^*}$ are different.

PCGLM selection: The partition tree $\mathcal{T}$ and the collection of models $\mathcal{C}$ have to be selected using ordering assumption among categories.

- *Nominal data*: the partition tree $\mathcal{T}$ is built by aggregating similar categories - such as the nested logit model of McFadden (1978) - and $\mathcal{C}$ contains only reference models, appropriate for nominal data; see Peyhardi (2013).
- *Ordinal data*: we propose to adapt the Anderson's indistinguishability procedure (1984) for PCGLM selection.
- *Partially ordered data*: the partial ordering assumption among categories can be summarized by an Hasse diagram. Zhang and Ip (2012) defined an algorithm to build the partition tree $\mathcal{T}$ automatically from the Hasse diagram; see figure 1 with the pear tree dataset. It should be remarked that every partially-ordered variable Y can be expressed in terms of elementary ordinal and nominal variables $\tilde{Y}_i$ (with at least one ordinal variable). We propose to build the partition tree $\mathcal{T}$ directly from these latent variables $\tilde{Y}_i$ to obtain a more interpretable structure. For these two methods of partition tree building, the main idea is to recursively partition the J categories in order to use a simple (ordinal or nominal) GLM at each step.

3 Application to pear tree dataset

Dataset description: In winter 2001, the first annual shoot of 50 one-year-old trees was described by node. The presence of an immediate axillary shoot was noted at each successive node. Immediate shoots were classified into four categories according to their length and transformation or not of the apex into spine (i.e. definite growth or not). The final dataset was thus constituted of 50 bivariate sequences of cumulative length 3285 combining a categorical variable Y (type of axillary production selected from among latent bud (l), unspiny short shoot (u), unspiny long shoot (U), spiny short shoot (s) and spiny long shoot (S)) with an interval-scaled variable X_1 (internode length).

Results: A higher likelihood and simpler interpretations were obtained using partial ordering information. The axillary production Y of pear tree can be decomposed into two levels. Production first follows a sequential mechanism (ordinal model), giving latent bud, short shoot or long shoot

(first level of hierarchy; figure 1), which is strongly influenced by the internode length X_1 (the longer the internode, the longer the axillary shoot). The axillary shoot apex then differentiates or not into spine (second level of hierarchy; figure 1) depending on distance to growth unit end (second explanatory variable X_2 expressed in number of nodes).

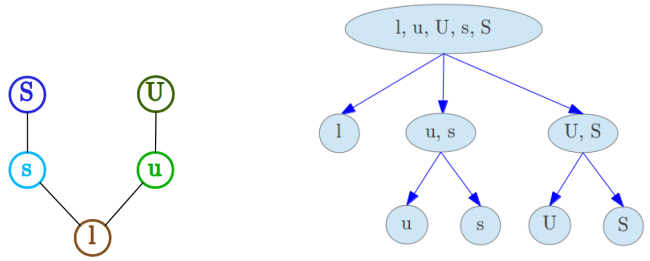


FIGURE 1. Hasse diagram and corresponding partition tree.

References

Anderson, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, **46**, 1 – 30.

McFadden, D. et al. (1978). Modelling the choice of residential location. *Institute of Transportation Studies, University of California.*

Peyhardi, J. (2013). A new GLM framework for analysing categorical data; application to plant structure and development. *PhD thesis.*

Peyhardi, J., Trottier, C. and Guédon, Y. (2013). A unifying framework for specifying generalized linear models for categorical data. *In 28th International Workshop on Statistical Modeling*, 331 – 335

Tutz, G. (1989). Compound regression models for ordered categorical data. *Biometrical Journal*, **31**, 259 – 272.

Zhang, Q. and Ip, E.H. (2012). Generalized linear model for partially ordered data. *Statistics in Medicine.*

Nodewise graphical modeling using the Focused Information Criterion for 'p larger than n' settings.

Eugen Pircalabelu[1], Gerda Claeskens[1], Sara Jahfari[2], Lourens Waldorp[3]

[1] KU Leuven, ORSTAT and Leuven Statistics Research Center, Belgium
[2] University of Amsterdam, Cognitive Science Center, The Netherlands
[3] University of Amsterdam, Department of Psychological Methods, The Netherlands

E-mail for correspondence: eugen.pircalabelu@kuleuven.be

Abstract: We present a new method of estimating graphical models with the clear goal of providing models that minimize the mean squared error of certain parameters of interest to the researcher. The method is applicable to undirected as well as to mixed graphs containing both directed and undirected edges. Quadratic approximations to several well studied penalties deal with problems where the number of nodes is greater than the number of samples. Extensions of the current application include a dynamical image of graphs based on consecutive focus points and estimating graphs where information is borrowed among subjects.

Keywords: Undirected Markov networks; Temporal chain graphs; Focused information criterion; Model selection; fMRI

1 Introduction and motivation

Motivated by a resting-state fMRI study and by the definitory characteristics of such data (a complete dataset of n repeated measurements for each of p brain regions per subject) we propose a new method for model selection for undirected as well as mixed graphical models (with both directed and undirected edges), in situations where the number of nodes in the graph is larger than the number of cases. The method is constructed to have good mean squared error properties and it is based on the focused information criterion (FIC) developed in Claeskens and Hjort (2003).
The reason for modeling the data in this way is two-fold: first, the graphical representation offers insight into the complex relations that might exist between the nodes (brain regions in this case) revealing thus patterns of

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

functional connectivity within the brain (Bullmore and Sporns, 2009). Second, we wish to accommodate distinct research objectives and select models tailored to those interests. Unlike traditional model selection criteria (such as AIC or BIC), the FIC allows to selecting individual models, tailored to a specific research purpose (the *focus*), as opposed to attempting an identification of a single model that should be used for all purposes.

2 Method description

In the context of graphical modeling (see Lauritzen, 1996), given multivariate data, the goal is to estimate plausible positions of edges in the graph, or equivalently conditional independencies between variables. A temporal chain graph $G(E, V)$, includes both directed and undirected edges for which the multivariate random vector Y at time t follows

$$Y_t|Y_{t-1} \sim N(\Gamma Y_{t-1}, \Sigma).$$

A directed edge between nodes i and j belongs to E if $\Gamma_{ij} \neq 0$ and an undirected one is placed if $\Sigma^{-1}_{ij} \neq 0$.
Working under a nodewise local misspecification framework assumes that the true model is in a 'neighborhood' of the least complex model one is willing to assume, that is for a particular node Y, given a sample of n cases, it assumes that

$$Y_k \text{ has density } f(y_k|w_k, z_k, \theta, \gamma_0 + \delta/\sqrt{n}),$$

where f is two times continuously differentiable in a neighborhood of the vector (θ_0, γ_0). We define a *focus parameter*, i.e. $\mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$ as a function of the parameters of the model density, and potentially of a user-specified vector of covariate values, for any particular node in the graph $G(E, V)$. The vector θ corresponds to all parameters that are included in every model (if a node X is known a priori to influence Y, then its effect is included in θ, and we call it a protected node) while γ collects all parameters corresponding to the potential influential set of nodes. W and Z denote the sets of protected and unprotected nodes.
An estimator for (θ, γ) is obtained by optimizing

$$Q(\theta, \gamma) = \frac{1}{n}\sum_{k=1}^{n} \log f(y_k|w_k, z_k, \theta, \gamma) - \frac{\lambda}{n}\sum_{j=1}^{d_\gamma} \psi(|\gamma_j - \gamma_{j0}|), \qquad (1)$$

with respect to θ and γ for a given penalty function ψ (that is twice differentiable in 0) and that depends on an external value λ.
Subsequent steps involve specifying a collection of models S (based on which some γ elements are set at 0, while others are freely estimated) and (1) is optimized for parameters corresponding to the prespecified models. For $n \to \infty$ the quantity $\sqrt{n}(\hat{\mu}_S - \mu_{true}) \xrightarrow{d} \Lambda_S$ for which Λ_S is a normal random variable with a certain mean and variance.

Adding squared bias and variance, we immediately obtain the MSE expression of the estimator $\hat{\mu}_S$.
The FIC estimates the MSE for each of a collection of models S at each node, and selects the model with the smallest MSE value. A nodewise decomposable FIC score pertaining to the entire graph is then constructed as:

$$\text{FIC}(G(E_S, V)) = \sum_{l=1}^{p} \widehat{\text{MSE}}(\hat{\mu}_{l;S_l}).$$

Since our above argumentation was nodewise related, we construct an estimated graph as follows: at each node both the contemporaneous and dynamic effects of other nodes are used in order to determine a low-MSE model, and none of them is on a priori grounds protected. Once all nodewise models are selected (some might include only contemporaneous or dynamic effects, while others might contain both) we apply the following 'OR' rule adapted from Meinshausen and Bühlmann (2006):

$$\begin{aligned}
\hat{E}^{\lambda,\text{OR}}_{i-j} &= \left\{(i,j) \cup (j,i) : i_t \in \hat{ne}^{\lambda}_{j_t} \text{ OR } j_t \in \hat{ne}^{\lambda}_{i_t}\right\} \\
\hat{E}^{\lambda,\text{OR}}_{i\to j} &= \left\{(i,j) : i_{t-1} \in \hat{ne}^{\lambda}_{j_t}\right\} \\
\hat{E}^{\lambda,\text{OR}} &= \left\{\hat{E}^{\lambda,\text{OR}}_{i-j} \cup \hat{E}^{\lambda,\text{OR}}_{i\to j}\right\},
\end{aligned}$$

where $\hat{ne}^{\lambda}$ denotes the neighborhood of the considered node for a certain value of λ. The focus of the research (i.e. the purpose of the model), directs the selection and different focuses may lead to different selected models. In this way, one can obtain better selected models in terms of MSE, than obtained from a global model search. For this application of FIC for graphs, the focus is the expected value of a variable, reflecting interest in discovering a topology of the graph that produces a low MSE for this focus. To deal with situations where $n < p$ a quadratic approximation to penalties such ℓ_1, 'SCAD' or 'Bridge' is applied on the γ vector. We propose further the selection of the regularization level as the quantity optimizing the MSE expression as

$$\lambda_S = \frac{\omega^t G_S \delta 1_q^t (J^{11,S,0})^t \omega - \omega^t \delta 1_q^t (J^{11,S,0})^t \omega}{\omega^t J^{11,S,0} 1_q 1_q^t (J^{11,S,0})^t \omega} \frac{\sqrt{n}}{\psi''(0)},$$

where G_S and $J^{11,S,0}$ can easily be obtained from the Fisher information matrix, while ω involves both the derivatives of the focus parameter and parts of the Fisher information matrix and $\psi''(0)$ is the second derivative of the penalty function evaluated at 0.

3 fMRI data application

Figure 1 illustrates that for a new subject, for which certain brain regions (colored in red) have a large signal compared to the average, the FIC performs well in discovering the activated regions. Moreover, studying the

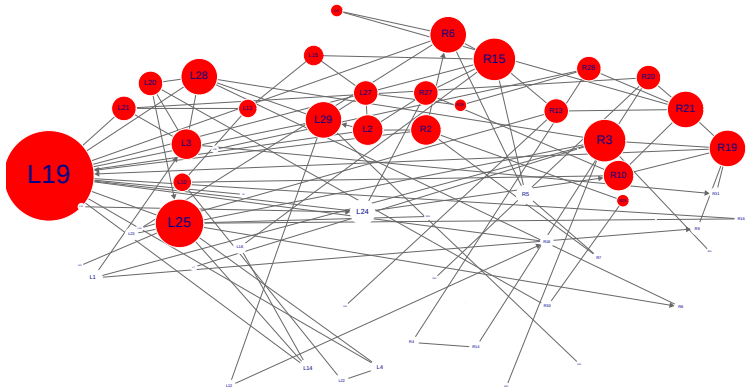


FIGURE 1. fMRI data. FIC graph based on ℓ_2 penalty. Undirected edges denote contemporaneous effects between brain regions while directed edges denote autoregressive dynamic effects. L/R denote the left/right hemisphere and larger labels correspond to high-degree regions.

evolution of the network over time by considering each time the measured signal as a focus, one can conclude that, for example, based on Figure 2 the small-worldness feature of the network which quite often is an interesting aspect of the network to look at, seems to be implausible for the two subjects in the analysis. Most of the estimated values seem to be below a threshold line at 1 and only in a relatively small number of cases the estimated values are larger than the cut-off value. Interestingly, it seems that for subject 1 the estimated networks at later time points produce larger values, while for subject 2 the networks estimated in the first and third part of the measurements produce larger index values.

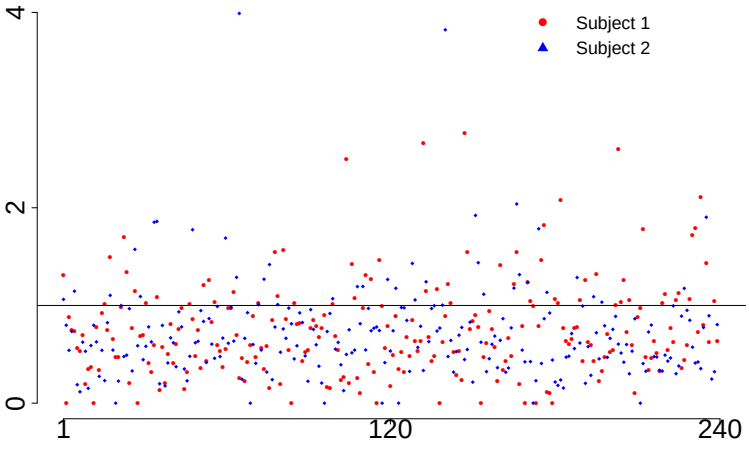


FIGURE 2. fMRI data. Estimated Small-World indexes for two subjects. At each time point, for each of the subjects, a graph based on the ℓ_2 penalty is estimated and based on the graph structure, the Small-World index is computed.

The basic method is then expanded towards three types of situations:

- where one wishes to bring in all information from all the subjects in the analysis and construct one estimated network;
- where one wishes to construct related networks for subjects by allowing for dependence between the measurements for one region of interest across multiple subjects;
- or where one is pooling all data and wishes to estimate networks taking into account subject specific effects.

4 Simulation study

To show the method's performance in a controlled study, we have generated data from 42 different settings using various sample sizes, number of nodes or true graph model and inspected the empirical MSE, sparsity index and a 'structure closeness' score F_1 that measures how close the estimated graph is to the true graph for 3 different focuses. Figure 3 presents averages of the empirical MSE value when the data obtained from 42 different simulation settings each with 300 simulation runs is pooled together. The FIC procedures with three types of penalties provide low MSE with the ℓ_2 penalty being the best performing one.
Depending on the focus and the simulation setting sometimes the FIC provides sparser graphs than competitors (sometimes denser) and sometimes the graphs estimated by the FIC are closer to the true data generating process, sometimes further away, but most importantly the estimated graphs perform well with respect to MSE as intended from construction.

5 Discussion

Using the FIC one has at disposal a powerful method to study evolutions over time of networks constructed to study the functional connectivity between brain regions. The method clearly identifies important brain regions which seem to be highly connected with others, acting as 'hubs' or 'informational gateways' and performs well on simulated data.

Acknowledgments: This research was partially supported by KU Leuven grant GOA FlexStatRob.

References

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, **10(3)**, 186 – 198.

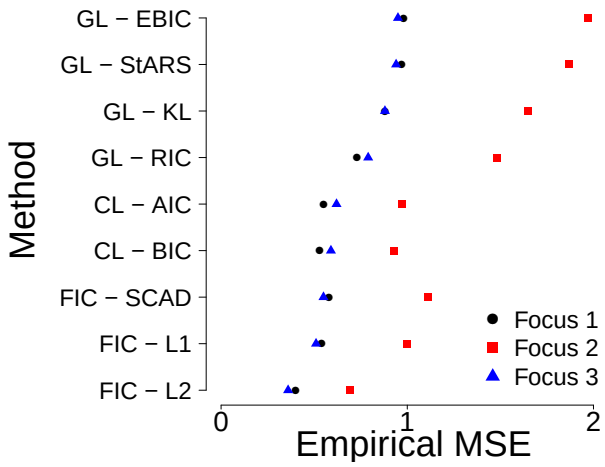


FIGURE 3. Simulated data. Averages of empirical MSE estimated across 300 simulation runs and 42 different settings for three focus points.

Claeskens, G. and Hjort, N. (2003). The focused information criterion (with discussion and a rejoinder by the authors). *JASA*, **98**, 900 – 916.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, **34(3)**, 1436 – 1462.

Regularization and Selection of Proportional Versus Nonproportional Effects in Sequential Logit Models

Wolfgang Pößnecker[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-University Munich, Germany

E-mail for correspondence: Wolfgang.Poessnecker@stat.uni-muenchen.de

Abstract: The sequential logit model for ordinal response is an important tool for the modelling of disease progression and discrete survival. A common issue is the problem of effect type selection: if the available predictor variables are specified with a constant effect across response categories, one obtains a rather inflexible model that is parsimonious and straightforward to interpret. Using category-specific effects leads to a flexible, but high-dimensional model that is difficult to handle. With reference to the effects of these parameterizations, they are also called "proportional" and "nonproportional" effects. In the literature, most authors decide on one type of effect and assume a priori that all covariate effects are of the chosen type. We investigate the downside of this status quo and develop a penalized likelihood approach based on a grouped fused lasso penalty that enables us to seperately choose between proportional and nonproportional effects for each covariate in an automatic, data-driven way. Moreover, our method provides a continuous transition from nonproportional to proportional effects. The usefulness of our approach is illustrated on data about the time to bankruptcy of newly founded companies.

Keywords: Ordinal regression; Discrete Survival Analysis; Regularization; Group Fused Lasso.

1 The Sequential Logit Model

Let $Y \in \{1, 2, \ldots, k\}$ denote the categorical response variable whose categories can be ordered and let $\boldsymbol{x}$ be a vector of (potential) predictors for Y. With $q = k-1$ denoting the amount of relevant category transitions, the sequential logit model in its traditional form is given by (Tutz, 2012)

$$P(Y = r|Y \geq r, \boldsymbol{x}) = \frac{\exp(\beta_{0r} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})}{1 + \exp(\beta_{0r} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})}, \quad r = 1, \ldots, q, \qquad (1)$$

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

which is equivalent to

$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y>r|\boldsymbol{x})}\right)=\beta_{0r}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}.$$

Thus, "continuation ratio logits" are the quantity that is linearly parameterized in sequential logit models. Simple derivations yield

$$\log\left(\frac{\frac{P(Y=r|\boldsymbol{x}_1)}{P(Y>r|\boldsymbol{x}_1)}}{\frac{P(Y=r|\boldsymbol{x}_2)}{P(Y>r|\boldsymbol{x}_2)}}\right)=(\boldsymbol{x}_1-\boldsymbol{x}_2)^{\mathrm{T}}\boldsymbol{\beta}, \tag{2}$$

$$\log\left(\frac{\frac{P(Y=r|\boldsymbol{x})}{P(Y>r|\boldsymbol{x})}}{\frac{P(Y=s|\boldsymbol{x})}{P(Y>s|\boldsymbol{x})}}\right)=\beta_{0r}-\beta_{0s}\quad\forall r\neq s. \tag{3}$$

Equation (2) demonstrates the advantage in interpretation offered by *global, category-unspecific coefficients*. Equation (3) shows that for model (1), the log-odds-ratios between two response categories are unaffected by the covariates. Hence, continuation ratio odds for all categories are proportional to one another and only differ by constant factors $\exp(\beta_{0r}-\beta_{0s})$. A similar proportionality concept also holds for cumulative models and for different link functions, for example, in the proportional hazards model (cf. Tutz, 2012). Therefore, we subsequently call this type of effect a *proportional effect*.

The major disadvantage of model (1) is that it *a priori* assumes all predictors to have proportional effects. A generalization is given by the sequential logit model with *category-specific* and thus *nonproportional effects*:

$$P(Y=r|Y\geq r,\boldsymbol{x})=\frac{\exp(\beta_{0r}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_r)}{1+\exp(\beta_{0r}+\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_r)},\quad r=1,\ldots,q. \tag{4}$$

This model offers much more flexibility and avoids the proportionality assumption, but its interpretability suffers since (2) and (3) no longer hold. Thus, one should only use nonproportional effects $\boldsymbol{\beta}_{\bullet j}=(\beta_{1j},\ldots,\beta_{qj})^{\mathrm{T}}$ for those predictors x_j that actually require it and assign proportional effects otherwise. Additionally, the large number of parameters in model (4) can lead to instability or even non-existence of parameter estimates. For these reasons, it is highly desirable to be able to perform variable selection and to simultaneously decide between proportional and nonproportional effects. In the next section, we develop a penalty approach that performs these selection tasks in an automatic and data-driven way.

2 A Grouped Fusion Penalty for the Selection of Proportional Effects

To perform variable and effect type selection, penalty approaches maximize a penalized loglikelihood of the form $l_{\mathrm{pen}}(\boldsymbol{\beta})=l(\boldsymbol{\beta})-J(\boldsymbol{\beta})$, where

$l(\boldsymbol{\beta})$ denotes the loglikelihood of the chosen model, here based on either (1) or (4), and $J(\boldsymbol{\beta})$ is a functional that penalizes the coefficient vector. Selection of variables in ordinal models has been considered in Archer & Williams (2012) and Zahid & Tutz (2013). The first authors use a proportional sequential logit model like (1) in conjunction with a lasso-type penalty (Tibshirani, 1996) of the form $J(\boldsymbol{\beta}) = \lambda \sum_{r=1}^{q} \sum_{j=1}^{p} |\beta_{rj}|$ to model the influence of genes on disease progression. Zahid & Tutz (2013) use a boosting approach for the proportional odds model.

First, we consider variable selection. In the nonproportional model, each predictor influences the response via a vector $\boldsymbol{\beta}_{\bullet j}$. Following Tutz, Pößnecker & Uhlmann (2012), these coefficient vectors should be penalized jointly to achieve variable selection. Therefore, we suggest to use a penalty similar to the sparse group lasso of Simon et al. (2013):

$$J(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} \left(\psi \sqrt{\boldsymbol{\beta}_{\bullet j}^{\mathrm{T}} \boldsymbol{\beta}_{\bullet j}} \; + \; (1-\psi) \sum_{r=1}^{q} |\beta_{rj}| \right). \tag{5}$$

The tuning parameter $\lambda > 0$ controls the degree of penalization, $\psi \in [0, 1]$ balances variable selection and selection of atomic coefficients.

To select proportional versus nonproportional effects, we add a grouped fusion penalty to (5), obtaining:

$$J(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} \left(\psi \sqrt{\boldsymbol{\beta}_{\bullet j}^{\mathrm{T}} \boldsymbol{\beta}_{\bullet j}} \; + \; (1-\psi) \sum_{r=1}^{q} |\beta_{rj}| \right) + \lambda_2 \sum_{j=1}^{p} \sqrt{\boldsymbol{\beta}_{\bullet j}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{\beta}_{\bullet j}}, \tag{6}$$

with $\boldsymbol{\Omega} = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{D}$ and

$$\boldsymbol{D} = \begin{pmatrix} -1 & 1 & & & \mathbf{0} \\ & -1 & 1 & & \\ & & \ddots & & \\ \mathbf{0} & & & -1 & 1 \end{pmatrix}.$$

Since sequential models implicitly assume that response categories can only be reached successively, only differences between adjacent coefficients must be penalized. Our penalty (6) is able to yield solutions in which $\beta_{1j} = \beta_{2j} = \ldots = \beta_{qj}$ holds and thus shrinks nonproportional effects to proportional ones. Hence, we can always start with the more flexible model (4) since our penalty automatically removes unimportant variables from the model and selects proportional effects whenever possible. Even if a covariate effect is estimated to be nonproportional, the "degree of nonproportionality" is still reduced by (6). Thus, our penalty term provides a continuous transition from nonproportional to proportional effects.

3 Application to the Munich founder study

To illustrate the benfit of our approach, we consider data about the survival of newly founded companies in Munich, Germany. Their survival time, defined as the time to bankruptcy, is measured in intervals of six months, so,

e.g., $y_i = 3$ means bankcruptcy occured between 12 and 18 months after company foundation. Companies that survived more than three years are pooled in a seventh response category. Explanatory variables are economic sector, legal form, location, whether the new company was built up from scratch or resulting from a takeover, starting, equity and debt capital, target market, type of customer, number of employees as well as the degree, sex, age (metric) and business experience of the company's founder.
We fit a sequential logit model with nonproportional effects and penalty (6) to this dataset. Note that the quantity $P(Y = r|Y \geq r, \boldsymbol{x})$ can here be interpreted as a discrete hazard rate. Hence, nonproportional and proportional effects here correspond to time-varying and (time)-constant effects, respectively.
The tuning parameters λ_1 and λ_2 are chosen via 10-fold crossvalidation, ψ is set to 0.5. The parameter estimates for this model are given in Table 1. Proportional effects are assigned to the variables economic sector, target market, customer type and age, hence their effect is constant over time. Legal form, start and debt capital as well as the number of employees have a nonproportional and thus time-varying effect on the companies' survival. The remaining six variables are removed from the model. Therefore, this dataset exhibits a mixture between irrelevant, proportional and nonproportional effects.

4 Outlook

Simulation results prove that our approach distinctly outperforms all existing techniques if the true model contains a mix of proportional and nonproportional effects. We will investigate the consequences and importance of effect type selection in greater detail using both simulated and real data. The connection between sequential models and survival analysis for discrete-time will also be treated in greater detail.

References

Archer, K., Williams, A. (2012). L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine*, **31**, 1464 – 1474.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, **22**, 231 – 245.

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

Tutz, G., Pößnecker, W., Uhlmann, L. (2012). Variable Selection in General Multinomial Logit Models. *Technical Reports, Department of Statistics, LMU*, **126**.

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

Zahid, F., Tutz, G. (2013). Proportional Odds Models with High - dimensional Data Structure. *International Statistical Review*, **81**, 388 – 406.

TABLE 1. Penalized coefficient estimates for the Munich founder study.

Response category	**1**	**2**	**3**	**4**	**5**	**6**
Intercept	-1.76	-1.84	-2.22	-2.19	-2.50	-2.71
sector2	0.52	0.52	0.52	0.52	0.52	0.52
sector3	0.49	0.49	0.49	0.49	0.49	0.49
legalform2	-0.33	-0.49	0.19	-0.47	0.46	-1.26
legalform3	-1.81	-0.99	-1.41	-1.69	-0.76	-1.32
legalform4	-0.38	-0.36	0.11	-0.15	1.14	-0.31
location2	0	0	0	0	0	0
takeover	0	0	0	0	0	0
startcapital2	-1.62	-0.58	0.25	-0.43	-0.62	0.41
startcapital3	-2.92	-0.92	-0.48	-0.16	-1.24	-0.15
equity2	0	0	0	0	0	0
equity3	0	0	0	0	0	0
equity4	0	0	0	0	0	0
debtcapital2	0.79	0.41	0	0.18	0	-0.59
market2	-0.19	-0.19	-0.19	-0.19	-0.19	-0.19
customertype2	-0.24	-0.24	-0.24	-0.24	-0.24	-0.24
degree2	0	0	0	0	0	0
degree3	0	0	0	0	0	0
sex	0	0	0	0	0	0
experience2	0	0	0	0	0	0
experience3	0	0	0	0	0	0
employees2	-0.31	0.23	-0.27	-0.01	-0.53	-0.66
employees3	-0.74	0.07	-0.54	-0.54	0.26	-1.94
age	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01

Beyond the Gauss' principle

Pedro Puig[1]

[1] Universitat Autònoma de Barcelona, Spain

E-mail for correspondence: ppuig@mat.uab.cat

Abstract: The Gauss' principle is reformulated using the concept of asymptotic efficiency. It allows to obtain new characterizations of symmetric error distributions.

Keywords: Asymptotically efficient estimator; Characterization of distributions; Symmetric location models; Trimmed mean.

1 Introduction

In 1809 Gauss proved that the only location model (under mild conditions) such that the sample mean is the maximum likelihood estimator (MLE) of the location parameter is the normal distribution. For this reason this property is known as Gauss' principle (see Puig 2008 and references therein). Of course at Gauss' time maximum likelihood estimation was not invented yet, and he expressed in other words by saying:
It has been customary certainly to regard as an axiom the hypothesis that if any quantity has been determined by several direct observations, made under the same circumstances and with equal care, the arithmetical mean of the observed values affords the most probable value, if not rigorously, yet very nearly al least, so that it is always most safe to adhere to it.
(Book 2, section 177 in Theoria Motus, 1809)
This important result has been the starting point of many maximum likelihood characterizations of distributions, some of them described in the recent paper of Duerinckx et al. (2014). In particular, it is remarkable the work of other pioneer, Poincaré, who characterized in 1896 the one parameter families satisfying the Gauss' principle obtaining an exponential family with a density function of the form,

$$f(x;\theta) = h(x)\exp(A(\theta)x + B(\theta)),$$

where $\theta A'(\theta) + B'(\theta) = 0$.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

According to the Gauss' spirit and all the subsequent research, the key points in the line of reasoning of MLE characterizations can be summarized as follows:

1. The center of interest is certain statistic: sample mean, median, circular mean, trimmed mean, etc.
2. The objective is to find a specific distribution, inside a general family, such that the chosen statistic is the "best" estimator of its population counterpart.
3. MLE is the "best" in the sense that it is asymptotically unbiased and efficient. So, it is usually imposed that the chosen statistic has to be the MLE of the corresponding population parameter.

However, if the statistic of interest is an unbiased estimator of its population counterpart, it seems reasonable to impose only the condition of being asymptotically efficient, not necessarily to be the MLE of the population parameter. It leads to interesting characterizations and even more new distributions, as we'll see in the next section.

2 Characterization of symmetric error distributions

In 1809 Gauss considered the most simple measurement model, where it is assumed that the observations y_i satisfy the relation $y_i = \mu + \varepsilon_i$, $i = 1, ..., n$, where ε_i are independent and identically distributed random errors, and μ represents the true value of the magnitude that have to be estimated. Gauss also assumed that this random error was symmetric around zero and, consequently, the observations could be described by symmetric location models. Therefore, we shall now consider statistical models defined on the real line with a density function of the form $f(x - \mu)$, where μ is the location parameter and $f(t) = f(-t)$. Under enough regularity conditions, the Cramér-Rao lower bound for any unbiased estimator of μ is,

$$\sigma^2_{CR}(f) = \frac{1}{2n \int_0^\infty (f'(x))^2 / f(x) dx}. \tag{1}$$

Given an unbiased estimator $\hat{\mu}$ of the location parameter, having an asymptotic variance $\sigma^2_{\hat{\mu}}(f)$, solving the functional equation $\sigma^2_{CR}(f) = \sigma^2_{\hat{\mu}}(f)$ we shall find all the distributions for which the chosen estimator of the location parameter $\hat{\mu}$ is asymptotically efficient. Sometimes the solutions of this functional equation can be obtained with a tricky application of the Cauchy-Schwarz inequality. Next, we are going to present some results.

2.1 The Hodges-Lehmann estimator

The Hodges-Lehmann estimator (HLE) is the median of all pairwise means of the observations, that is,

$$\hat{\mu} = median \left\{ \frac{y_i + y_j}{2},\ i, j = 1, ..., n \right\}.$$

This is a robust and unbiased estimator of the location parameter for any symmetric location model, and it is very easy to calculate. Its asymptotic variance is,

$$\sigma^2_{HL}(f) = \frac{1}{12n(\int_{\mathbb{R}} f^2(x)\,dx)^2}. \tag{2}$$

Equating (1) and (2) and solving the resulting functional equation, we prove that the only symmetric distribution with a strictly positive density, for which the HLE is an asymptotically efficient estimator of the location parameter μ is the logistic distribution (Damilano and Puig, 2006). It is important to remark that, for the logistic distribution, the HLE is not the MLE of μ. Both are asymptotically equivalent but they are not the same.

2.2 Linear combination of the median and the sample mean

About 200 years ago, Laplace proposed to estimate the population mean μ of a symmetric distribution using a linear combination of the sample mean $\bar{y}$ and the sample median $\tilde{y}$ of the form $\hat{\mu} = w\bar{y}+(1-w)\tilde{y}$, being w a constant (Stigler, 1973). It is immediate to see that this is an unbiased estimator of μ. Laplace proposed to choose w in such a way that its asymptotic variance were minimized. Assuming that the distribution has a continuous density positive at zero, the asymptotic variance of the Laplace estimator (LE) has the form,

$$\sigma^2_{LE}(f) = w^2\frac{v^2}{n} + \frac{w(1-w)\tau}{f(0)n} + \frac{(1-w)^2}{4f^2(0)n}, \tag{3}$$

where $v^2 = \int_{-\infty}^{\infty} t^2 f(t)dt$ and $\tau = \int_{-\infty}^{\infty} |t| f(t)dt$. Equating (1) and (3) and solving the resulting functional equation, we obtain a three-parameter density of the form,

$$f_\theta(x;\mu,\sigma) = \frac{\phi(\theta)}{2(1-\Phi(\theta))\sigma}\exp(-\theta\frac{|x-\mu|}{\sigma} - \frac{(x-\mu)^2}{2\sigma^2}), \tag{4}$$

where $\phi(.)$ and $\Phi(.)$ denote the standard normal density and cumulative distribution functions, respectively, μ and σ are location and scale parameters, and θ is a shape parameter (Damilano and Puig, 2004). Moreover, fixing θ, the constant w of the linear combination remains $w = w(\theta) = (1-\Phi(\theta))/(1-\Phi(\theta)+\theta\phi(\theta))$. The density is unimodal for $\theta \geq 0$ and bimodal for $\theta < 0$ as can be seen in Figure 1. Note that when $\theta = 0$ it is the normal distribution. When θ and σ tend to ∞ such a way that θ/σ tends to a constant, the limiting density is that of the Laplace distribution. This density can be successfully used to model currency exchange interbank rates US Dollar/Euro (Damilano and Puig, 2004).

2.3 The trimmed mean

Trimmed mean was first documented in an anonymous work in 1821:
...to determine the mean yield of a property of land, there is a custom to observe this yield during twenty consecutive years, to remove the strongest

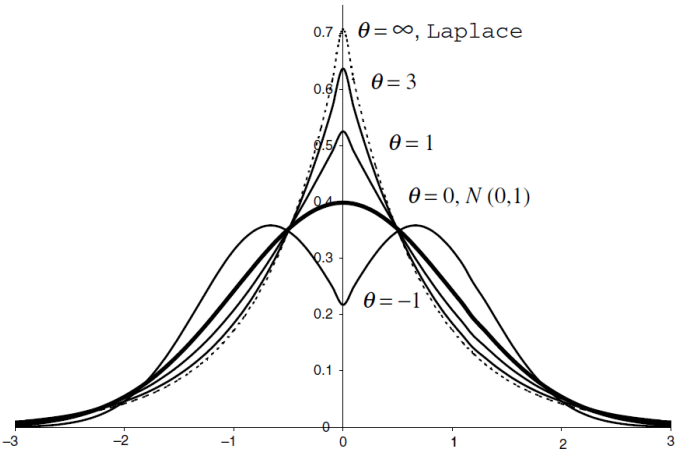


FIGURE 1. Standardized densities for some values of θ.

and the weakest yield and then to take one eighteenth the sum of the others. (Annales de Mathmatiques pures et appliques, tome 12 (1821-1822), p. 181-204. Translated by Huber in 1972.)

The α%-trimmed mean (TM_α) of the observations is calculated by sorting all the values, discarding α% of the smallest and α% of the largest values, and computing the average of the remaining values. Note that TM_0 is the sample mean and $\mathrm{TM}_{0.5}$ is the sample median. It is known that for symmetric distributions this a unbiased estimator of the population mean μ, with an asymptotic variance expressed as,

$$\sigma^2_{TM_\alpha}(f) = \frac{1}{n(1-2\alpha)^2}\left(\int_{-a}^{a} x^2 f(x)dx + 2\alpha a^2\right), \tag{5}$$

where, $-a = F^{-1}(\alpha)$ and $a = F^{-1}(1-\alpha)$.

Similarly to the preceding examples, equating the asymptotic variance (5) to (1) and solving the corresponding functional equation, the following three parameter density is obtained:

$$h(x;\mu,a,b) = \frac{b^2 e^{b^2}}{a(erf(b)\sqrt{\pi}b + e^{-b^2})} \begin{cases} \exp\left(-\frac{((x-\mu)^2+a^2)b^2}{a^2}\right) & |x-\mu| \le a \\ \exp\left(-\frac{2b^2|x-\mu|}{a}\right) & |x-\mu| > a \end{cases}$$

where $erf(.)$ is the error function. Note that the profile of this sliced density is like a Gaussian density in the interval $|x-\mu| \le a$, and like a Laplace density otherwise. The location parameter (population mean) is indicated as μ, and b is a shape parameter that directly determines the truncation percentage α% of the trimmed mean and the kurtosis of the distribution as well. Direct calculations show that b can be expressed as a function of α solving numerically the equation,

$$erf(b)\sqrt{\pi}b\exp(b^2) = \frac{1}{2\alpha} - 1.$$

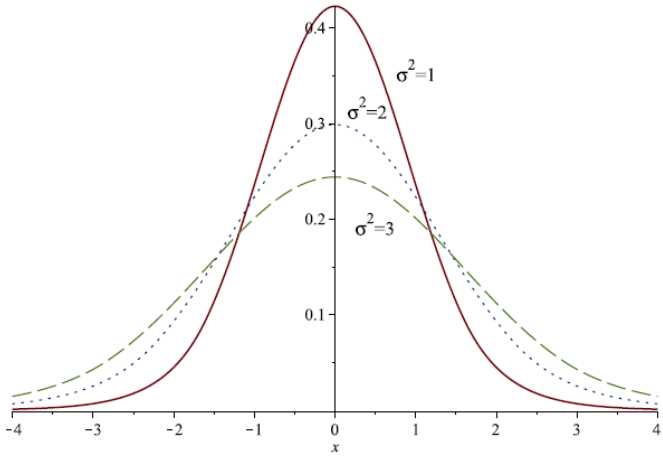


FIGURE 2. Standardized densities ($\mu = 0$) for $\alpha = 5\%$ and some different values of the variance σ^2.

For instance, when $\alpha = 0.05$ (this is just the case considered in the former publication of 1821) we obtain $b = 1.2273$. It can be shown that, when b is fixed, the variance of the distribution σ^2 is a lineal function of a^2. In particular, for $\alpha = 0.05$ the variance has the expression $\sigma^2 = 0.38718a^2$. Figure 2 illustrates the profiles of these densities for $\mu = 0$ and variances equal to 1, 2 and 3.

When $\alpha = 0$, b tends to ∞ and $h(x; \mu, a, b)$ tends to be a Gaussian density. On the other hand, when $\alpha = 0.5$, b tends to 0 and $h(x; \mu, a, b)$ tends to be a Laplace density. Figure 3 shows the corresponding standardized densities for several values of α.

As far as we know, this density has not been considered or studied before. Anyway, we are not optimistic about its utility in practical applications. However, this density can be useful in studies of robustness of location estimators, where the trimmed mean will be the "best" estimator of the location parameter for samples generated from this distribution.

Acknowledgments: This work was partially funded by the grant MTM2012 31118 from the Ministry of Economy and Competitiveness of Spain.

References

Damilano, G. and Puig, P. (2004). Efficiency of a Linear Combination of the Median and the Sample Mean: The Double Truncated Normal Distribution *Scandinavian Journal of Statistics*, **31**, 629 – 637.

Damilano, G. and Puig, P. (2006). A Note on the HodgesLehmann Estimator and the Logistic Distribution. *Communications in Statistics-Theory and Methods*, **35**, 257 – 261.

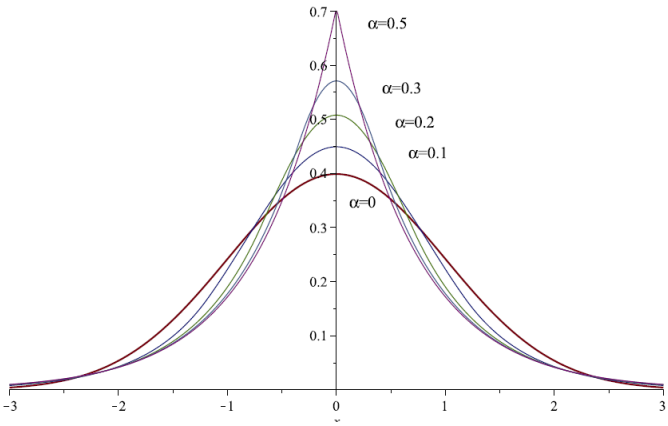


FIGURE 3. Standardized densities ($\mu = 0$, $\sigma^2 = 1$) for several values of the truncation proportion $\alpha = 0, 0.1, 0.2, 0.3, 0.5$.

Duerinckx, M., Ley, C. and Swan, Y. (2014) Maximum likelihood characterization of distributions. *Bernoulli* (forthcoming paper).

Puig, P. (2008). A note on the harmonic law: a two-parameter family of distributions for ratios. it Statistics and Probability Letters, **78**(3), 320 – 326.

Stigler, S.M. (1973). Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika*, **60**, 439 – 445.

Local Influence Diagnostics for Generalized Linear Mixed Models With Overdispersion

Trias Wahyuni Rakhmawati[1], Geert Molenberghs[1,2], Geert Verbeke[2,1], Christel Faes[1,2]

[1] I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium
[2] I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium

E-mail for correspondence: triaswahyuni.rakhmawati@uhasselt.be

Abstract: Since the seminal paper by Cook and Weisberg (1982), local influence, next to case deletion, has gained popularity as a tool to detect influential subjects and measurements for a variety of statistical models. For the linear mixed model the approach leads to easily interpretable and computationally convenient expressions, not only highlighting influential subjects, but also which aspect of their profile leads to undue influence on the model's fit (Verbeke and Lesaffre 1998). Ouwens, Tan, and Berger (2001) applied the method to the Poisson-normal generalized linear mixed model (GLMM). Given the model's non-linear structure, these authors did not derive interpretable components but rather focused on a graphical depiction of influence. In this paper, we consider GLMMs for binary, count, and time-to-event data, with the additional feature of accommodating overdispersion whenever necessary. For each situation, three approaches are considered, based on: (1) purely numerical derivations; (2) using a closed-form expression of the marginal likelihood function; and (3) using an integral representation of this likelihood. The methodology is illustrated in case studies of A Clinical Trial in Epileptic Patients.

Keywords: Boundary condition; Case deletion; GLMM; Combined model; Local Influence.

1 Introduction

Next to linear mixed models (LMM) for hierarchical Gaussian data (Verbeke and Molenberghs 2000), generalized linear mixed models (GLMM) have become a tool for routine use for the analysis of a hierarchical data of a variety of data types over the last twenty years (Molenberghs and Verbeke, 2005). Like with every statistical model, after formulating and fitting a model, an assessment of model fit and a diagnostic analysis is advisable. In this paper, we are concerned with the detection of influential subjects.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

A large variety of diagnostic tools is available for linear and generalized linear models. Cook and Weisberg and and Chatterjee and Hadi (1988) provide early treatises. In classical linear regression, Cook's distances (Cook 1977a, 1977b, 1979) have been used extensively. Linear mixed models, unlike linear models, generally do not allow for closed-form parameter estimators. Further, residual analysis is not straightforward, given the presence of both fixed-effect and random-effects covariates, so that even uniquely defining residuals is not possible. For these and related reasons, Lesaffre and Verbeke (1998) chose local influence (Cook 1986, Backman, Nachtsheim, and Cook 1987) to examine influence in linear mixed models.
In this study, we extend local influence for the GLMM in several ways. First, we consider outcomes of binary, count, and time-to event type. Second, using the extension proposed by Molenberghs, Verbeke, and Demétrio (2007) and Molenberghs *et al* (2010), we flexibly allow for overdispersion in the GLMM, by introducing conjugate random effects, in addition to normal ones. This model is referred to as the combined model. Third, apart from numerical derivations of local influence, we examine two alternative routes: (a) closed forms for the marginal likelihood such as proposed in Molenberghs *et al* (2010) and (b) the marginal likelihood with integral form. The closed forms in (a) do not always exist; while they are available for the probit-(beta-)normal, Poisson-(gamma-)normal, and Weibull-(gamma-)normal, they are not for the logit-(beta-)normal. Even when they do, they may be somewhat unwieldy and therefore, route (b) is more promising.

2 Local Influence for GLMM

Local influence was presented by Cook (1986) and used by several authors since. The impact of individuals and measurements on the analysis is assessed by comparing standard maximum likelihood estimates with those resulting from slightly perturbing the contribution of an individual or a measurement. Lesaffre and Verbeke (1998) introduced an influence assessment paradigm for the linear mixed model.
Cook (1986) derived a convenient computational scheme.Let $\boldsymbol{\Delta}_i$ be the s-dimensional vector of second-order derivatives of log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$, w.r.t. perturbation ω_i and all components of $\boldsymbol{\theta}$, and evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Also, write Δ for the $s \times r$ matrix with $\boldsymbol{\Delta}_i$ in the ith column. Let $\ddot{L}$ denote the $s \times s$ matrix of second derivatives of $\ell(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. For any unit vector $\boldsymbol{h}$ in Ω, it follows that: $C_h = 2 \left| \boldsymbol{h}' \Delta' \ddot{L} \Delta \boldsymbol{h} \right|$. Various choices for $\boldsymbol{h}$ have received attention. First, as will be done here, one can focus on subject i only, by choosing $\boldsymbol{h} = \boldsymbol{h_i}$, the zero vector with a sole 1 in the ith position. Local influence then is $C_i \equiv C_{h_i} = 2 \left| \boldsymbol{\Delta}_i' \ddot{L} \boldsymbol{\Delta}_i \right|$. Lesaffre and Verbeke (1998) showed that local influence C_i can be re-expressed as

$$C_i = 2||\ddot{L}||\ ||\boldsymbol{\Delta_i}||^2 \cos(\varphi_i), \tag{1}$$

where φ_i is the angle between $\text{vec}(-\ddot{L})$ and $\text{vec}(\boldsymbol{\Delta_i \Delta_i'})$.

The integral-based approach can be used as an alternative way to alleviate complexities with the explicit marginal likelihood expressions. The marginal density corresponding to the linear mixed model is defined as: $\widetilde{f}(\boldsymbol{y_i}) = \int \widetilde{f}(\boldsymbol{y_i}|\boldsymbol{\beta}, \boldsymbol{b}_i)\widetilde{f}(\boldsymbol{b}_i|D)\ d\boldsymbol{b}_i$, with the log-likelihood contribution of the ith individual takes the form: $\ell_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \widetilde{f}(\boldsymbol{y_i})$.
For count data,the first derivative of log-likelihood contribution for ith subject as followed:

$$\frac{\partial \ell_i(\boldsymbol{\beta}, D)}{\partial \boldsymbol{\beta}} = \sum_{j=1}^{n_i} \{y_{ij} - E(y_{ij}|\boldsymbol{b}_i)\}\, \boldsymbol{x}_{ij} = \sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij}, \tag{2}$$

$$\frac{\partial \ell_i(\boldsymbol{\beta}, D)}{\partial d_{jk}} = -\frac{1}{2}(2 - \delta_{jk}) \left\{ (D^{-1})_{jk} - (D^{-1}D^{-1})_{jk} \sum_{k=1}^{q} \mathrm{Var}(b_{ik}) \right\} \tag{3}$$

where d_{jk} is a component of D and δ_{jk} is one if j is equal to k, and zero otherwise. Interpretable expressions can now be derived using (1). It showed

$$\begin{aligned} ||\boldsymbol{\Delta}_i||^2 &= \left(\sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij} \right) \left(\sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij} \right)' \\ &+ \sum_{k,l} \left\{ -\frac{1}{2}(D^{-1})_{kl} + \frac{1}{2}(D^{-1}D^{-1})_{kl}\mathrm{Var}(\boldsymbol{b}_i) \right\}^2. \end{aligned}$$

Let $C_i = C_{1i} + C_{2i}$ with:

$$C_{1i} = 2||\ddot{L}||\ ||\boldsymbol{r}_i\boldsymbol{x}_i||^2 \cos(\varphi_i), \tag{4}$$

$$C_{2i} = \frac{1}{2}||\ddot{L}||\ ||(D^{-1})_{kl} - (D^{-1}D^{-1})_{kl}\mathrm{Var}(\boldsymbol{b}_i)||^2 \cos(\varphi_i), \tag{5}$$

where $\boldsymbol{r}_i\boldsymbol{x}_i = \sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij}$. Note that C_{1i} and C_{2i} are the contributions of subject i to local influence C_i from $\boldsymbol{\beta}$ and D, respectively. Reconstructing the component C_{1i} and C_{2i} leads to the interpretable components that can be described local influence. Hence, the interpretable components of C_i in the case of the Poisson-normal model can be described using the 'length of the fixed effect' ($||\boldsymbol{x}_i\boldsymbol{x}_i'||$), the 'squared length of the residual' ($||\boldsymbol{r_i}||^2$), and the 'squared of random effect variability' ($\mathrm{Var}(\boldsymbol{b}_i)^2$).
In binary cases, the local influence for both probit and logit normal models have been derived.The first derivatives for probit-normal model are:

$$\frac{\partial \ell_i(\boldsymbol{\xi}, D)}{\partial \boldsymbol{\xi}} = [I - (\boldsymbol{X}_i\boldsymbol{\xi})^{-1}]\boldsymbol{X}_i, \tag{6}$$

$$\frac{\partial \ell_i(\boldsymbol{\xi}, D)}{\partial d_{jk}} = \frac{3}{2}L^{-1}\left(I_{n_i} - Z_iM_iM_i'(D^{-1}D^{-1})_{jk}Z_i'\right), \tag{7}$$

where $M_i = \left(D^{-1} + Z_i'Z_i\right)^{-1}$. It also follows that

$$||\boldsymbol{\Delta}_i||^2 = [I - (\boldsymbol{X}_i\boldsymbol{\xi})^{-1}]^2\boldsymbol{X}_i\boldsymbol{X}_i' + \sum_{k,l} \frac{9}{4L^2}\left(I_{n_i} - Z_iM_iM_i'(D^{-1}D^{-1})_{jk}Z_i'\right)^2.$$

Thus, also for this case, the components $||\boldsymbol{X_i}||^2$ and $||\boldsymbol{Z_i Z_i'}||^2$ turn up. Evidently, the same binomial expression is used, but now with $\text{logit}(\lambda_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$.The derivatives of logit-normal model take the form:

$$\frac{\partial \ell_i(\boldsymbol{\xi}, D)}{\partial \boldsymbol{\xi}} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \int \frac{1}{1+\exp(\mu_{ij})} \tilde{\tau}(\boldsymbol{b}_i|\boldsymbol{y}_i) d\boldsymbol{b}_i, \quad (8)$$

$$\frac{\partial \ell_i(\boldsymbol{\xi}, D)}{\partial d_{jk}} = -\frac{1}{2}(2-\delta_{jk}) \left\{ (D^{-1})_{jk} - (D^{-1}D^{-1})_{jk} \text{Var}(\boldsymbol{b}_i) \right\}, \quad (9)$$

where $\mu_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$. It also follows that

$$||\boldsymbol{\Delta}_i||^2 \propto \left(\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \right) \left(\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \right)' + \sum_{k,l} \left(-\frac{1}{2}(D^{-1})_{kl} + \frac{1}{2}(D^{-1}D^{-1})_{kl} \text{Var}(\boldsymbol{b}_i) \right)^2.$$

Reconstructing the fixed- and random-effects components, respectively, like in the Poisson case, leads to $C_{1i} = 2||\ddot{L}||\ ||\boldsymbol{x}_i||^2 \cos(\varphi_i)$ and C_{2i} as in (5). Hence, the interpretable components of C_i for the logit-normal model can be described using the length of fixed effect ($||\boldsymbol{x}_i||^2$) and the squared random-effects variability, $\text{Var}(\boldsymbol{b}_i)^2$ (i.e., the sum of all variances), in analogy with the Poisson-normal model. The same is true for the Weibull-normal model, as will be seen next.
The first derivative for Weibull case take the form:

$$\frac{\partial \ell_i(\boldsymbol{\xi}, D)}{\partial \boldsymbol{\xi}} = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} - \lambda \sum_{j=1}^{n_i} y_{ij}^{\rho} \boldsymbol{x}_{ij} \exp(\boldsymbol{\mu}_{ij}), \quad (10)$$

$$\frac{\partial \ell_i(\boldsymbol{\xi}, D)}{\partial d_{jk}} = -\frac{1}{2}(2-\delta_{jk}) \left[(D^{-1})_{jk} - (D^{-1}D^{-1})_{jk} \text{Var}(\boldsymbol{b}_i) \right], \quad (11)$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. It further follows that

$$\begin{aligned} ||\boldsymbol{\Delta}_i||^2 &= \left(\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \right) \left(\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \right)' - 2 \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \boldsymbol{Q}'_i + \boldsymbol{Q}_i \boldsymbol{Q}'_i \\ &+ \sum_{k,l} \left\{ -\frac{1}{2}(D^{-1})_{kl} + \frac{1}{2}(D^{-1}D^{-1})_{kl} \text{Var}(\boldsymbol{b}_i) \right\}^2, \end{aligned}$$

where $\boldsymbol{Q}_i = \lambda \sum_{j=1}^{n_i} y_{ij}^{\rho} \boldsymbol{x}_{ij} \exp(\boldsymbol{\mu}_{ij})$.
Like in the Poisson-normal and binary-normal cases, a decomposition $C_i = C_{1i} + C_{2i}$ follows, with

$$C_{1i} = 2||\ddot{L}||\ \left\{ ||\boldsymbol{x}_i||^2 - 2\boldsymbol{x}_i \boldsymbol{Q}_i + ||\boldsymbol{Q}_i||^2 \right\} \cos(\varphi_i)$$

and C_{2i} as in (5). Hence, interpretable components analogous to the earlier settings arise.

3 Application

The Epileptic dataset consisted of 89 patients with 2 different group of treatments, placebo and a new anti-epileptic drug (AED). Patients were followed (double-blind) during 16 weeks (some patients until 27 weeks). The outcome of interest is the number of epileptic seizures experienced during the most recent week. Poisson-normal (P-N) model as well as the combined model with gamma random effect (PGN) have been fitted as follow:

$$\ln(\boldsymbol{\lambda}_{ij}) = \begin{cases} (\beta_{00} + b_i) + \beta_{01} t_{ij} & \text{if placebo} \\ (\beta_{10} + b_i) + \beta_{11} t_{ij} & \text{if treated,} \end{cases} \tag{12}$$

where Y_{ij} represent the number of epileptic seizures patient i experiences during week j, t_{ij} is the time point at which Y_{ij} has been measured and the random intercept $b_i \sim N(0, d)$. Parameter estimates are given in Table 1. Figure 1 contain index plots (versus patient ID) for various local influence analyses. The top row represents local influence for (P-N) model, yet in below rows for (PGN) model. Patients #38, #49, and #62 stand out with large total influence C_i when compared to other patients. Importantly, influences show a major drop when switching from (P-N) to (PGN). To get further insight as to why these subject have higher influence than others, plots with interpretable components are given in Figure 2: 'squared length of the fixed effects' $||\boldsymbol{x}_i \boldsymbol{x}_i'||$, 'squared length of the residual' $||\boldsymbol{r_i}||^2$, and 'random-effect variability' $\text{Var}(b_i)^2$. It is hardly surprising that #38 stands out in terms of $||\boldsymbol{r_i}||^2$. Influences on #49 and #62 are less pronounced.

4 Discussion

It has been showed from this study that the influential subject for hierarchical model can be detect using local influence approach. And it was found that the combined model can be used to reduced the influence effect of the subject. Moreover, the interpretable components can be use as the tools to evaluate in which way the influence subject affect the estimation in modeling process.

References

Cook, R.D (1986). Assessment of Local Influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133 – 169.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**, 570 – 582.

Molenberghs G. *et al* (2010). A Family of Generalized Linear Models for Repeated Measures with Normal and Conjugate Random Effects. *Statistical Science*, **25(3)**, 325 – 347.

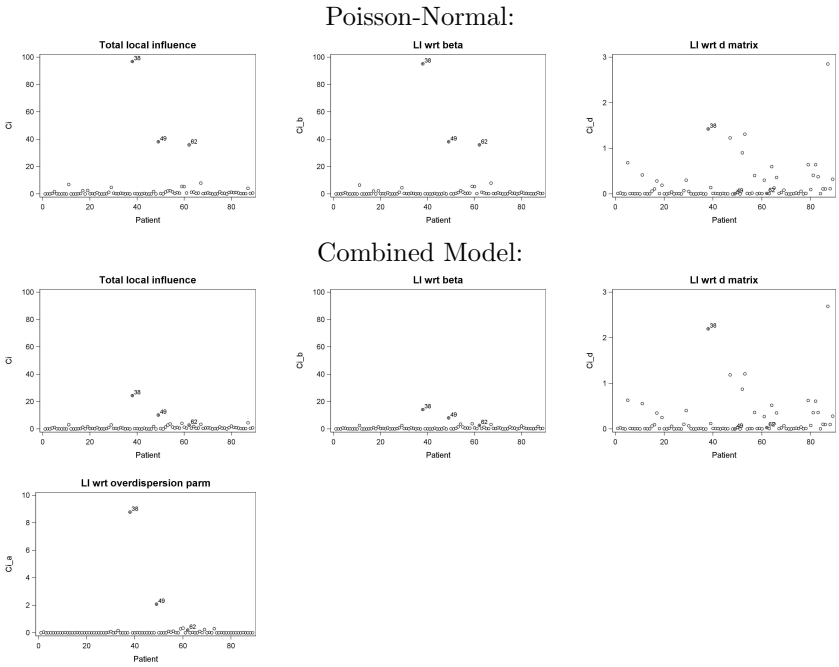


FIGURE 1. Plot of Local Influence

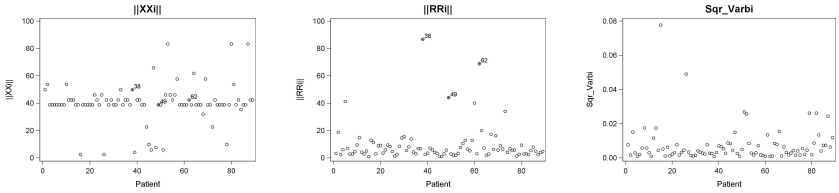


FIGURE 2. Plot of Interpretable components of Local Influence

TABLE 1. Parameter estimates (standard errors) for the P-N and PGN models.

Effect	Par.	P-N	PGN
Interc. plac.	β_{00}	0.818(0.168)	0.911(0.176)
Slope plac.	β_{01}	-0.014(0.004)	-0.025(0.008)
Interc. treat.	β_{10}	0.648(0.170)	0.656(0.178)
Slope treat.	β_{11}	-0.012(0.004)	-0.012(0.007)
Treat. eff.	$\beta_{11} - \beta_{10}$	0.002(0.006)	0.013(0.011)
Treat. eff.	β_{11}/β_{10}	0.840(0.398)	0.475(0.335)
Std. rand. int.	σ	1.076(0.086)	1.063(0.087)
Overdisp. par.	α		2.464(0.211)

Improved Methods for Nearest Neighbor Imputation

Shahla Ramzan[1], Gerhard Tutz[1], Christian Heumann [1]

[1] Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: shahla.ramzan@stat.uni-muenchen.de

Abstract: Missing data is an important issue in almost all fields of quantitative research. A nonparametric procedure that has been shown to be useful is the nearest neighbor imputation method. We suggest a weighted nearest neighbor imputation method based on L_q-distances. The weighted method is shown to have smaller imputation error than available NN estimates. In addition we consider weighted neighbor imputation methods that use selected distances. The careful selection of distances that carry information on the missing values yields an imputation tool that outperforms competing nearest neighbor methods distinctly.

Keywords: kernel function; weighted nearest neighbors; weighted imputation; MCAR; metric data

1 Nearest Neighbors

Missing data has always been a challenging topic for researchers. Ignoring all the missing cases is not an advisable way to deal with missing values. The methods for filling the incomplete data matrix can be divided into two main categories; single imputation and multiple imputation (Little and Rubin (1987)). In the literature many techniques have been suggested to impute data when the values are missing completely at random (MCAR), nearest neighbor imputation is one of them (Troyanskaya et al. (2001)). It is based on the average of the k *nearest neighbors* which are computed based on some distance measure. In a comparative study on gene expression data, Troyanskaya et al. (2001) concluded that k nearest neighbor imputation performs better than mean imputation and singular value decomposition techniques.
We suggest a new imputation estimate based on weighted average of k nearest neighbors using L_q distance. Our proposed method does not depend

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

on k, rather it uses almost all the information in the neighbors. For the high-dimensional cases, we suggest a new distance that uses weights that are proportional to the correlation among variables. This method automatically selects the relevant variables.

2 Weighted Neighbors

Let data be collected in (n, p) matrices, $\boldsymbol{X}=\{\boldsymbol{x_i}\} = (x_{is})$, $i = 1, \ldots, n$, with x_{is} denoting i^{th} observation of the s^{th} variable, and let $\boldsymbol{O} = (o_{is})$, with $o_{is} = 1$ denoting that it has been observed, $o_{is} = 0$ if it is missing. If x_{is} is a missing value ($o_{is} = 0$), one determines the k nearest neighbors from the $(\tilde{n}, p)$ reduced data set $\tilde{\boldsymbol{X}} = \{\boldsymbol{x}_j, o_{js} = 1\}$ obtaining

$$\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(k)} \quad \text{with} \quad d(\boldsymbol{x}_i, \boldsymbol{x}_{(1)}) \leq \cdots \leq d(\boldsymbol{x}_i, \boldsymbol{x}_{(k)})$$

Distances between two observation $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, which represent rows in the data matrix, can be computed by using the L_q-metric for the observed data

$$d_q(\boldsymbol{x}_i, \boldsymbol{x}_j) = [\frac{1}{m_{ij}} \sum_{s=1}^{p} |x_{is} - x_{js}|^q I(o_{is} = 1) I(o_{js} = 1)]^{1/q}, \tag{1}$$

where $m_{ij} = \sum_{s=1}^{p} I(o_{is} = 1) I(o_{js} = 1)$ denotes the number of valid components in the computation of distances. The indicator function $I(a)$ = 1, if a is true and 0 otherwise. The actually used components in the computation of neighbors is given by $C_{ij} = \{s : I(o_{is} = 1).I(o_{js} = 1) = 1\}$. The *weighted imputation estimate* for x_{is} considered here is

$$\hat{x}_{is} = \sum_{j=1}^{k} w(\boldsymbol{x}_i, \boldsymbol{x}_{(j)}) x_{(j)s} \tag{2}$$

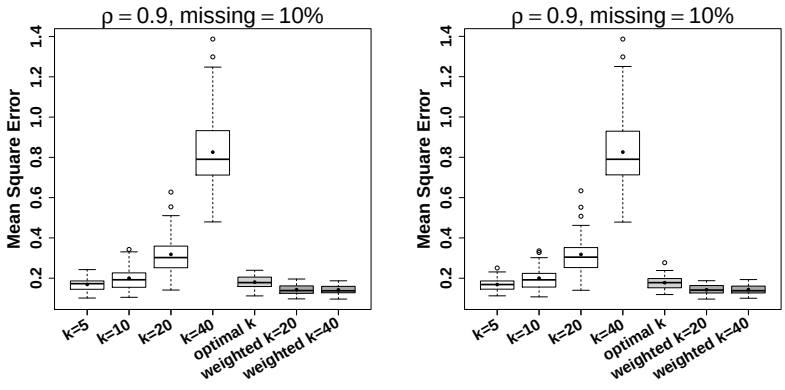
with weights

$$w(\boldsymbol{x}_i, \boldsymbol{x}_j) = K(d(\boldsymbol{x}_i, \boldsymbol{x}_j)/\lambda) / \sum_{l=1}^{k} K(d(\boldsymbol{x}_i, \boldsymbol{x}_j)/\lambda), \tag{3}$$

where $K(.)$ is a kernel function and λ is a tuning parameter. If $k = \tilde{n}$, where $\tilde{n}$ is the number of all *available* nearest neighbors of x_{is}, then the widow-width λ is the only tuning parameter. The optimal value of λ is chosen by cross validation. The performance of different NN imputation methods is compared on the basis of mean square error (MSE) computed from original and imputed values (Troyanskaya et al. (2001)).

2.1 Simulation Study

For a pre-defined number of settings, the data are generated from $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the correlation matrix. The data are missed completely at random with a probability of p. The weighted NN imputation estimates for



(a) L_1-metric (2) L_2-metric

FIGURE 1. Comparison of NN imputation with fixed k (white boxes), optimal k (light grey), and weighted NN estimates (dark grey).

L_1-metric (q=1) and L_2-metric (q=2) are compared with unweighted approach (see, as an example, Figure 1 for ρ=0.9 with 10% missing data). Simulation results (not presented here) suggest that the Gaussian kernel provides smaller MSE as compared to other kernel functions. The optimal λ is chosen with MSE via cross validation. It is seen that for larger values of k, weighting results in a reduced MSE.

3 Weighted Neighbors with Selection of Predictors

An extended version of (1) uses weights for the computation of distances that are linked to the correlation. Consider imputation for $\boldsymbol{x}_i$ in component s ($o_{is} = 0$). When computing distances from the reduced data set $\{\boldsymbol{x}_j, o_{js} = 1\}$ we propose to use use an additional weight. More concrete, for L_q-distance we compute component-specific distances by

$$d_{q,C}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \{\frac{1}{m_{ij}} \sum_{l=1}^{p} |x_{il} - x_{jl}|^q I(o_{is} = 1) I(o_{js} = 1) C(r_{sl})\}^{1/q}, \quad (4)$$

where r_{sl} is the empirical correlation between covariates s, l and $C(.)$ is a convex function defined on the interval $[-1, 1]$ that selects the components that are correlated with component s (or weights according to the strength of the correlation). Thus, components that are strongly correlated with component s contribute to the computation of distances, components that are not correlated do not contribute to the distance.
For $C(.)$, we used $C(r) = 0$ if $|r| \leq c$, $C(r) = |r|/(1-c) - c/(1-c)$ if $|r| > c$, where $c > 0$. Thus if $|r_{sj}| \leq c$ the component s does not contribute to the distance. Of course c is an additional tuning parameter. A smoother approach is $C(r) = |r^m|$, where m is the additional tuning parameter. Here r_{sl} cannot be computed since missing values are present in the data. For simplicity we use a simple first step imputation, namely unweighted

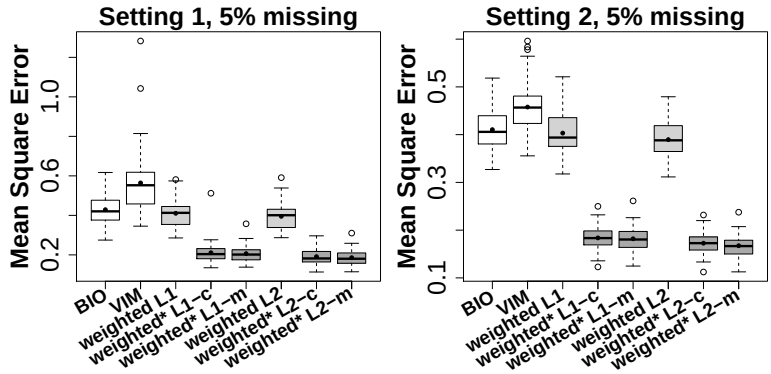


FIGURE 2. Comparison of weighted NN imputation with selection of variables(*) (dark grey), weighted L_1 & L_2 metrics (light grey), and two `R` packages as benchmarks (white); `impute` from Bioconductor (BIO) & `VIM` (VIM).

five nearest neighbors from which the correlations are computed. Then the actual imputation is carried out by using the computed correlation. The optimal values of the tuning parameters are again selected by cross validation on a double grid, i.e., the pair (λ, c) or (λ, m) with minimal MSE is selected as optimal.

3.1 Simulation Study

We generate data with predefined settings from $N_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the correlation matrix of a prespecified structure with pairwise correlation ρ. The data is replaced with missing values (missing completely at random) with a specified probability p. The average MSE is obtained from the imputation estimates, using distances (1) and (4) at their corresponding optimal tuning parameters chosen via cross validation. As an example, the weighted NN imputation results with an *autoregressive* type correlation structure $\mathbf{\Sigma}$ with ρ=0.9, 5% missing for $n = 50$ (setting 1) and $n = 100$ (setting 2) are shown in Figure 2. We use as benchmarks two methods: (i) the function `impute.knn` from Bioconductor `R package version 1.36.0` (Hastie et al.), which uses k-nearest neighbors to impute the missing expression values (ii) the function `kNN` from the `R` package `VIM` (Templ et al. (2013)). Figure 2 shows that the weighted NN imputation procedure provides the smaller mean square error. Moreover, the selection of predictors has improved the weighted NN imputation. This selection is more effective when the sample size is large, e.g., in setting 2 with $n = 100$, the smallest average MSE is 0.05, whereas with n=50 the smallest average MSE is obtained at 0.2 as shown in Figure 2.

Next we consider the dependence of MSE on the probability of the occurrence of missing data. Therefore, we perform another simulation study for different predefined simulation settings. To compare the performance of imputation methods, the mean square error is computed with tuning parameters chosen by cross validation. As example, the average MSE obtained

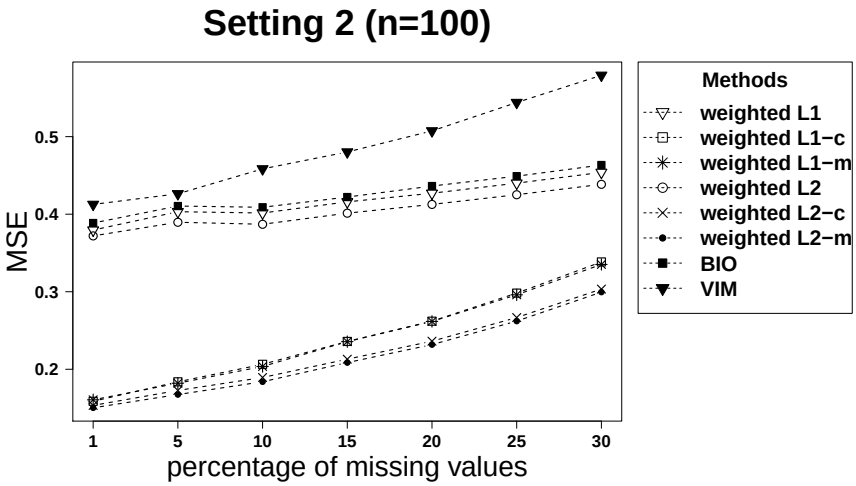


FIGURE 3. Comparison of average MSE of different NN imputation methods

for $n = 100$ at different percentages of missing data, is shown in Figure 3. It is obvious that the suggested imputation procedure with weighted distances performs remarkably well even when the number of missing data is high. When the probability of missing data is small, the weighted selection of predictors using L_1 and L_2 metrics provide nearly similar MSE. This difference increases as the percent of missing data increases. But ultimately L_2 metric provides the smallest average MSE.

4 Application

As application we use gene expression data on three different types of human tumor namely, diffuse large B-cell lymphomas (DLBCL), leukemia and brain tumor. We consider 100 genes f associated with these three types of tumor. The data can be accessed at `http://www.gems-system.org`. The NN imputation techniques were applied to these data after replacing 5% values by missing (MCAR). The missing values were imputed using weighted kNN with selection of variables. The values of tuning parameters were selected via cross validation. For the benchmark methods, the number of the nearest neighbors k was also selected by cross validation. The cross validation process was repeated 10 times. Table 1 shows the average of MSE results.

TABLE 1. Application on gene expression data

	L_1 **metric**		L_2 **metric**		**Benchmark**	
DATA	**opt-c**	**opt-m**	**opt-c**	**opt-m**	**BIO**	**VIM**
DLBCL	0.13740	**0.12802**	0.15443	0.14241	0.15445	0.14278
Brain Tumor	0.18433	0.17274	0.18195	**0.16648**	0.18996	0.16845
Leukemia	0.07515	**0.06602**	0.08136	0.06988	0.08946	0.07845

In all three case studies, the minimum value of MSE is obtained by one of the suggested imputation methods. For DLBCL data, the minimum, 0.12802, was obtained for the L_1 metric with the second convex function showing better performance. The `VIM` procedure had an MSE of 0.14278, and the `impute` function in Bioconductor package produced the highest value of MSE, 0.15445. For the brain tumor data, the smallest MSE, 0.16648, was found for the weighted procedure using the L_2-metric and convex function with optimal m. Similar results are found for the Leukemia. Thus, for all three case studies the weighted imputation procedure including selection of variables showed the best performance.

5 Concluding Remarks

The main objective of this study was to find an improved nearest neighbor procedure for imputation of missing values. When the data are missing completely at random (MCAR), the simulation results, in general, show that the suggested *weighted imputation estimate* performs better than unweighted approach. The comparison of L_1 and L_2 metrics in the weighted imputation of missing data shows L_2 metric to be slightly better than L_1 metric. For high dimensional data, a weighted selection of predictors for imputation is suggested that uses cross validation for selection of optimal tuning parameters. The simulation results show that when the data are highly correlated, the proposed NN imputation procedure shows promising results.

References

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. *impute: Imputation for microarray data*. R package version 1.36.0.

Little, R. J. and Rubin, D. B. (1987). *Statistical analysis with missing data*. Volume 539. New York: Wiley.

Templ, M., Alfons, A., Kowarik, A., and Prantner, B. (2013). *VIM: Visualization and Imputation of Missing Values*. R package version 4.0.0.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17 (6)**, 520 – 525.

Bayesian inference in random-effects meta-analysis

Christian Röver[1], Tim Friede[1]

[1] Department of Medical Statistics, University Medical Center, Göttingen, Germany

E-mail for correspondence: christian.roever@med.uni-goettingen.de

Abstract: Meta-analyses are commonly conducted using a simple random-effects model. We show that within a Bayesian framework part of the necessary integration can be done analytically, allowing to implement numerical strategies to accurately perform all necessary integrations with little computational effort. We illustrate the approach via an example.

Keywords: Meta-analysis; Bayesian analysis; numerical integration.

1 Random-effects meta-analysis

Meta analyses are commonly performed using the *random effects model*. Parameter estimates from several studies are combined, taking into account their corresponding standard errors, to yield an overall, joint estimate of the parameter of interest. Potential discrepancies between study results are anticipated and accounted for using an additional variance parameter, the *heterogeneity* τ. Assuming we have k parameter estimates y_i with associated standard errors σ_i, the model may be stated as

$$y_i \;\sim\; \mathrm{Normal}\big(\theta, \sigma_i^2 + \tau^2\big) \tag{1}$$

i.e., each estimate y_i deviates from the true parameter value θ by a measurement error (quantified through a known standard error σ_i) *plus* an additional offset due to *between-study variation*, whose expected magnitude is given through the heterogeneity τ. So for given data $(y_i,\, \sigma_i,\, i=1\ldots,k)$ there are two unknowns, the overall effect θ, the parameter of primary interest, and the heterogeneity τ, which constitutes a nuisance parameter. If the heterogeneity τ was known, the (conditional) maximum likelihood estimate of θ would be a weighted average

$$\hat{\theta}_\tau \;=\; \frac{1}{\sum_i \frac{1}{\tau+\sigma_i}} \sum_{i=1}^{k} \frac{y_i}{\tau+\sigma_i}. \tag{2}$$

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

The uncertainty in τ however complicates the problem (Hedges and Olkin, 1985; Hartung et al., 2008).

2 The Bayesian approach

2.1 Prior

We assume that the prior proability density may be factored into $p(\tau, \theta) = p(\tau) \times p(\theta)$. For the prior on the overall effect, $p(\theta)$, we assume either a uniform or a normal prior, leaving the heterogeneity prior $p(\tau)$ unspecified for the moment.

2.2 Likelihood

The likelihood follows from the random effects model specification (1) as

$$\log\big(p(\vec{y}|\theta,\tau,\vec{\sigma})\big) \;=\; -\tfrac{1}{2}\left(k\log(2\pi)+\sum_{i=1}^{k}\log\big(\tau^2+\sigma_i^2\big)+\sum_{i=1}^{k}\frac{(y_i-\Theta)^2}{\tau^2+\sigma_i^2}\right). \quad (3)$$

2.3 Marginal likelihood

The likelihood may be marginalized over θ analytically; using an improper uniform prior over the entire real line for θ we get the marginal likelihood

$$\begin{aligned}\log\big(p(\vec{y}|\tau,\vec{\sigma})\big) \;=\; -\frac{1}{2}\Bigg((k-1)\log(2\pi)+\sum_{i=1}^{k}\log\big(\tau^2+\sigma_i^2\big)\\ +\sum_{i=1}^{k}\frac{\big(y_i-\hat{\theta}_\tau\big)^2}{\tau^2+\sigma_i^2}+\log\Bigg(\sum_{i=1}^{k}\frac{1}{\tau^2+\sigma_i^2}\Bigg)\Bigg), \quad (4)\end{aligned}$$

where $\hat{\theta}_\tau$ is the conditional posterior mean of $\theta|\tau$ (see eqn. (2)). A similar expression results when assuming a normal prior distribution for θ.

2.4 Marginal and joint posterior

Having derived the marginal likelihood, the (one-dimensional) marginal posterior density of τ is

$$p(\tau|\vec{y},\vec{\sigma}) \;\propto\; p(\vec{y}|\tau,\vec{\sigma})\times p(\tau). \quad (5)$$

Integration can now easily be done numerically for arbitrary priors $p(\tau)$. Re-writing the joint posterior as

$$p(\theta,\tau|\vec{y},\vec{\sigma}) \;=\; p(\theta|\tau,\vec{y},\vec{\sigma})\times p(\tau|\vec{y},\vec{\sigma}), \quad (6)$$

it is evident that inference on θ and τ is possible based on the marginal and conditional posterior distributions of τ and $\theta|\tau$. As the conditional posterior of $\theta|\tau$ again is normal, computations become relatively straightforward.

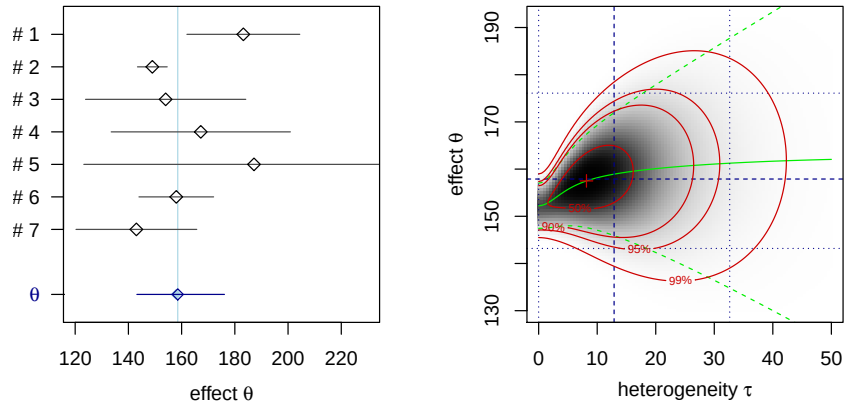


FIGURE 1. *Left panel*: a forest plot of the data. The estimate, shown at the bottom in blue, corresponds to the marginal posterior median and 95% confidence interval. *Right panel*: the joint posterior density of the two unknowns τ and θ. Blue lines indicate the marginal posterior medians and 95% credibility intervals of both parameters. The red contour lines approximately correspond to 50%, 90%, 95% and 99% credibility regions. The green lines show the conditional mean and 95% credibility interval of θ as a function of τ.

3 Application / example

As an example application we will analyze the data by Cochran (1954, Example 3, p. 199). The data are comprised of 7 estimates with associated standard errors, which we assume to be precisely known here. For the following analysis, we assume improper uniform priors on both τ and μ. Figure 1 illustrates the data along with the joint posterior density of the two unknowns θ and τ. The joint posterior has its mode at ($\tau = 8.17$, $\theta = 157.5$), marked by a red cross; for the uniform prior this coincides with the maximum likelihood estimate. The red contour lines approximately correspond to credibility regions as labelled in the figure. The solid green line shows the conditional posterior mean of $\theta|\tau$ as a function of the heterogeneity τ, along with corresponding (conditional) 95% confidence limits (dashed lines).

Integrating over single parameters then yields the marginal distributions of θ and τ; marginalization over the effect τ may be done using the above formula, and marginalization over θ may be implemented e.g. via a simple grid approximation, by averaging over the normal conditionals along a discrete set of τ values. Figure 2 shows the two marginal posterior probability densities for the example data. Dashed lines indicate medians and dotted lines show the shortest 95% credibility intervals. The same values are also marked by blue lines in the joint density plot in Figure 1.

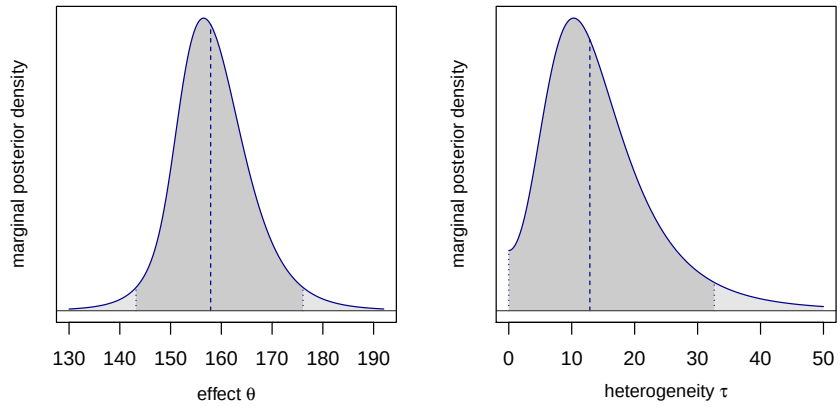


FIGURE 2. Marginal posterior densities of the two unknowns τ and θ. Dashed lines show marginal posterior medians, and dotted lines indicate the shortest 95% credibility intervals.

4 Conclusions

A range of "frequentist" approaches are available for combining (possibly) heterogeneous estimates within the random-effects meta-analysis framework, providing a range of estimates for the two involved parameters (Sidik and Jonkman, 2007). These usually require the determination of a unique heterogeneity estimate (possibly involving a significance test) before proceeding to infer the effect of primary interest.
The Bayesian approach on the other hand leads to unique procedures once the prior distribution is specified. Marginalization allows to infer single parameters while accounting for uncertainty in the other. In the random-effects model, integration may be done semi-analytically, providing coherent results with limited computational effort.

References

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, **10**(1):101-129.

Hartung, J., Knapp, G., and Sinha, B.K. (2008). *Statistical meta-analysis with applications*. Hoboken, NJ: Wiley.

Hedges, L.V., and Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Sidik, K., and Jonkman, J.N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, **26**(9):19641981.

Nonlinear mixed-effects models for bioequivalence on pharmacokinetic data

C. M. Russo[1], S. P. Willemsen[2], D. Leão[1], E. Lesaffre[23]

[1] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, CEP: 13566-590, Brazil,
[2] Department of Biostatistics, Erasmus Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands
[3] L-BioStat, KU Leuven, Kapucijnenvoer 35 blok d - box 7001 3000 Leuven, Belgium

E-mail for correspondence: cibele@icmc.usp.br

Abstract: A bioequivalence assessment of two drugs compares their therapeutic effectiveness. The biological motivation behind this problem often leads to highly nonlinear mixed-effects models (NLMEMs) with interpretable parameters. However, most of these models are difficult to fit and frequently involve complex algorithms, whose convergence depend on the initial values and often requires data transformation. We propose a flexible alternative to the usual parametric models, inspired by a Multivariate SuperImposition by Translation and Rotation (MSITAR) model. A fully parametric nonlinear mixed-effects model is also considered for comparison. To this end we consider a Bayesian approach and illustrate it on a real data set where the interest lies in contrasting the average bioequivalence of a test and reference formulation of an antihypertensive drug.

Keywords: SITAR model; nonlinear mixed-effects models; bioequivalence.

1 Introduction

The importance of mixed-effects models for analysing longitudinal data and repeated measurements is unquestionable. For nonlinear mixed-effects models (NLMEMs), numerical integration is usually required to obtain the likelihood, as well as iterative algorithms with reasonable initial values and possibly data transformations. Challenging applications are found in the pharmacokinetic area, where highly nonlinear models are used to model the absorption and elimination of a substance in the body.
In this context, compartment models are popular to model the pharmacokinetic characteristics of the absorption-elimination process of a substance in the body. An important application area is assessing the bioequivalence

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

of two drugs. In this respect, regulatory agencies FDA and EMEA recommend the comparison of bioavailability parameters, i.e. the maximum drug concentration achieved (C_{max}) and the area under the drug concentration-time curve (AUC) (Chen and Huang, 2013). Test and reference drugs are then considered bioequivalent if the ratio of these parameters for the two drugs is close to 1 and the uncertainty is small enough to exclude all relevant differences. The most common techniques to assess bioequivalence are non-compartmental analysis (NCA) and nonlinear mixed-effects models (NLMEM). A comparison between these two methods was presented in Dubois et al. (2010). A semi-parametric Bayesian test for bioequivalence was proposed by Ghosh and Gönen (2008). More recently, a stochastic approximation expectation-maximisation (SAEM) algorithm was proposed by Dubois et al. (2011), as well as a Wald test to assess bioequivalence.
In this paper we propose a Bayesian analysis that is based on template derived from the median curve of the data, instead of a model that is inspired by a biological theory. Our approach is based on the Multivariate SITAR model, recently proposed by Willemsen et al. (2014) in the context of growth curve modelling problem. This model is an extension of the SuperImposition by Translation and Rotation (SITAR) model (Cole et al., 2010). An important advantage of the SITAR model is that it often provide better fits to the longitudinal data than purely parametric models, which may lack flexibility. We argue that our approach may also provide a better fit to the pharmacokinetic profiles and therefore will enjoy a greater reliability of the bioequivalence conclusion. In addition, tests on bioequivalence may be easily performed by using the (Bayesian) credible intervals for the bioavailability parameters C_{max} and AUC. As a disadvantage, the SITAR regression parameters are no longer interpretable as in the parametric biological model.

2 Motivating data set

A crossover randomized study with 24 volunteers was set up to compare two formulations of the antihypertensive Losartan. The drug concentration was observed at fixed time points, as shown in Figure 1. Reference and test formulations of the drug were administered to each individual on separate days, in a randomised order.

3 Nonlinear models with flexible random components

The bioequivalence of drugs is usually assessed by a non-compartmental analysis (NCA) or NLMEMs. In this section we motivate the first-order compartment problem with the usual full parametric model. A flexible multivariate semi-parametric SITAR-type model is proposed as an efficient alternative.

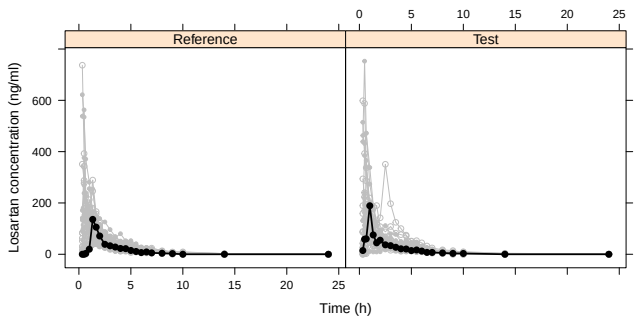


FIGURE 1. Measurements of Losartan concentration in 24 volunteers using two drug formulations, with the measurements of a randomly chosen individual observation highlighted.

3.1 The first order compartment full parametric model

The problem of modeling absorption and elimination of a substance by the body is usually tackled with a compartment model, where the body represents the compartment where the substance is injected and from where it is naturally eliminated afterwards. A popular pharmacokinetic model for the median concentration of the substance Y at the time T is

$$\mathrm{E}(Y) = \exp(lK_a + lK_e - lC_l)\frac{[\exp(-e^{lK_e}T) - \exp(-e^{lK_a}T)]}{e^{lK_a} - e^{lK_e}},$$

with parameters that are interpreted as the logarithm of the substance absorption rate (lK_a), the logarithm of the substance elimination rate (lK_e) and the logarithm of plasma clearance (lC_l).
Denote by $\mathbf{y}_{ik} = (y_{i1k}, \ldots, y_{imk})^\top$ the vector of m observed substance concentrations in the ith subject after the administration of the kth drug, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. In the motivating example, $K = 2$ since two different formulations were administered to each individual on different days. In this case the measurements obtained on the i-th individual of the reference drug concentration $\mathbf{y}_{i1}$ represent the first vector of responses while the measurements of the test drug $\mathbf{y}_{i2}$ represent the second vector of responses. Thus a possible mixed-effects model for the jth response of the kth drug formulation of the ith individual is given by

$$y_{ijk} = \exp(\varphi_{i1k}+\varphi_{i2k}-\varphi_{i3k})\frac{[\exp(-e^{\varphi_{i2k}}T_{ij}) - \exp(-e^{\varphi_{i1k}}T_{ij})]}{e^{\varphi_{i1k}} - e^{\varphi_{i2k}}}+\epsilon_{ijk}, \quad (1)$$

where $\varphi_{i1k} = lK_{ak} + b_{i1k}$, $\varphi_{i2k} = lK_{ek} + b_{i2k}$ and $\varphi_{i3k} = lC_{lk} + b_{i3k}$; lK_{ak}, lK_{ek} and lC_{lk} are the fixed-effects related to the k-th response; b_{i1k}, b_{i2k} and b_{i3k} are the respective random effects. The usual assumption for the vector of random effects related to the i-th individual is $\mathbf{b}_i \sim \mathbf{N}(\mathbf{0}, D_b)$, where $\mathbf{b}_i = (\mathbf{b}_{i1}^\top, \ldots, \mathbf{b}_{iK}^\top)^\top$ and $\mathbf{b}_{ik} = (b_{i1k}, b_{i2k}, b_{i3k})^\top$ is the vector of random effects related to the k-th response of the i-th individual. The random errors $\epsilon_{ijk} \sim N(0, \sigma_e^2)$ are assumed to be mutually independent and independent

of the random effects. To assess bioequivalence C_{max} and AUC will be compared. In the Bayesian context, credible intervals for the differences of C_{max} and AUC for both the drugs formulations may be considered.

3.2 The SITAR model

The SITAR model proposed by Cole et al. (2010) provides a versatile tool for fitting highly non-linear longitudinal data and can be applied to different types of data sets. The original SITAR model is based on vertically and/or horizontally shifting and/or stretching a mean curve in different directions. This enables the model to take into account the correlation between measurements and delivering flexible fitted profiles being guided by the data variability. A multivariate version of the SITAR (MSITAR) model was proposed recently by Willemsen et al. (2014). In this paper we propose a different MSITAR-type model with now two stretch effects (vertical and horizontal) and one shifting effect.
Namely, a possible model for the kth response at time t_{ij} of individual i, y_{ijk}, is:

$$\begin{array}{l} y_{ijk}|\boldsymbol{\gamma}_{ik}, \boldsymbol{\beta}_k, \sigma_k^2 \sim N(\exp(\gamma_{i2k})[T_{ijk}^\top \boldsymbol{\beta}_k], \sigma_k^2) \\ T_{ijk} = B(\exp(\gamma_{i3k})(t_{ij} + \gamma_{i1k})) \end{array} \tag{2}$$

where T_{ijk} is a matrix with bases of cubic splines, $\boldsymbol{\gamma}_{ik} = (\gamma_{1ik}, \gamma_{2ik}, \gamma_{i3k})^\top$ is the vector of random-effects for the kth response of the ith individual, $i = 1, \ldots, N$. It is assumed that $\boldsymbol{\gamma}_i = (\boldsymbol{\gamma}_{i1}^\top, \ldots, \boldsymbol{\gamma}_{iK}^\top)^\top \sim \mathbf{N}(\mathbf{0}, D_\gamma)$.
It is important to notice that, although the SITAR-type model fits a model guided by the data, the parameters are no longer interpretable as for the popular parametric nonlinear mixed-effects models.

4 Application

The Bayesian approach is considered to fit the parametric and SITAR models and assess the bioequivalence of the curves. For the full parametric nonlinear mixed-effects models, independent vague normal priors were considered for the fixed-effects parameters $\beta_i \sim N(0, 1000)$. Further, an inverse gamma was assumed for the measurement error variance, i.e. $\sigma_e^2 \sim IG(0.001, 0.001)$ and an inverted Wishart non-informative prior was considered for the covariance matrix of the random effects vector, $D_b \sim IW(J \times I, J)$, where J is the dimension of the random effects vector. For the MSITAR model, the same independent vague normal priors were considered for the fixed-effects parameters, an inverse gamma was assumed for the measurement error variance, $\sigma_k^2 \sim IG(0.000001, 0.000001)$ and an inverted Wishart non-informative prior , $D_\gamma \sim IW(J * I, J)$ for the covariance matrix of the random effects vector, where J is the dimension of the random effects vector. MCMC sampling was used to estimate the parameters, making use of JAGS. The models were compared using the deviance information criterion - DIC. The parametric fit produced DIC=10645.1

and the MSITAR model produced DIC = 10268.5, which implies a considerable improvement of the fit for the MSITAR model. For the kth drug, samples of $C_{max,k}$ and AUC_k are obtained within the MCMC procedure, with the estimated median curve with a grid of 200 uniformly-spaced values for the time. The trapezoidal rule is used to approximate the area under the curve. Credible intervals for the C_{max} ratio, $C_{max,1}/C_{max,2}$ and AUC ratio, AUC_1/AUC_2 are constructed. The drugs are considered bioequivalent if the value 1 belongs to both of the credible intervals. Figures 2 and 3 illustrate the median fitted curves obtained with parametric and SITAR fits, respectively.

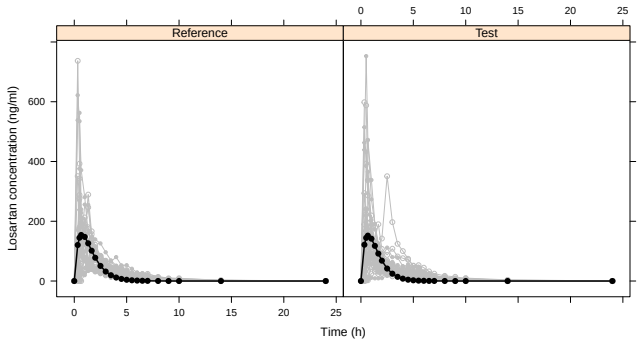


FIGURE 2. Median curves obtained by the parametric model fitted to the pharmacokinetic data.

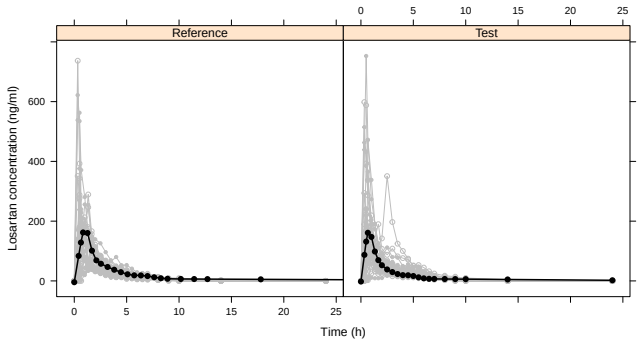


FIGURE 3. Median curves obtained by the SITAR model fitted to the pharmacokinetic data.

5 Discussion

To assess bioequivalence, we propose in this paper the use of a multivariate SITAR model. Our approach is data-driven, i.e. the MSITAR model is

based on a template curve that is obtained from a smooth fit to the data as opposed to following a parametric form based on biology. Such parametric models, while having a theoretical basis might provide a poor fit to the data, but definitely are quite difficult to handle computationally in contrast to the MSITAR model. A Bayesian approach to the bioequivalence problem simplified the comparison of the two formulations of the test drug.

References

Cole, T. J., Donaldson, M. D. & Ben-Shlomo, Y. (2010). 'SITAR a useful instrument for growth curve analysis', International journal of epidemiology 39, 1558–1566.

Chen, Y. I. and Huang, C. S. (2013). New approach to assess bioequivalence parameters using generalized gamma mixed-effect model (model-based asymptotic bioequivalence test). Statistics in Medicine. DOI: 10.1002/sim.5978.

Cole, T. J., Donaldson, M. D. and Ben-Shlomo, Y. (2010). Sitar – a useful instrument for growth curve analysis, *International journal of epidemiology* **39**(6), 1558–1566.

Dubois, A., Gsteiger, S. and Pigeolet, E. and Mentré, F. (2010). Bioequivalence tests based on individual estimates using non-compartmental or model-based analyses: evaluation of estimates of sample means and type I error for different designs, *Pharmaceutical research* **27**, 92–104.

Dubois, A., Lavielle, M., Gsteiger, S., Pigeolet, E. and Mentré, F. (2011)., Model – based analyses of bioequivalence crossover trials using the stochastic approximation expectation maximisation algorithm, *Statistics in Medicine* **30**, 2582–2600.

Ghosh, P. and Gönen, M. (2008). Bayesian modeling of multivariate average bioequivalence, *Statistics in Medicine* **27**, 2402-2419.

Willemsen, S. P., Eilers, P. H., Steegers-Theunissen, R. and Lesaffre, E., (2014). A multivariate bayesian model for human growth, submitted to *Statistics in Medicine*.

Statistical Estimation of Pollution Reductions for Meeting Government Targets

Carl Scarrott[1]

[1] University of Canterbury, New Zealand

E-mail for correspondence: carl.scarrott@canterbury.ac.nz

Abstract: Regional Councils in New Zealand (NZ) are responsible for ensuring that government environmental standards for air quality are met. The standard for the daily average PM_{10} is $50\mu g/m^3$, with the number of exceedances to be no more than three per year by 2016, and one by 2020. Currently, emission reductions are set to ensure an average second or fourth largest concentration per year meets this standard. This method ignores the uncertainty due to the large inter-annual meteorological variation and will underestimate the reductions required to be confident the standards are met. A generalised additive mixed model is developed to describe the meteorological and historical trends in the PM_{10} concentrations in a large rural township Timaru, New Zealand. A simulation study using bootstrapped meteorological conditions is implemented to determine the percentage reductions needed to achieve these future targets with a certain likelihood. The results are used to inform air quality management policy and planning.

Keywords: generalised additive mixed model; block bootstrap; simulation; pollution modelling.

Many of New Zealand's cities and towns suffer from severe air pollution over winter, predominantly due to the use of solid fuel burners for home heating. The concentration of airborne particulate matter of the order of 10 micrometres or less (PM_{10}) regularly exceed the World Health Organisation (WHO) guideline of a 24 hour average of $50\mu g/m^3$. Such air pollution has a significant impact on health, increases morbidity and has a high economic cost. Kuschel *et al.* (2012) estimated the cost to the NZ economy of the social effects (health, lost work days, premature deaths, etc.) of anthropogenically sourced air pollution to be over NZ$4 billion (euro 2.5 billion) per annum.

The National Environmental Standards for Air Quality (NESAQ) set targets for concentrations of many such pollutants. The PM_{10} target level is the WHO guideline of a $50\mu g/m^3$ with at most 3 exceedances per year by 1 September 2016 (i.e. end of winter) and at most one exceedance by 1 September 2020, which are equivalent to the fourth and second largest

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

concentration being at most $50\mu g/m^3$. The NESAQ specifies that the 24 hour average is defined as midnight to midnight, which is inconsistent with the typical air pollution generating and dispersal process. Determining the required reduction in emissions is challenging due to the impact of meteorology both on particulate emissions (e.g. influencing burner usage) and directly on the concentrations, thus leading to substantial intra- and inter-annual variability in the number of exceedances. The target reductions must therefore account for the distribution of possible exceedances observable across both 'good' and 'bad' winters.

An objective statistical basis is outlined for determining the target reductions in concentrations, in order to meet these future standards with a prescribed probability. The major source of uncertainty in setting these target reductions is the meteorological variation. Therefore, the first step is to develop a generalised additive mixed model (Wood, 2006) to describe the impacts of meteorology and historical trends on the concentrations, including capturing a lag one autocorrelated error structure in the residuals. A Monte Carlo simulation study then uses the fitted model, with block bootstrapped meteorology, to determine the total reduction needed to achieve the targets with a prescribed likelihood in 2016/2020.

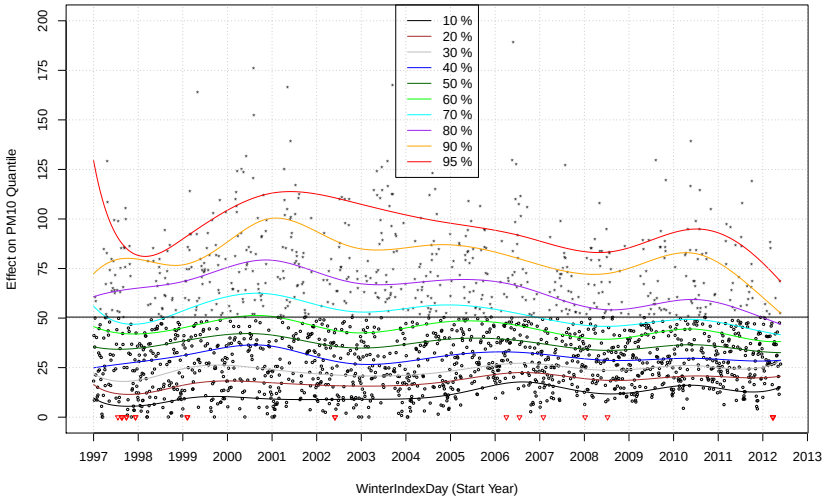


FIGURE 1. Daily average PM_{10} concentrations in $\mu g/m^3$ for Timaru for each winter 1997 to mid-2012. Missing values are shown in red at zero concentration. Quantile regression based smooth trends shown by coloured lines.

1 Statistical Modelling and Simulation

A generalised additive mixed model (GAMM) is used to provide a flexible description of the observed meteorological effects, along with a slowly varying trend in the historical PM_{10} concentrations. As the majority of ex-

ceedances occur in winter, the model is fit to only the May-August concentrations which are depicted in Figure 1. The emissions from the non-winter periods are from different sources, so provide limited information for predicting meteorological impacts on winter concentrations. Estimated smooth trends in the quantiles are shown Figure 1 for visualisation purposes.
Alternative GAMM formulations, with different assumed distributions for the concentrations, meteorological variables and correlated errors were considered to ensure the results are robust to any specific set of assumptions. These were fit using the `mgcv` library in `R`. The results presented below assume a normal family with log-link function. The trend is assumed to be slowly varying in time, described by smooth thin plate splines (Wood, 2003) which use around 8 degrees of freedom over this time period. Appropriately transformed meteorological predictors are included as linear terms, as insufficient evidence was found for need for non-linear effects. The residual autocorrelation is captured well by a lag one autoregressive structure with estimated coefficient of 0.29. The meteorological predictors were chosen by stepwise selection, with external validation against variables selected from regression tree approaches and subject matter experts.

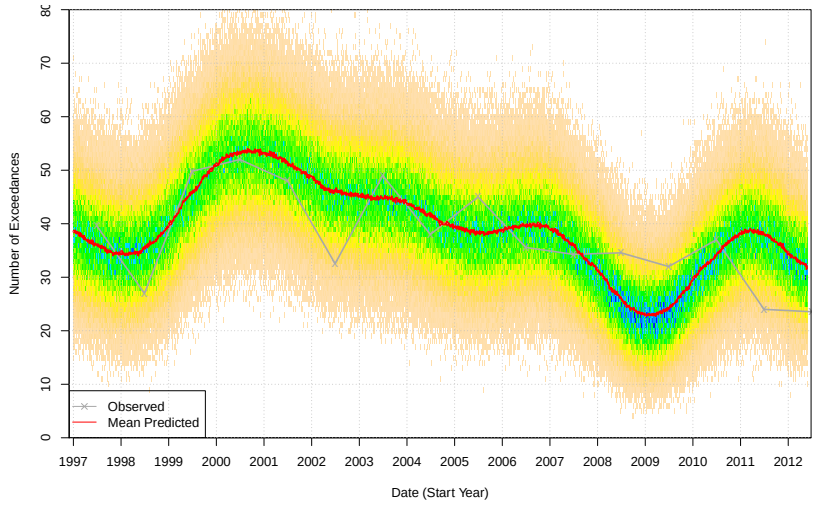


FIGURE 2. Distribution of predicted number of exceedances resulting from simulated winters assuming that particular day's trend applies over the whole year. The mean of the distribution is shown in red. The observed number of exceedances per year is shown by grey points in middle of year.

The fitted distribution of number of exceedances over time are shown in Figure 2, with the observed number per year in grey. Note that the observed number of exceedances in 2012 are low as it was only partially observed in the available dataset. The Monte Carlo simulations from the GAMM used to produce the estimated distribution of exceedances (shown by colour density image), are based on realisations of potential observed winter concentrations generated by:

1. resampling the meteorological conditions for each of the 123 days of winter using a randomised block size bootstrap using an average period of 5 days (Davison & Hinkley, 1997);

2. applying fitted GAMM with simulated meteorology and estimated trend for that day applied over the entire winter period; and

3. simulate observed concentrations from the assumed distribution to account for residual variation.

In the resampling in step 1, the measurements of all the meteorological variables within a day are resampled together, thus retaining the joint dependence between them.
The predicted number of exceedances for the 123 days of winter is extracted, from 10,000 realised target years to provide the distribution of the number exceedances. The fitted (conditional) distribution of the number of exceedances at each timepoint reasonably describes that observed, once the meteorological variation is accounted for. In particular, the estimated expected number of exceedances shown in red reasonably follows the observed number shown in grey.
The Monte Carlo simulations from the fitted GAMM are used to evaluate the total percentage reduction in the concentrations needed to meet the future targets with a prescribed probability. The thin plate spline based trend term in the log-link function of the GAMM only applies over historical timepoints, so cannot be extrapolated to the future. For setting the targets and limits detailed below the thin plate smoother is replace with a constant multiplicative trend on the log-link scale given by:

$$log(\mu) = \mathbf{X}\beta + \beta_{date} t_{date}$$

where $\mathbf{X}$ represents the simulated meteorological effects and the corresponding coefficients β are replaced with their estimated values from the GAMM. The variable t_{date} is the number of years since mid-winter 2012 and therefore $\exp(\beta_{date})$ is the annual reduction in the predicted concentrations. The simulation study determines the value of β_{date}, or more usefully, the implied total percentage reduction in the predicted concentrations such that the standard is achieved in each target year with a prescribed probability (50% and 95% chance considered below).
Simulated future winter concentrations in each target year (2016 & 2020) are generated using the above simulation scheme, with the trend estimate per winter day in step (2) replaced with the estimated target reduction. Target values for other annual characteristics are obtained, including the average concentration, number of exceedances, second and fourth largest. Limits on the possible variation in these annual statistics are also obtained, which indicate the range of tolerable variation from the annual targets if the airshed really is on target to meet these future targets.
The estimated total reduction in average PM_{10} concentrations to achieve the NESAQ target in 2016 with a 50% chance is estimated to be 51.4% as in

TABLE 1. Estimated targets for PM_{10} concentrations assuming a constant annual reduction rate, with 50% and 95% chance of meeting the 2016 target.

50% Chance of Meeting 2016 Target					
		2013	2014	2015	2016
	Annual % Reduction		***16.5***		
	Cumulative % Reduction	16.5	30.3	41.8	***51.4***
Target	# of Exceedances	21	13	7	3
Values	Fourth Largest	75	65	57	50.5
90% Upper	# of Exceedances	28	18	11	6
Limit	Fourth Largest	83	72	63	56
95% Chance of Meeting 2016 Target					
	Annual % Reduction		***21.6***		
	Cumulative % Reduction	21.6	36.6	51.9	***62.3***
Target	# of Exceedances	18	8	3	1
Values	Fourth Largest	71	59	50	44
90% Upper	# of Exceedances	24	13	6	3
Limit	Fourth Largest	79	66	56	49

Table 1, which is an annual reduction of 16.5%. Notice that the target value in 2016 is 3 exceedances and the fourth largest concentration is $50.5\mu g/m^3$, as you would expect as observed values of $50.5\mu g/m^3$ are rounded to the nearest integer value which would therefore make the fourth largest just obtain the definition of an exceedance.

The 90% upper limits are determined from the sample quantile of distribution of the simulated number of exceedances and fourth largest concentrations per year obtained from the process is on-target to reach the required reductions. Notice that the upper limits are of course higher then target levels. Of course the 10% chance of exceedance of these limits is an annual risk, so the probability of multiple exceedances in consecutive years will be correspondingly much smaller.

If the chance of meeting the future target is increased to 95% then the required reductions are higher at 21.6% per year and 62.3% in total by 2016. The target level for the fourth largest in 2016 has also reduced from $50.5\mu g/m^3$ to $44\mu g/m^3$. Essentially, these provide the required extra margin to account for meteorological uncertainty in the target years.

Similar behaviour of the targets and limits for the year 2020 are observed in Table 2. It is also clear that the 2016 NESAQ target is harder to achieve than the 2020 target, in the respect that a larger annual reduction in the concentrations is required. Notice in particular that the extra reduction margin to achieve the 2020 target, of at most one exceedance per year, with a 95% chance gives a target of no exceedances in 2019 and 2020.

Acknowledgments: The contributions of John Newell, Marco Reale, Shakira Suwan and Jacky Sung are acknowledged for their contribution to early work on the city of Christchurch. Advice from staff at Environment Can-

TABLE 2. Estimated targets for PM_{10} concentrations assuming a constant annual reduction rate, with 50% and 95% chance of meeting the 2020 target.

		2013	2014	2015	2016	2017	2018	2019	2020
50% Chance of Meeting 2020 Target									
Annual % Reduction					***10.8***				
Cumulative % Reduction		10.8	20.5	29.1	36.8	43.7	49.8	55.2	***60.1***
Target	# Exceed.	26	19	14	9	6	4	2	1
Values	Fourth Larg.	88	80	73	67	62	58	54	50.5
90% Up	# Exceed.	33	25	19	14	10	7	5	3
Limit	Fourth Larg.	99	90	83	76	70	66	61	58
95% Chance of Meeting 2020 Target									
Annual % Reduction					***15.3***				
Cumulative % Reduction		15.3	28.3	39.2	48.5	56.4	63.1	66.7	***73.5***
Target	# Exceed.	23	14	8	4	2	1	0	0
Values	Fourth Larg.	84	74	65	59	53	49	45	42
90% Up	# Exceed.	29	20	13	8	4	3	2	1
Limit	Fourth Larg.	95	84	74	66	60	55	52	48

terbury (Teresa Aberkane, Vincent Salomon, Angie Scott and Tim Mallett) is also gratefully acknowledged. This work was partially completed whilst on sabbatical at the National University of Ireland in Galway.

References

Davison A.C. and Hinkley D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Kuschel, G. Metcalfe, J., Wilton, E., Guria, J., Hales, S., Rolfe, K. and Woodward, A. (2012). *Updated Health and Air Pollution in New Zealand Study*. HAPINZ Technical Report `http://www.hapinz.org.nz`.

Politis, D.N. and Romano, J.P. (1994). The stationary bootstrap. Journal of the American Statistical Association, 89, 13031313.

Wood S. (2003). *Thin plate regression splines*. Journal of the Royal Statistical Society **B** 65(1), pp 95-114. CRC Press.

Wood S. (2006). *Generalised Additive Models: An Introduction with R*. CRC Press.

DIFboost: A Boosting Method for the Detection of Differential Item Functioning

Gunther Schauberger[1], Gerhard Tutz[1]

[1] Ludwig-Maximilians-Universität München, Germany

E-mail for correspondence: gunther.schauberger@stat.uni-muenchen.de

Abstract: Methods for the identification of Differential Item Functioning (DIF) in Rasch models are typically restricted to the case of two subgroups. A boosting algorithm is proposed that is able to handle the more general setting where DIF can be induced by several covariates at the same time. The covariates can be both metric and (multi-) categorical. The method works for a general parametric model for DIF in Rasch models and competes well with traditional DIF methods.

Keywords: Rasch model; Differential Item Functioning; Boosting; DIFboost.

1 Differential Item Functioning Model

In the binary Rasch model, the probability for a person to score on an item is determined by a parameter for the latent ability of the person and a parameter for the item difficulty. In the case of P persons and I items, the Rasch model is given by

$$\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) = \theta_p - \beta_i, \quad p = 1,\dots,P \quad ,i = 1,\dots,I, \tag{1}$$

where Y_{pi} represents the response of person p on item i. It is coded by $Y_{pi} = 1$ if person p solves item i and $Y_{pi} = 0$ otherwise. Both the person parameters θ_p and the item parameters β_i are unknown and have to be estimated. For identifiability, we set $\theta_P = 0$.
In item response models, DIF occurs if an item has different difficulties depending on characteristics of the participants. This concept can be formalized by

$$\log\left(\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}\right) = \theta_p - (\beta_i + \boldsymbol{x}_p^{\mathrm{T}}\boldsymbol{\gamma}_i), \tag{2}$$

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

where $\boldsymbol{x}_p^{\mathrm{T}} = (x_{p1}, \ldots, x_{pm})$ denotes a person-specific covariate vector of length m and, again, the restriction $\theta_P = 0$ is used. This general Differential Item Functioning Model (DIF model) has been used by Tutz und Schauberger (2013a). It is an extension of the Rasch model (1) allowing for person-specific item difficulties $\beta_i + \boldsymbol{x}_p^{\mathrm{T}}\boldsymbol{\gamma}_i$.
In the general DIF model (2), one has, additionally to the parameters from the Rasch model, $m \cdot I$ parameters that describe DIF. Classical estimation procedures tend to fail or yield extremely unstable estimates. Moreover, a general assumption of the model is that only some of the items show DIF. Therefore, only for these items item-specific parameters $\boldsymbol{\gamma}_i$ have to be estimated. It turns out that for this problem boosting procedures with automatic variable selection (Bühlmann and Hothorn, 2007) are a very useful tool to obtain efficient estimates.

2 The DIFboost Algorithm

The objective of our approach is to detect DIF by boosting the logistic DIF model (2). Since the person parameters θ_p and the item parameters β_i are essential for the interpretability of the model, it is sensible to start the boosting selection procedure after a basic Rasch model has been fitted. With the estimtates from the Rasch model, for a single observation a linear predictor $\hat{\eta}_{pi} = \hat{\theta}_p - \hat{\beta}_i$ can be calculated. The linear predictors for all person-item combinations are passed on to the further steps of the algorithm.
For the boosting steps, the Rasch model (1) is extended to the more general DIF model (2). The parameters of the DIF model determine the base learners that are used. In our case, three types of base learners are useful:

$$\tilde{\eta}(\boldsymbol{x}_p, p, i) = \begin{cases} \tilde{\theta}_p, & p = 1, \ldots, P-1 \\ \tilde{\beta}_i, & i = 1, \ldots, I \\ \boldsymbol{x}_p^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}_i, & i = 1, \ldots, I \end{cases} \tag{3}$$

In every boosting step, only one of the base learner is updated, namely the one which yields the strongest reduction of an adequate loss function $L(Y_{pi}, \tilde{\eta}_{pi})$, which can be denoted by

$$\tilde{\eta}^*(\boldsymbol{x}_p, p, i) = \underset{\tilde{\theta}_p, \tilde{\beta}_i, \boldsymbol{x}^{\mathrm{T}}\tilde{\boldsymbol{\gamma}}_i}{argmin} \sum_{p,i} L(Y_{pi}, \tilde{\eta}_{pi}).$$

The estimates for the single candidates of the base learner are obtained by fitting logit models where the linear predictor from the current model fit is used as known offset and the respective base learner is the only predictor. Therefore, in every step only the base learner with the highest gain of information is updated. An additional parameter ν, $0 < \nu < 1$, regulates the step size of the parameter updates. It is chosen sufficiently small ($\nu = 0.1$) and is used to prevent overfitting.

3 Stability Selection

For our goal of variable selection, we use the concept of stability selection proposed by Meinshausen and Bühlmann (2010). For a predefined number of replications B, subsamples of $\lfloor \frac{P}{2} \rfloor$ persons are drawn randomly from the original data set. For each of the subsamples, the boosting algorithm is executed until a (predefined) number of q different base learners have been selected. Then, one calculates relative frequencies on how often a specific base learner was selected at each specific step. The frequencies are illustrated by so-called stability paths along the boosting steps (see Figure 1). Finally, all base learners with stability paths beyond a certain threshold value π_0 are selected and a final DIF model is fitted with the selected items.

4 Application

The method was applied to data from the Intelligence-Structure-Test 2000 R (I-S-T 2000 R), developed by Amthauer et al. (2001). We considered a subtest on the topic sentence completion, consisting of 20 items. The data origin from a test on 273 students from different faculties from the university of Marburg, Germany, aged between 18 and 39 years. The data have first been used in Bühner et al. (2006). Three covariates were used as possibly DIF inducing covariates, gender (0: male, 1: female), age (in years) and the interaction between gender and age.

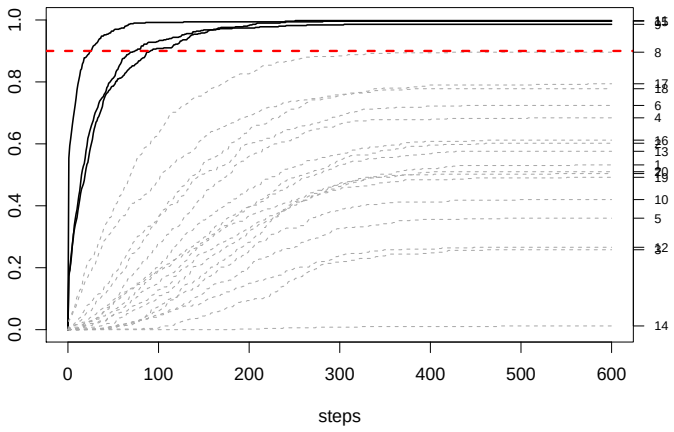


FIGURE 1. Stability paths for all items; dashed line represents the threshold $\pi_0 = 0.9$; items 9, 11 and 15 are diagnosed as DIF items

For our analysis, $B = 500$ subsamples were drawn. Figure 1 shows the stability paths for DIFboost with stability selection. It is seen, that three

items (9, 11, 15) are selected in almost all replications and, therefore, are identified to have DIF.
We illustrate the coefficients of the DIF-items by effect stars (Tutz and Schauberger, 2013b). Since the logit link is used, the exponentials of the coefficients represent the effects of the covariates on the odds $\frac{P(Y_{pi}=1)}{P(Y_{pi}=0)}$. The length of the rays correponds to the exponentials of the respective coefficients. The circle around each star has a radius of $\exp(0) = 1$ and, therefore, represents the no-effect case. Both gender and age were standardized prior to the analysis so that the size of the coefficient estimates will be comparable. Figure 2 shows the effect stars for the estimated coefficients.

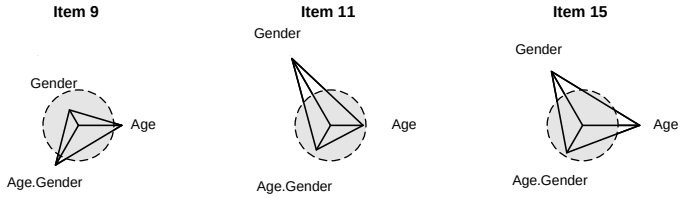


FIGURE 2. Effect stars for the detected DIF items from the subtest sentence completion

Generally, a ray beyond the circle represents positive coefficients. With positive coefficients, the difficulty of the respective item is increased if the corresponding covariate is increased while the probability to solve the item is decreased. Item 9, for example, has a negative coefficient for gender. Therefore, this item is easier for female participants as female is encoded by 1. After all, since also the interaction between gender and age is considered, one has to look at all coefficients at a time. With growing age, the difficulty increases for female participants.
Figure 3 shows for each DIF-item the effects of both gender and age on the probability to score on the respective item. Separately for male (solid lines) and female (dashed lines) participants, the probability to score on the respective item is depicted along the covariate age. For simplicity, the plots refer to a person with a 'mean' ability according to the estimates of the θ parameters.
Figure 3 clarifies the effect of the interaction term. As the probabilities to score on an item can intersect, the main effects of age or gender should not be interpreted separately but always with respect to the interaction term. The ability to include interaction terms in this manner can be seen as a big improvement compared to existing methods of DIF detection allowing for new insights on the occurence of DIF. In extreme cases, both the main effects for gender and age could be neglectible but the interactions term could still be influential.
Therefore, item 9 can not generally be assumed to be easier for female participants. This only holds for participants younger than 30 years while

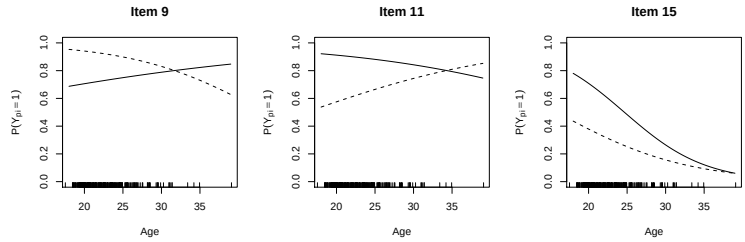


FIGURE 3. Probabilities to score on items depending on gender and age for all DIF-items. Solid lines represent male, dashed lines represent female participants.

it is inversely for older participants. Items 11 and 15 are, in general, easier for male participants, in particular if they are rather young. For growing age this difference slowly vanishes, in item 11 the effect is even reversed for higher age.

References

Amthauer, R., Brocke, B., Liepmann, D., and Beauducel, A. (2001). Intelligenz-Struktur-Test 2000 R (IST 2000 R). *Göttingen: Hogrefe.*

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, **22**, 477 – 505.

Bühner, M., Ziegler, M., Krumm, S., and Schmidt-Atzert, L. (2006). Ist der IST 2000 R Rasch-skalierbar? *Diagnostica*, **52(3)**, 119 – 130.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, **72**, 417 – 473.

Tutz, G. and Schauberger, G. (2013a). A Penalty Approach to Differential Item Functioning in Rasch Models. *Psychometrika*, to appear

Tutz, G. and Schauberger, G. (2013b). Visualization of Categorical Response Models - from Data Glyphs to Parameter Glyphs. *Journal of Computational and Graphical Statistics*, **22(1)**, 156 – 177.

Functional Additive Mixed Models

Fabian Scheipl[1], Ana-Maria Staicu[2], Sonja Greven[1]

[1] Department of Statistics, Ludwig-Maximilians-Universität München, Germany
[2] Department of Statistics, North Carolina State University, USA

E-mail for correspondence: sonja.greven@stat.uni-muenchen.de

Abstract: We propose an extensive framework for additive regression models for correlated functional responses, allowing for multiple partially nested or crossed functional random effects with flexible correlation structures for, e.g., spatial, temporal, or longitudinal functional data. Additionally, our framework includes linear and nonlinear effects of functional and scalar covariates that may vary smoothly over the index of the functional response. It accommodates densely or sparsely observed functional responses and predictors which may be observed with additional error and includes both spline-based and functional principal component-based terms. Estimation and inference in this framework is based on standard additive mixed models, allowing us to take advantage of established methods and robust, flexible algorithms. We provide easy-to-use open source software in the pffr() function for the R-package refund. We evaluate our approach in simulations and two applications with spatially and longitudinally observed functional data.

Keywords: Functional data analysis, functional principal component analysis, P-splines, Smoothing, Varying coefficient models.

1 Introduction and Model

Scientific studies increasingly collect functional data with correlation structures. We are motivated by a longitudinal diffusion tensor imaging study on multiple sclerosis and the benchmark functional data set on weather stations spatially distributed across Canada. We propose regression models for functional responses that accommodate general correlation structures via functional and scalar random effects as well as linear or nonlinear effects of scalar and functional covariates.
We consider structured additive regression models of the general form

$$y_i(t) = \sum_{r=1}^{R} f_r(\mathcal{X}_{ri}, t) + \epsilon_i(t), \tag{1}$$

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

for functional responses $y_i(t), i = 1, \ldots, n$, observed over a domain $\mathcal{T}$. Each term in the additive predictor $\sum_{r=1}^{R} f_r(\mathcal{X}_{ri}, t)$ is a function of a) the index t of the response and b) a subset $\mathcal{X}_r$ of the complete covariate set $\mathcal{X}$ including scalar and functional covariates and (partially) nested or crossed grouping factors.
We assume $\epsilon_i(t)$ to be independent $N(0, \sigma_\varepsilon^2)$ variables for each t. Functional random effects $b_g(t)$ for a grouping variable g with M levels are modeled as realizations of a mean-zero Gaussian random process on $\{1, \ldots, M\} \times \mathcal{T}$ with a general covariance function smooth in t.
Most existing work on functional random effects has considered only special cases such as the functional random intercept model or a two or three-level hierarchy. Morris and Carroll (2006) and subsequent work by this group propose a general Bayesian functional linear mixed model based on a wavelet transformation of (usually very spiky) data observed on an equidistant grid. The proposed model includes correlation between different random effects and heterogeneous residual errors, which we do not. Our approach, on the other hand, is well suited to smooth underlying curves and allows a more general mean structure than previous functional linear mixed models; in particular we are able to estimate smooth nonlinear or linear effects of scalar and/or functional covariates within the same framework. In addition, we are able to handle data on non-equidistant or sparse grids. To the best of our knowledge, our proposal is the first publicly available implementation that allows such a high level of flexibility for a functional regression model.

2 Estimation

For notational simplicity, we assume $y_i(t)$ to be observed on identical grids $t = (t_1, \ldots, t_T)^T$, but irregular/sparse grids are naturally accommodated in the rephrased model formulation given below. Then, model (1) can be expressed as

$$y_{il} = \sum_{r=1}^{R} f_r(\mathcal{X}_{ri}, t_l) + \epsilon_{il}, \quad \epsilon_{il} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, n, l = 1, \ldots, T. \quad (2)$$

The smoothness assumption on $\mathrm{E}(y_i(t))$ is preserved implicitly by enforcing smoothness across $\mathcal{T}$ for all $f_r(\mathcal{X}_r, t)$. Let $y = (y_{11}^T, \ldots, y_{1T} \ldots, y_{nT})^T$ and let $\mathcal{X}_r$ denote the vector or matrix with rows of observations $\mathcal{X}_{ri}$. Let $f(t) = (f(t_1), \ldots, f(t_T))$ and let $f(x, t)$ denote the vector of evaluations of f for each combination of rows in x and t. Let 1_d denote a d-vector of ones. For an $m \times a$ matrix A and an $m \times b$ matrix B denote the row tensor product by $A \odot B = (A \otimes 1_b^T) \cdot (1_a^T \otimes B)$, with element-wise multiplication $\cdot$.
We approximate each $f_r(\mathcal{X}_r, t)$ as a linear combination of basis functions on the product space for $\mathcal{X}_r$ and t, with corresponding marginal penalties. We use the versatile row tensor product of marginal bases (Wood, 2006,

ch. 4.1.8),

$$f_r(\mathcal{X}_r, t) \approx (\Phi_{xr} \odot \Phi_{tr})\theta_r = \Phi_r \theta_r. \tag{3}$$

Φ_{xr} and Φ_{tr} contain the evaluations of suitable marginal bases for the covariate(s) in $\mathcal{X}_r$ and in t, respectively. We choose sufficiently large bases for flexibility, but use a penalized likelihood approach for regularization. The corresponding penalty is the Kronecker sum of marginal penalties P_{xr} and P_{tr} (ch. 4.1, Wood, 2006),

$$\text{pen}(\theta_r|\lambda_{tr}, \lambda_{xr}) = \theta_r^T(\lambda_{xr} P_{xr} \otimes I_{K_t} + \lambda_{tr} I_{K_x} \otimes P_{tr})\theta_r = \theta_r^T P_r(\lambda_{tr}, \lambda_{xr})\theta_r.$$

P_{xr} and P_{tr} are known and fixed positive (semi-)definite penalty matrices and $\lambda_{tr}, \lambda_{xr} > 0$ are smoothing parameters controlling the trade-off between goodness of fit and smoothness of $f_r(\mathcal{X}_r, t)$ in $\mathcal{X}_r$ and t, respectively. In the following, we motivate and define $\Phi_{xr}, \Phi_{tr}, P_{tr}$ and P_{xr} for different effect types. Constant effects over t are associated with $\Phi_{tr} = 1_{nT}$ and $P_{tr} = 0$, while Φ_{tr} and P_{tr} can be chosen freely for terms varying over t.
For a functional intercept $\alpha(t)$ (i.e., $\mathcal{X}_r = \emptyset$), $\Phi_{xr} = 1_{nT}$ and $P_{xr} = 0$. For effects linear in scalar covariates z like $f_r(z, t) = z\delta$ or $f_r(z, t) = z\delta(t)$, the marginal basis reduces to $\Phi_{xr} = z \otimes 1_T$ where $z = (z_1, \ldots, z_n)^T$, with penalty $P_{xr} = 0$. For nonlinear effects of scalar covariates like $f_r(z, t) = \gamma(z)$ or $f_r(z, t) = \gamma(z, t)$, Φ_{xr} is a suitable marginal spline basis matrix over z and P_{xr} is the associated penalty.
For linear effects of functional covariates $f_r(x_i(s), t) = \int_{\mathcal{S}} x_i(s)\beta(s, t)ds$, we follow Ivanescu et al (2013) and model $\beta(s, t)$ using tensor product splines with basis functions $\Phi_{k_s}(s), k_s = 1, \ldots, K_x$, over $\mathcal{S}$ and a basis over $\mathcal{T}$, so $\Phi_{xr} = [x \,\text{diag}(w)\Phi_s] \otimes 1_T \approx [\int_{\mathcal{S}} x_i(s)\Phi_{k_s}(s)ds]_{i=1,\ldots,n;k_s=1,\ldots,K_x} \otimes 1_T$, where $x = [x_i(s_h)]_{i=1,\ldots,n;h=1,\ldots,H}$, $w = (w_1, \ldots, w_H)^T$ contains quadrature weights for numerical integration over $\mathcal{S}$, and

$$\Phi_s = [\Phi_{k_s}(s_h)]_{h=1,\ldots,H;k_s=1,\ldots,K_x}.$$

P_{xr} is the penalty associated with the $\Phi_{k_s}(s)$. By defining suitable weight matrices $w_{i,l}$ with zero entries for $s_h < l_i(t_l)$ and $s_h > u_i(t_l)$ this can be extended to terms like $\int_{l_i(t)}^{u_i(t)} x_i(s)\beta(s, t)ds$. In the limit, this includes the concurrent model $f_r(x_i(t), t) = x_i(t)\beta(t)$.
Non-linear function-on-function effects $\int_{\mathcal{S}} F(x_i(s), s, t)ds$, which generalize McLean et al (2014) from scalar to functional responses, can be similarly represented, please see the full paper (Scheipl et al, 2014) for details.
Functional random effects $b_g(t)$ are represented as smooth functions in t for each level m of g. Functional random intercepts are associated with a marginal basis $\Phi_{xr} = [\delta_{g(i)m}]_{i=1,\ldots,n;m=1,\ldots,M} \otimes 1_T$, where $g(i)$ denotes the level of g for observation i. This yields an incidence matrix mapping the observations to the different levels of the grouping variable. For a functional random slope in z, $\Phi_{xr} = [z_i\delta_{g(i)m}]_{i=1,\ldots,n;m=1,\ldots,M} \otimes 1_T$. The marginal penalty P_{xr} for functional random effects is a $M \times M$ precision matrix that defines the dependence structure between the levels of g. The full penalty

induces a mean zero Gaussian process assumption for $b_g(t)$. Any combination of spline basis, smoothness penalty and between-subject correlation can be used to construct multiple (partially) nested or crossed functional random effects.
For alternative FPC-based representations of linear or nonlinear function-on-function effects as well as functional random effects, we refer to the full paper (Scheipl et al, 2014).
Using the tensor product representation, model (1) can be re-written as

$$y = \Phi\theta + \epsilon; \qquad \epsilon \sim N(0, \sigma_\epsilon^2 I_{nT}), \tag{4}$$

where $\Phi = [\Phi_1|\ldots|\Phi_R]$ and $\theta = (\theta_1^T, \ldots, \theta_R^T)^T$. To clear up notation, we assign a sequential index $v = 1, \ldots, V$ to the smoothing parameters $\lambda_{xr}, \lambda_{tr}$. We pad $P_{tr} \otimes I_{K_x}$ and $I_{K_t} \otimes P_{xr}$ with rows and columns of zeros, denoting these matrices $\tilde{P}_{v_1}$ and $\tilde{P}_{v_2}$, such that $\theta_r^T(\lambda_{tr} P_{tr} \otimes I_{K_x} + \lambda_{xr} I_{K_t} \otimes P_{xr})\theta_r = \lambda_{v_1}\theta^T\tilde{P}_{v_1}\theta + \lambda_{v_2}\theta^T\tilde{P}_{v_2}\theta$. The penalized likelihood criterion to be minimized then becomes

$$\frac{1}{\sigma_\epsilon^2}\|y - \Phi\theta\|^2 + \sum_v \frac{\lambda_v}{\sigma_\epsilon^2}\theta^T\tilde{P}_v\theta \to \min. \tag{5}$$

Let $\tau_v = \sigma_\epsilon^2/\lambda_v$ and obtain the solution $\widehat{\theta}$ of (5) as the best linear unbiased predictor in the linear mixed effects model

$$y \sim N\left(\Phi\theta, \sigma_\epsilon^2 I_{nT}\right); \qquad \theta \sim N\left(0, \left(\sum_v \tau_v^{-1}\tilde{P}_v\right)^-\right), \tag{6}$$

where S^- denotes the generalized inverse of positive semi-definite covariance matrix S, and $N(0, S^-)$ is a partially improper Gaussian distribution. The smoothing parameters $\lambda_v = \sigma_\epsilon^2/\tau_v$ can now be estimated as variance ratios using restricted maximum likelihood (REML), which has been shown to be more stable and result in somewhat lower MSE than generalized cross-validation (GCV) (Reiss and Ogden, 2009). Since the fit criterion (6) corresponds to that of conventional additive mixed models (AMMs) for scalar data, confidence intervals, tests, model selection etc. directly transfer. The full framework for functional AMMs is implemented in the pffr-function in the refund package for R. The underlying inference engine is the mgcv package for generalized additive models.

3 Simulation Study

We simulate data with repeated measures for four scenarios including different combinations of functional random intercepts, functional random slopes, functional and scalar covariates. We vary the number of subjects M, observations per subject n_i, grid points T for t, signal-to-noise ratio and relative importance of the random effects. We here summarize the main results.

Important effects that contribute relevantly to the predictor are estimated with good to excellent accuracy. Only a single replicate resulted in a relative integrated mean squared error for $y(t)$ greater than 0.1 - even in the most challenging data situations with few noisy observations and small group sizes, our approach is able to reproduce the true structure of the data well. Our results indicate that estimation accuracy of covariate effects is affected most strongly by changes in the signal-to-noise ratio and especially the relative importance of the random effects, and less strongly by changes in the available number of observations M, n_i and T. The patterns of relative change in accuracy are identical for simple functional regression coefficients, index-varying smooth effects or effect surfaces for functional covariates. The estimation accuracy of the functional random effects is affected strongly by the relative importance of the random effects and the group size n_i, and little by the number of groups M. FPC-based random effects seem to require a sufficiently large number of groups and low noise level to obtain usable FPC estimates. Spline-based approaches yielded superior results to FPC-based and wavelet-based (Morris and Carroll, 2006) approaches, but it should be noted that the data-generating process for the simulation study was spline-based itself. Overall, the observed coverage of the approximate pointwise confidence intervals was very close to the nominal level except for very small or noisy data.

4 Application: Canadian weather

Due to space, we here do not discuss the longitudinal imaging study, but focus on results for the Canadian weather data. The Canadian weather data consists of temperature and precipitation curves, measured as the monthly averages over several years at 35 Canadian weather stations. The data has been used extensively in the functional data analysis literature. As it is available as part of the R-package fda, we can make our analysis fully reproducible. We will here focus on both the functional relationship between temperature and precipitation profiles as well as on the spatial nature of the data, clearly visible from the locations of the weather stations (Figure 1, middle).

We model log-precipitation $y_i(t)$ using a functional intercept $\alpha_{g_i}(t)$ per climate region g_i, a linear effect of temperature $x_i(s)$ and smooth spatially correlated residual curves $e_i(t)$,

$$y_i(t) = \alpha_{g_i}(t) + \int x_i(s)\beta(s,t)ds + e_i(t) + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2). \tag{7}$$

Results in Figure 1 indicate differences between climate zones, with strongest seasonality in precipitation in the continental region, and remaining spatial correlation between stations after accounting for region. Higher temperatures in fall and winter go hand-in-hand with higher precipitation levels throughout the whole year, while the reverse is true for higher temperatures in spring and summer.

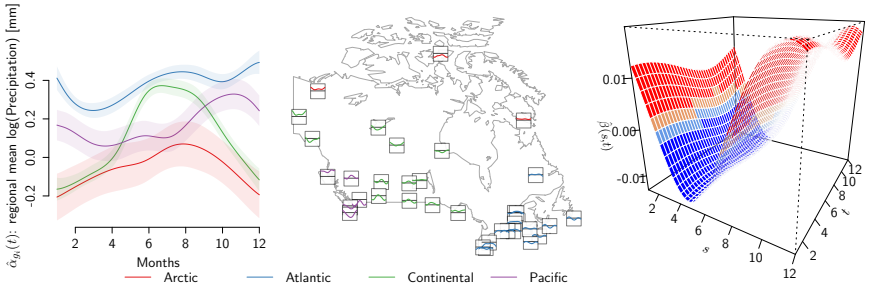


FIGURE 1. Regional mean log-precipitation (left), spatially correlated smooth residual curves (middle) and temperature effect on log-precipitation (right) for 35 Canadian weather stations.

Acknowledgments: A longer version of this paper was previously accepted for publication in the ASA publication Journal of Computational and Graphical Statistics, see Scheipl et al (2014). This work was funded by Emmy Noether grant GR 3793/1-1 from the German Research Foundation (FS, SG) and U.S. National Science Foundation grant number DMS 1007466 (AMS).

References

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

Ivanescu, A.E., Staicu, A.-M., Scheipl, F. and Greven, S. (2013). Penalized function-on-function regression. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, Working Paper 254.

McLean, M., Hooker, G., Staicu, A.-M., Scheipl, F. and Ruppert, D. (2014). Functional Generalized Additive Models. *Journal of Computational and Graphical Statistics*, 23(1), 249-269.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B* 68(2), 179-199.

Reiss, P. T. and Ogden, T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B* 71(2), 505-523.

Scheipl, F., Staicu, A.-M. and Greven, S. (2014). Functional Additive Mixed Models. *Journal of Computational and Graphical Statistics*, in press, DOI 10.1080/10618600.2014.901914.

Expectile smoothing for big data

Sabine K. Schnabel[1], Paul H.C. Eilers[1,2]

[1] Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands
[2] Erasmus Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: sabine.schnabel@wur.nl

Abstract: In the last years expectiles have become a popular alternative to quantiles. This is partly due to their fast and easy computation based on least squares. However, for large data sets the estimation is still computationally intensive. Here, we are introducing expectile smoothing for big data. First, we summarize the data as a two-dimensional histogram and proceed with the midpoints of the bins as pseudo-observations combined with a set of weights. We apply discrete smoothing to these data. Our approach is illustrated with an example from the field of plant genetics: the relation between the decay of linkage disequilibrium and the distance of SNPs (single nucleotide polymorphisms) on chromosomes of maize.

Keywords: Expectiles; SNPs; Big Data; Discrete Smoother

1 Introduction

Expectiles are an interesting and useful alternative to quantiles (Schnabel and Eilers, 2009). One of their attractions is fast and easy computation using iteratively asymmetrically weighted least squares. Yet for large datasets the computational effort might be heavy. In one application from (plant) genetics we visualize the decay of linkage disequilibrium (LD decay) against the distance of SNPs on chromosomes. The basic idea is to compute the correlation between pairs of SNPs. With thousands of SNPs we get millions of pairs and an equal amount of correlations and distances.
To get acceptable computation times we first summarize the data on a grid covering the two-dimensional domain of distance and correlation. With a 100 by 100 grid we get a maximum of 10^4 pseudo-observations, independent of the number of initial data pairs. In addition, we can use fast matrix calculations in R. Moreover, we do not need a spline basis, but for further simplification we can use a discrete smoother.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

2 Theory and application

Expectile smoothing produces curves that visualize the conditional distribution of data pairs (x, y). The curves are constructed with a generous set of B-splines and a roughness penalty is used to tune smoothness (Schnabel and Eilers, 2009). If B is the B-spline basis, the following objective function is minimized:

$$Q = (y - B\alpha)^{\mathrm{T}} W_p (y - B\alpha) + \lambda ||D\alpha||^2. \tag{1}$$

Here W_p is a diagonal matrix with

$$\begin{array}{ll} w_{ii} = p & \text{if } y_i > \hat{y}_i; \\ w_{ii} = 1 - p & \text{otherwise,} \end{array} \tag{2}$$

where $\hat{y} = B\hat{\alpha}$ are the fitted values. The parameter p $(0 < p < 1)$ is called the asymmetry. The matrix D forms (second order) differences and with λ one can tune smoothness. To estimate α, the system

$$(B^{\mathrm{T}} W_p B + \lambda D^{\mathrm{T}} D)\alpha = B^{\mathrm{T}} W_p y \tag{3}$$

is solved repeatedly, updating the elements of W_p after each iteration. This is the least asymmetrically weighted squares (LAWS) algorithm. It converges quickly to the minimum of Q, which is a convex function of α. For visualization, one computes and plots the curves for a set of values of p.
This all works well for data sets up to tens of thousands of observations. For larger data sets computation speed and memory use can become problematic. Assuming we have a million data points and use 20 B-splines. The basis matrix B has 20 million elements and the inner product $B'W_pB$ has to be computed about five times for each value of p. To minimize memory use and computation time, we propose a very simple solution: summarize the data on a grid as a two-dimensional histogram, and use the midpoints of the bins as pseudo-data, with prior weights equal to the counts in the bins.
We could apply "classic" expectile smoothing to these pseudo-data, but further streamlining is possible. Instead of B-splines we can apply discrete smoothing immediately. This is equivalent to a B-spline basis of degree zero (the identity matrix) and one element of α for each bin on the x-axis. Let the bins be indexed by j (for x) and k (for y) and let $U = [u_{jk}]$ contain the counts. The vector $t = [t_k]$ contains the midpoints of the bins along the y-axis. Let $\tilde{\alpha}_j$ be the current approximation to the point on the expectile curve in bin j. Weights are computed in $V = [v_{jk}]$, with

$$\begin{array}{ll} v_{jk} = p & \text{if } t_k > \tilde{\alpha}_j; \\ v_{jk} = 1 - p & \text{otherwise.} \end{array} \tag{4}$$

Let $s_j = \sum_k u_{jk} v_{jk}$ and $r_j = \sum_k u_{jk} v_{jk} t_k$. Then a new α is found by solving

$$(S + \lambda D^{\mathrm{T}} D)\alpha = r, \tag{5}$$

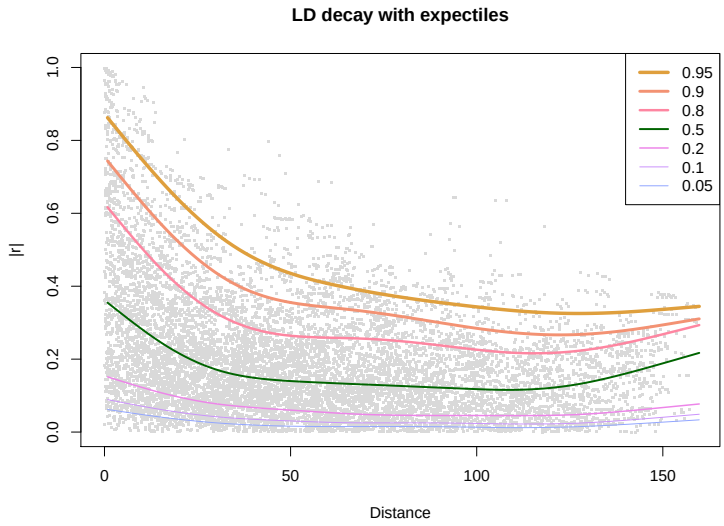


FIGURE 1. Decay of linkage disequilibrium (LD) with distances (Morgan) on chromosome 7 of maize. Expectile curves (with smoothing parameter $\lambda = 10^5$) are plotted for seven asymmetries p.

where $S = \text{diag}(s)$ and D again is a (second order) difference matrix.

The two-dimensional histogram can be constructed quickly in `R`. The simplest approach is to just run a loop over all observations. That takes 3 to 5 seconds for one million observations. Compilation to byte code —using `cmpfun` in the `compiler` package— reduces this time by a factor five. The fastest solution however is to apply the `hist` function. Let j_i and k_i be the row and column indices of observation i. Let $h_i = j_i + nk_i$ be a combined index. The one-dimensional histogram of h with n^2 bins can be redimensioned to the desired n-by-n matrix. This speeds up the calculations by another five times.

To illustrate our algorithm, we use the correlations between almost 12000 pairs of SNPs on chromosome 7 of maize. The data is available in the `R` package `synbreed` (Wimmer et al., 2012). This is not a very large data set as we want to avoid filling the plot with too many data points. Typically data in plant as well as in animal and human genetics contain much larger number of SNPs. Figure 1 shows the data and the estimated expectile curves. In both directions 100 bins were used. Even when using a loop for forming a histogram, the graphs appears without noticeable delay after the correlations have been computed.

The density of the dots in Figure 1 is just acceptable. With many more dots, the plane would be filled completely in many places. An alternative is the scatterplot smoother (Eilers and Goeman, 2004). It also computes a two-dimensional histogram, smoothes the counts and shows them as an image with a color map. As a by-product it delivers the raw histogram, which can be used directly for expectile smoothing. The resulting expectile

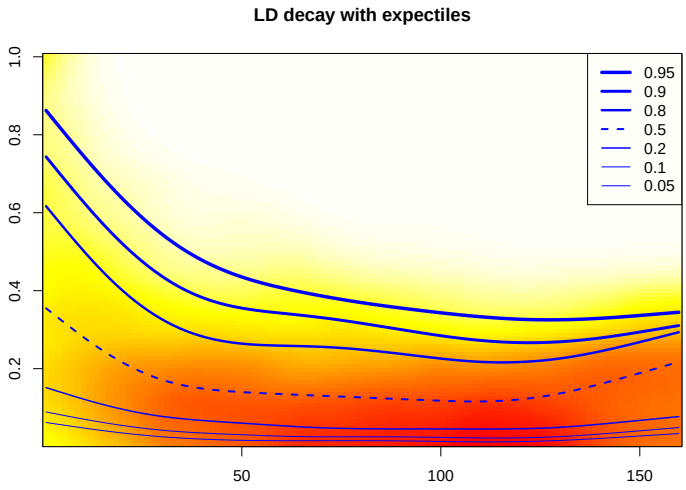


FIGURE 2. Decay of linkage disequilibrium (LD decay) with distances (Morgan) on chromosome 7 of maize. Expectile curves for seven asymmetries p (with smoothing parameter $\lambda = 10^5$) are plotted over the scatterplot smoother.

curves can be plotted over the images in a contrasting color (Figure 2).
It is possible to adapt the grid algorithm to the estimation of quantile curves, using the algorithm of Schlossmacher (1973). Our experiments show that this works, but that convergence is rather slow, in contrast to the estimation of expectile curves. Figure 3 shows results.
An alternative approach is to derive conditional densities from the expectiles, as explained by Schnabel and Eilers (2013). The densities can be used to compute quantiles. Exploration of this approach will be reported elsewhere.

3 Discussion

We have presented an algorithm to apply expectile smoothing to (very) large data sets. After construction of the two-dimensional histogram, which can be done very quickly, the expectile curves are obtained almost instantly. The construction of the the histogram itself takes less than a second for ten million observations.
For visualization purposes a grid of 100 bins on the x-axis is fine enough. The number of bins in the y-direction might need more attention in special cases. If the data cloud is not horizontal and narrow, it might be covered locally by only a small number of bins. Then one might need 200 or more bins for y. An alternative is to first remove the trend, compute expectile curves for the residuals and combine these with the trend.
Many aspects need further study. One is the automatic choice of λ, e.g. by cross-validation or variance-components estimation based on the Schall

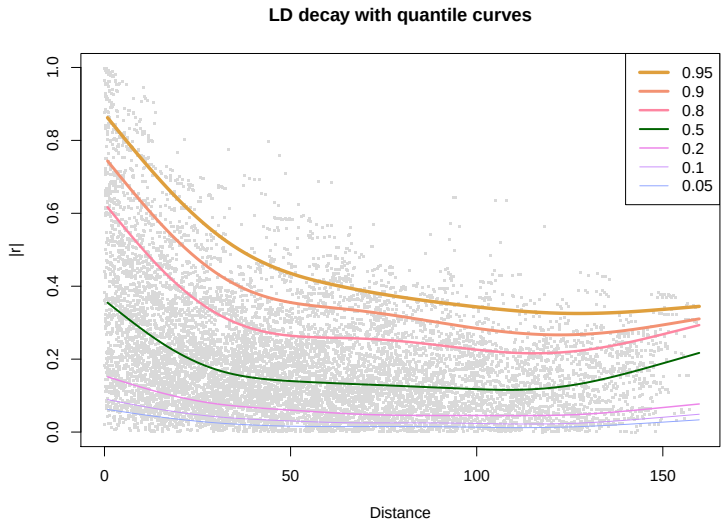


FIGURE 3. Decay of linkage disequilibrium (LD) with distances (Morgan) on chromosome 7 of maize. Seven τ-quantile curves (with smoothing parameter $\lambda = 10^6$) are displayed.

algorithm (Schnabel and Eilers, 2009). Also we will combine the grid approach with more advanced expectile smoothing techniques such as the location-scale model (Schnabel and Eilers, 2013) and expectile sheets (Schnabel and Eilers, 2014). On the other hand the need for such extensions, which were inspired by the problem of crossing expectile curves, is less when there are many observations.

It will also be useful to introduce shape constraints. It is generally assumed that LD decays monotonically with distance. In the graphs we see a tendency for the estimated curves to increase at the right end. The desired behaviour can be realized by adding an asymmetric penalty (Eilers, 2005).

References

Eilers, P.H.C (2005). Unimodal Smoothing. *Journal of Chemometrics*, **19**, 317 – 328.

Eilers, P.H.C. and Goeman, J.J. (2004). Enhancing scatterplots with smoothed density. *Bioinformatics*, **20**, 623 – 628.

Schlossmacher, E.J. (1973). An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association*, **68**, 857 – 859.

Schnabel, S.K. and Eilers, P.H.C. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, **53**, 4168 – 4177.

Schnabel, S.K. and Eilers, P.H.C. (2013). A location-scale model for non-crossing expectile curves. *Stat*, **2**, 171 – 183.

Schnabel, S.K. and Eilers, P.H.C. (2014). A note on expectile sheets. *In revision.*

Wimmer V., Albrecht T., Auinger H.J., and Schoen C.C. (2012). synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*, **28**, 2086 – 2087.

Semiparametric quantile regression using a mixed models representation

Fabian Sobotka[1], Thomas Kneib[1]

[1] Chair of Statistics, Department of Economics, University of Goettingen, GERMANY

E-mail for correspondence: fabian.sobotka@wiwi.uni-goettingen.de

Abstract: While a simple mean regression attempts to describe the expectation of a response as a function of the covariates, the results of a quantile regression offer a much broader view. In principle, a dense set of quantiles allows for an analysis of the complete conditional distribution of the response. This can lead to new insight into the dependency between the response and its covariates. In our work, we allow for additive regression models with nonlinear as well as spatial effects. Models of this flexibility are previously unavailable to frequentist quantile regression. We achieve the inclusion of B-splines and Markov random fields, for example, with their mixed model representation and a LASSO penalty. We also investigate possible improvements in coverage rates.

Keywords: quantile regression; semiparametric models; mixed models; LASSO

1 Introduction

Quantile regression (Koenker and Bassett, 1978) is on the verge of becoming a standard tool in modern regression analysis. A regression quantile is normally estimated using linear programming. Although there are some attempts to provide flexible estimates for nonlinear or spatial effects, quantile smoothing splines (Koenker et al., 1994) and triograms do not provide the same tempting smooth results as semiparametric models in mean or expectile regression. Further, the inclusion of a Markov random field is not possible until now. Different estimation procedures like Bayesian quantile regression or quantile boosting inherently support the most flexible semiparametric models. In this paper, we propose an efficient semiparametric quantile regression (SPQR) using the fast linear programming procedure and the asymptotic normality.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

2 Semiparametric Regression

The most flexible regression models include parametric effects as well as nonparametric effects $f(z)$ for metric covariates and for spatial covariates or interaction terms. Additionally, random effects can be included. The regression equation can then be written as

$$y = \beta_0 + \sum_{j=1}^{r} f_j(z) + \varepsilon. \tag{1}$$

In order to make all effects available for estimation, we need to parameterise all unknown functions as $f_j = \mathbf{Z}_j\beta_j$ with a design matrix $\mathbf{Z}$ and a vector of regression coefficients β. This simplifies the regression to

$$y = \beta_0\mathbf{1} + \mathbf{Z}_1\beta_1 + \ldots + \mathbf{Z}_r\beta_r + \varepsilon.$$

Overparameterisation is avoided by adding penalty terms $\mathbf{K}$ for the coefficient vector. These penalty terms are generally designed to work with a least squares estimate. We can, however, transform all design matrices $\mathbf{Z}$ such that their penalty has a diagonal form as described in Fahrmeir et al.(2004). All design matrices have the representation $\mathbf{Z} = (\mathbf{X}, \mathbf{B})$ with an unpenalised part $\mathbf{X}$ and penalised random effects in $\mathbf{B}$. For any basis matrix $\tilde{\mathbf{Z}}$ we receive this representation by calculating $\mathbf{Z} = \tilde{\mathbf{Z}}\mathbf{\Gamma}(\mathbf{\Gamma}\mathbf{\Gamma}')^{-1}$ if the penalty matrix is given by $\mathbf{K} = \mathbf{\Gamma}'\mathbf{\Gamma}$.

3 Quantile Regression

In quantile regression we assume that the τ-quantile of the error distribution equals zero, i.e.

$$P(\varepsilon_{i\tau} \leq 0) = \tau.$$

This implies that the predictor $\mathbf{Z}\beta_\tau$ corresponds to the τ-quantile of the response y.
Computationally, regression quantiles with a LASSO penalty for random effects are obtained by minimising an asymmetrically weighted absolute residuals criterion

$$\sum_{i=1}^{n} w_\tau(y_i, \mathbf{Z}_i\beta_\tau)|y_i - \mathbf{X}_i\alpha_\tau - \mathbf{B}_i\gamma_\tau| + \lambda|\gamma_\tau| \tag{2}$$

with asymmetric weights

$$w_\tau(y_i, \mathbf{Z}_i\beta_\tau) = \begin{cases} 1-\tau & y_i \leq \mathbf{Z}_i\beta_\tau \\ \tau & y_i > \mathbf{Z}_i\beta_\tau, \end{cases}$$

a response y and a quantile-specific predictor $\mathbf{Z}_i\beta_\tau$ consisting of the unpenalised effects $\mathbf{X}_i\alpha_\tau$ and the penalised random part $\mathbf{B}_i\gamma_\tau$ where the regression coefficients $\beta' = (\alpha, \gamma)'$. This loss function can be subject to a linear program as provided by Koenker(2005). Hence, we can now estimate regression quantiles for all semiparametric models.

4 Empirical Evaluation

In order to assess the possible gains of the mixed model approach besides the additional available modelling choices, we compare P-splines and quantile smoothing splines in artificial data scenarios.

4.1 Design

We replicate 1000 data sets drawing a uniform covariate $X \sim U(0,1)$ and heteroscedastic random errors from a normal $\varepsilon \sim N(0,(0.2+|x-0.5|)^2)$ or an exponential $\varepsilon \sim Exp(1/(0.2+|x-0.5|))$ distribution. The data is then of the form

$$y = \sin(2(4x-2)) + 2\exp(-16^2(x-0.5)^2) + \varepsilon.$$

For the estimation we choose a linear programming quantile regression with a P-spline basis (SPQR) or quantile smoothing splines (quantreg). Both times the smoothing parameter is optimised via a grid search for the lowest BIC. As a further reference we also employ quantile boosting from the package mboost (Hothorn et al., 2013). The results are compared in terms of a root mean squared error (RMSE) and the coverage rates along the covariate. The latter is not available for boosting.

4.2 Results

The error of the estimation in Figure 1 shows the efficient estimates that come with the use of P-splines in the boosting algorithm, especially in the outer 5% of the distribution. At the same quantiles, the quantile smoothing splines of quantreg fail to produce a reliable estimate. However, the combination of P-splines with the linear programming of quantreg produces the overall smallest errors for all quantiles.
Based on the asymptotic normality of the quantile regression estimate, we can compute pointwise confidence intervals for the estimates. Our simulations also show that the use of P-splines and a LASSO penalty provides a better coverage of the true function than quantile smoothing splines, as shown in Figure 2. This is also true for the median where the RMSE of the two approaches was very similar.
We continue our simulations for geoadditive models and try to add a comparison with Bayesian quantile regression.

References

Bühlmann, P., Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science* 22(4), 477-505.

Fahrmeir, L., Kneib, T., Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica* 14, 731761.

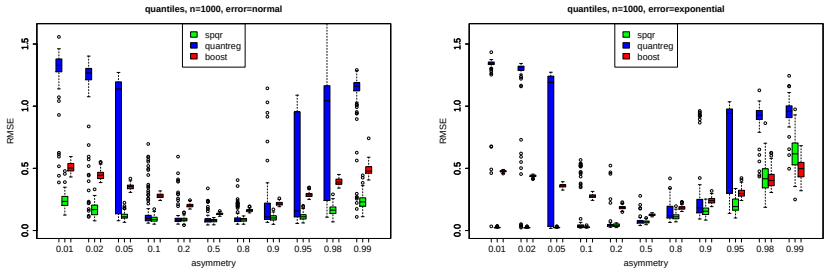


FIGURE 1. Root mean squared errors for a selected set of 11 quantiles comparing standard quantreg methods (dark), quantile boosting (med) and the proposed SPQR (light) for normal and exponential errors.

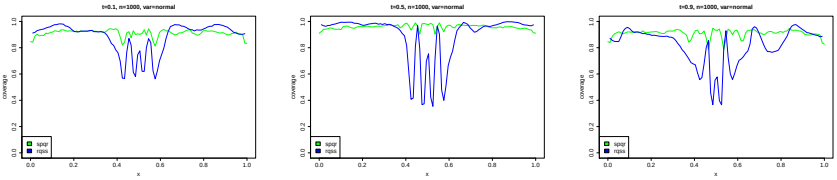


FIGURE 2. Coverage rates for quantile smoothing splines by quantreg (dark) and the proposed SPQR (light) for pointwise 95% confidence intervals along the covariate. Exemplary for normal errors and quantiles of 10%, 50% and 90%.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., Hofner, B. (2013). mboost: Model-Based Boosting. R package version 2.2-1.

Koenker, R., Bassett, G. (1978). Regression Quantiles. *Econometrica* 46(1), 33-50.

Koenker, R., Ng, P., Portnoy, S. (1994). Quantile Smoothing Splines. *Biometrika* 81(4), 673-680

Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press.

Improved quantitative analysis of tissue characteristics in PET studies with limited uptake information.

James Sweeney[1], Finbarr O'Sullivan[2], David Hawe[2]

[1] University College Dublin,Ireland
[2] University College Cork,Ireland

E-mail for correspondence: james.sweeney@ucd.ie

Abstract: Quantitative analysis of FDG uptake is important in oncologic PET studies in order to determine whether a tissue is malignant or benign, and in attempting to predict the aggressiveness of an individual tumour. Following injection of the FDG tracer, a series of scans are taken of a biological region of interest; this provides dynamic information from which we draw inference on the metabolic parameters describing the evolution of tracer activity and hence underlying tissue characteristics. However, in certain cases only a single static scan of a region of interest may be available. A prime example is a "late uptake scan" where a body region is analysed for only a very brief period, providing an extremely sparse and potentially noisy data set from which it is difficult to draw conclusions on underlying tissue characteristics. This article focuses on the impact and benefit of incorporating prior tissue information, via penalty structures in our data models, to improve prediction outcomes in the presence of limited uptake information. Specifically, we display that our proposal appears to offer an extremely promising alternative to existing, competing methodologies.

Keywords: Positron Emission Tomography,Dynamic reconstruction; Nonlinear Regression; Estimation of Flux.

1 Introduction

In the field of oncologic research, CT or MRI scans are often used in the detection of anamolous growths or lumps in the body regions of individual patients. A drawback of these imaging techniques however, is that they provide very limited information on living tissue characteristics; specifically, we cannot identify if the anamolous tissue is malignant, requiring immediate invasive surgical action, or benign, in which case surgery is not required.

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

In dynamic mode, Positron Emission Tomography (PET) imaging provides a solution to this problem. This is a powerful diagnostic imaging technique that provides a unique opportunity to probe the status of healthy and pathological tissue by examining how it processes substrates. In the case of FDG PET imaging, a radioactively laced glucose substrate is ingested by the patient. The patient is then placed in an imaging machine where a time sequence of PET images is obtained, allowing us to monitor the interaction of the tracer molecules with the bodys physiological processes. For example, malignant cancerous growths will have a high uptake of the substrate; such tissue regions will "hoard" the radio-laced tracer manifesting as a "hotspot" on a PET image. The nature of these hotspots pose many additional quantitative questions, chief amongst them being the rate of influx of a tracer to the tissue, which is an indicator of vigour for the cancerous growth.
The quality of inferences on tissue characteristics obtained from a PET scan is proportional to the length of the imaging time; the scanning period is of the order of 90 minutes. However, in many cases only a "late uptake scan" is available, where a given tissue region is studied for a 15 minute period. The output of such an experiment is an extremely sparse and potentially noisy data set from which it is difficult to draw conclusions on underlying tissue characteristics. In this article we propose the incorporation of penalty structure in our data model, and illustrate how this can dramatically improve the quality of inference on metabolic parameters.

2 Data

The available data consists of a set of 34 FDG PET imaging scans, each of duration 90 minutes, concerning patients identified with brain tumour lesions. Each 90 minute scan is recorded as a set of 31 images, which consist of information on the tracer uptake of brain tissue for each of the brain regions, namely grey matter, white matter and cancerous tissue.

3 Methods

The fundamental equation of dynamic PET radiotracer imaging is:

$$C_T(t) \quad = \quad v_B C_p(t) + \int_0^t R(t-s) C_p(s) ds \tag{1}$$

where $C_T(t)$ represents the accumulated concentration of substrate in a tissue region at time t, v_B the volume of blood in the region and $C_p(t)$ the concentration of substrate within arterial blood at time t. $R(t)$ represents a *Residue function* which models how the subrate is processed metabolically by the tissue. The 4 parameters governing $R(t)$ (henceforth denoted θ) provide information on the metabolic activity of the tissue region and are

used to estimate the rate of influx of tracer subrate, which is the primary question of interest. Hawe et al. (2012) provide a more formal introduction. In terms of statistical inference on θ, we assume that observed concentration of tracer substrate in a tissue region in image k, z_k, is approximately Gaussian distributed with variance proportional to the tracer concentration at time k.

$$z_k \quad = \quad C_T(t_k|\theta) + w_k(\theta)\epsilon_k \tag{2}$$

We seek to minimise the weighed residual sum of squares:

$$WRSS(\theta) \quad = \quad \sum_{k=1}^{N} w_k(\theta) \left[z_k - C_T(t_k|\theta)\right]^2 \tag{3}$$

In this article we will consider the case where only 15 minutes of noisy scan information is available, consisting of 3 observations(as opposed to 31) for each tissue region. Thus, there are only 3 observations available with which to estimate the 4 parameters comprising θ of the residue model. The solution we pose to this ill-posed problem, is to embed the inference procedure in a Bayesian framework and "borrow" information from "known" tumour tissue residues. The posterior distribution for the unknown model parameters, conditional on the data (Z) and known tissue information, is:

$$\pi(\theta, \sigma^2|Z, R^r, C_p) \quad \propto \quad \pi(Z|\theta, C_p, \sigma^2)\pi(\sigma^2)\pi(\theta|C_p, R^r) \tag{4}$$

Here σ^2 is a parameter governing fidelity to the data, as opposed to weighing on the prior information, and must also be inferred from the data. R_r represent reference residue structures obtained by harnessing the best estimates of the θ parameters from sample studies, and are assumed known (subject to quantifiable statistical error). In order to make inference on the unknown parameters, distributions must be specified for both the data $\pi(z_k|...)$, as well as priors on both σ^2 and the residue parameters $\pi(\theta|....)$.

$$\pi(z_k|\theta, C_p, \sigma^2) \quad \sim \quad N(v_B c_p(t) + \int_0^t R(t-s)C_p(s)ds, \sigma^2) \tag{5}$$

We assume that the data distribution is Gaussian, with constant variance σ^2. Whilst the variability in the data is heterogenous, in practice the assumption of contant variance seems appropriate given the relative proximity of the sample points in time.

$$\pi(\theta|C_p, R^r) \quad \sim \quad SkewNormal(f(\overline{R}(\theta))) \tag{6}$$

The prior on θ is data driven; it is specified on the residue structure implied by the parameters of θ as opposed to the individual parameters themselves.

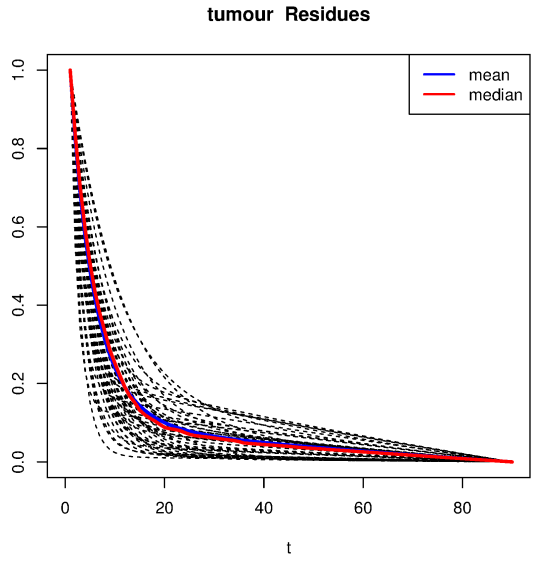


FIGURE 1. Plot of standardised tumour residues, constrained to 1 at time $t = 0$ and 0 at $t = 90$.

This prior is derived from the available data; the residue structures are standardised ($\overline{R}$) so as not to impact on the estimated flux values - this is achieved by constraining the residues to take the value 1 at time $t = 90$, and 0 at time $t = 0$. As illustrated in Figure 1, this constraint induces heterogenity in the variability associated with the residue estimates at each timepoint, and thus a skew Normal distribution is assigned with the median as it's centre of location. Finally, σ^2 is assigned a flat noninformative uniform prior.

4 Results

In Figure 2 & Figure 3 we illustrate the potency of the procedure presented in Equations (4)-(6). Flux values for the cancerous brain tissues of all 34 patients are obtained using the full imaging datasets, with the θ parameters estimated via maximum likelihood. A subset of the data is then considered, namely the imaging scans from the period 45-60 mins after the injestion of tracer substrate. For our novel framework inference procedures are Empirical Bayes' based; σ^2 is optimised across all 34 cases, with optimised estimates of θ specific to each study also obtained . In Figure 2 we present the inferred residue curves for a subset of cases; these residue structures are obtained from just 3 sample data points and appear to match the "unobserved" values quite well. In Figure 3, the real power of the approach is illustrated; we compare the flux estimates obtained by our novel approach to those supplied by a competing methodology ("Hunters method" - see

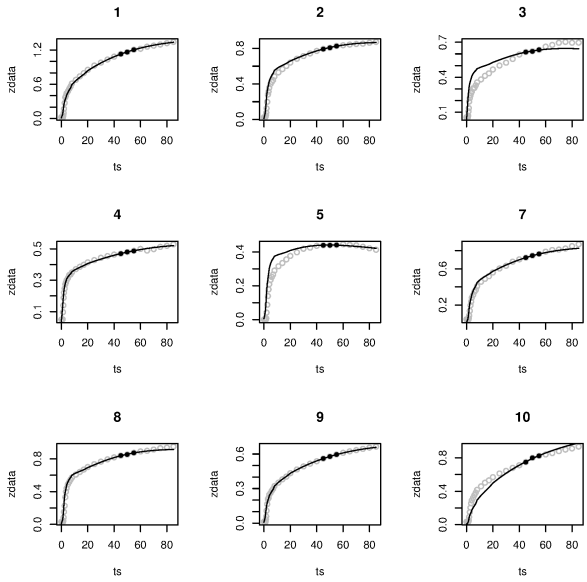


FIGURE 2. Estimated residue curves (R), inferred from the 3 sample points (in bold) available in each study. Omitted datapoint unused for model fitting are displayed in grey.

Hunter et al. (1996)). The results, displaying less bias and variability than the competing method, indicate that accurate inferences can be made on metabolic tissue parameters, even in the presence of limited uptake information, via the incorporation of information from known prior studies into the model framework.

Acknowledgments: Support of Science Foundation Ireland [11/PI/1027] is gratefully acknowledged.

References

Hawe, D., Huang, J., Wolsztynski, E. and O'Sullivan, F. (2012). Kinetic Analysis of Dynamic Positron Emission Tomography Data using Open Source Image Processing and Statistical Inference Tool. *WIREs Comp Stat*, **4**, 316 – 322.

Hunter, G.J., Hamberg, L.M., Alpert, N.M., Choi, N.C. and Fischman, A. (1996). Simplified Measurement of Deoxyglucose Utilization Rate. *Journal of Nuclear Medicine*, **37**, 950 – 955.

Patlak, C., Blasberg, R. and Fenstermacher, J. (1983). Graphical Evaluation of Blood to Brain Transfer Constants from Multiple Time Uptake Data *Journal of Cerebral Blood Flow and Metabolism*, **3**, 1 – 7.

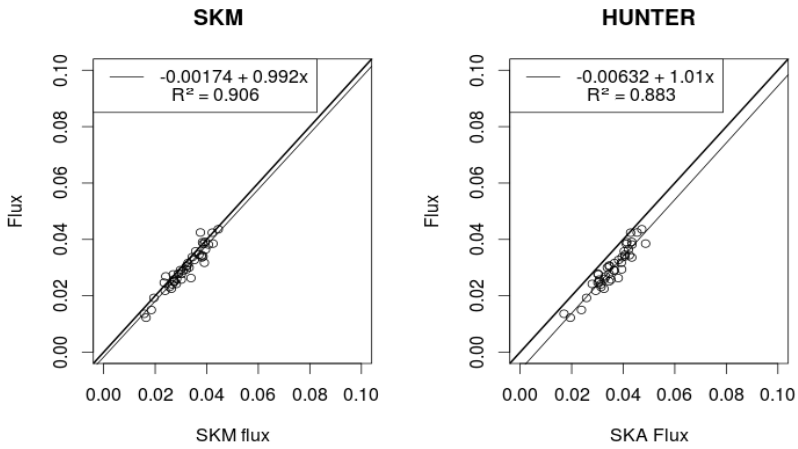


FIGURE 3. Plot of estimated values of flux for approximate methods.

Spence, A.M, Muzi, M.,Mankoff, D., and O'Sullivan, F. (2004). 18F-FDG PET of Gliomas at Delayed Intervals:Improved Distinction Between Tumor and Normal Gray Matter. *The Journal Of Nuclear Medicine*,**45**, 1653 – 1659.

Sundaram,S.K.,Freedman, N.,Carrasquillo, J. and Carson,J. (2004). Simplified Kinetic Analysis of Tumor 18F-FDG Uptake: A Dynamic Approach. *The Journal Of Nuclear Medicine*,**45**, 1328 – 1333.

Time-dependency in multi-state models: specification and estimation

Ardo van den Hout[1]

[1] Department of Statistical Science, University College London, UK

E-mail for correspondence: ardo.vandenhout@ucl.ac.uk

Abstract: Continuous-time multi-state survival models can be used to describe health-related processes over time. In the presence of interval-censored times for transitions between the living states, the likelihood is constructed using transition probabilities. Model specification and maximum likelihood estimation using the scoring algorithm are extended to allow for flexible parametric modelling of the time-dependency of the process. Within one multi-state model, transition-specific hazards can be defined using Gompertz and Weibull formulations. Data from the English Longitudinal Study of Ageing are analysed to illustrate the methods.

Keywords: Cognitive function; Gompertz distribution; Markov model.

1 Introduction

The application of interest is change of cognitive function over time in the older population. Data stem from the English Longitudinal Study of Ageing (ELSA). The longitudinal response variable is the number of words remembered in a recall from a list of ten. Of interest is the effect of age and gender on cognitive change over time when controlling for education. Four states are defined by the number of words an individual can remember, see Figure 1. The dead state is the fifth state. The transition times between the living states are interval-censored, but death times are known.
Using age as the time scale, a continuous-time five-state survival model will be specified to analyse the data. Parametric time-dependency of the transition process is approximated piecewise-constantly, and an extension of the scoring algorithm in Kalbfleisch and Lawless (1985) is used to estimate the model parameters.
Within one multi-state model, transition-specific hazards can be defined using Gompertz and Weibull formulations. This flexible parametric modelling and the corresponding scoring algorithm extend current methods for multi-state analysis of interval-censored data (*cf.* Jackson 2011).

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

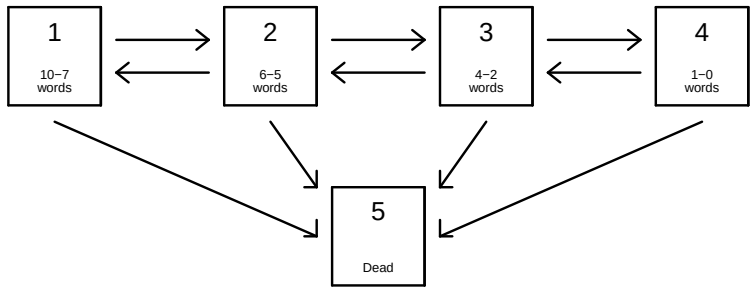


FIGURE 1. Five-state model for longitudinal data in ELSA on number of words remembered in a recall.

2 The Model

For a continuous-time Markov chain $Y(t)$ on state space S, time-homogeneous *transition probabilities* are given by

$$p_{rs}(t) = P\big(Y(t) = s | Y(0) = r\big),$$

for $r, s \in S$ and $t \geq 0$. Matrix $\mathbf{P}(t)$ contains these probabilities such that the rows sum up to 1. The Chapman-Kolmogorov equation is $\mathbf{P}(u+t) = \mathbf{P}(u)\mathbf{P}(t)$. The *transition intensities* (or *hazards*) are given by

$$q_{rs} = \lim_{\delta \downarrow 0} \frac{P\big(Y(t+\delta) = s | Y(t) = r\big)}{\delta},$$

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* $\mathbf{Q}$. Given $\mathbf{Q}$, the solution for $\mathbf{P}(t)$ subject to $\mathbf{P}(0) = \mathbf{I}$ is $\mathbf{P}(t) = \exp(t\mathbf{Q})$.
A hazard regression model for transition intensities combines baseline hazards with log-linear regression and is given by

$$q_{rs}(t) = q_{rs.0}(t) \exp\left(\boldsymbol{\beta}_{rs}^{\top} \mathbf{x}\right),$$

where $\mathbf{x}$ is the covariate vector without an intercept. Transition-specific time dependency can be introduced via the baseline hazards. Parametric examples are

$$\begin{array}{lllll} \text{Weibull:} & q_{rs.0}(t) & = \lambda_{rs}\tau_{rs}t^{\tau_{rs}-1} & \lambda_{rs}, \tau_{rs} > 0 \\ \text{Gompertz:} & q_{rs.0}(t) & = \lambda_{rs}\exp(\xi_{rs}t) & \lambda_{rs} > 0\ . \end{array}$$

Consider the time interval $(t_1, t_2]$ with observed states at t_1 and t_2. Working with constant hazards, $\mathbf{Q}$ is defined for time t_1 using the regression model, and the transition matrix $\mathbf{P}(t_2 - t_1)$ is subsequently defined for elapsed time $t_2 - t_1$ using $\mathbf{Q}$. Notation: $\mathbf{Q}(t_1)$ and $\mathbf{P}(t_1, t_2)$.
In longitudinal data, trajectories consist of repeated observations of the state. A piecewise-constant approximation of the time dependency can be

defined as follows. If states are observed at t_1, t_2, and t_3, the transition matrix for $(t_1, t_3]$ is given by $\mathbf{P}(t_1, t_3) = \mathbf{P}(t_1, t_2)\mathbf{P}(t_2, t_3)$, where both matrices at the right-hand side are derived using constant hazards. In this example, the grid for the piecewise-constant hazards is defined by the data. Using a similar approach, it is also possible to define an approximation using an imposed grid which is independent from the data.

3 Maximum likelihood

In the presence of interval-censoring, the likelihood is constructed using transition probabilities. Let the states by denoted by $1, 2, .., D$, with D the dead state. Consider an individual with observation times $t_1, ..., t_n$, where the state at t_n is allowed to be right-censored. The observed trajectory is the series of states $y_1, ..., y_n$. The likelihood contributions for intervals $(t_{j-1}, t_j]$ are given by

$$L_{ij} = \begin{cases} P\left(Y_j = y_j | Y_{j-1} = y_{j-1}, \boldsymbol{\theta}\right) & \text{for} \quad j = 2, ..., n-1 \\ C(y_n | y_{n-1}) & \text{for} \quad j = n \ , \end{cases} \tag{1}$$

where $\boldsymbol{\theta}$ is the vector with all the model parameters. If a living state at t_n is observed, then $C(y_n|y_{n-1}) = P\left(Y_n = y_n | Y_{n-1} = y_{n-1}, \boldsymbol{\theta}\right)$. If the state is right censored at t_n, then $C(y_n|y_{n-1}) = \sum_{s=1}^{D-1} P\left(Y_n = s | Y_{n-1} = y_{n-1}, \boldsymbol{\theta}\right)$. In the case of known age at death,

$$C(y_n|y_{n-1}) = \sum_{s=1}^{D-1} P\left(Y_n = s | Y_{n-1} = y_{n-1}, \boldsymbol{\theta}\right) q_{sD}(t_{n-1}|\boldsymbol{\theta}).$$

Given N individuals, the likelihood is given by $L = \prod_{i=1}^{N} \prod_{j=2}^{n_i} L_{ij}$, where n_i is the number of observation times for individual i.
Maximising the likelihood can be undertaken by using a scoring algorithm. The central part of the algorithm is the first derivative of the log-likelihood, i.e., the score function. The specification (1) and the theory for the underlying Markov chain shows that the crucial step is to derive $\partial \mathbf{P}(t_{ij}, t_{ij+1})/\partial\theta_k$. The important aspects are as follows. *(i)* Because of the piecewise-constant approximation, basic formulas for the time-homogeneous case (Kalbfleisch and Lawless 1985) apply for the constituent intervals with constant hazards. *(ii)* By using an eigenvalue decomposition, only the derivatives $\partial \mathbf{Q}(t_{ij})/\partial\theta_k$ are needed. *(iii)* Also for the model with the parametric time-dependency, these derivatives are straightforward to derive.
The algorithm is implemented in R in such a way that it is easy to vary transition-specific choices for parametric shapes. An example of such a model is explored in the application, where Gompertz hazards are defined for moving between living states, and Weibull models for death.

4 Application

For the current analysis, a random sample is used of 1000 individuals in ELSA. Measures have been taken by the data provider to prevent identi-

TABLE 1. State table for the ELSA data: number of times each pair of states was observed at successive observation times. The four living states are defined by number of words remembered

	To				
From	10-7 words	6-5 words	4-2 words	1-0 words	Dead
10-7 words	121	104	31	5	3
6-5 words	117	348	196	30	24
4-2 words	36	204	440	82	32
1-0 words	3	26	73	82	29

fication of the individuals. One of those measures is the censoring of age above 90 years. We remove the six individuals in our sample who are older than 90 during the follow-up. The resulting sample contains 157 individuals with only one record. Because these individuals do not provide follow-up information, their records are also ignored in the analysis. The resulting sample size is 837.

The interval-censored observations are summarised by the frequencies in Table 1. The sum of the transitions into the dead state is equal to the number of deaths in the sample, i.e., 88. The diagonal of the 4×4 subtable for the living states dominates. This shows that if there is change over time, then this change is slow relative to the follow-up times in the ELSA study. Table 1 also shows that the process is mainly progressive in the sense that the main trend over time is towards the states with fewer words.

Model estimation is undertaken by using the scoring algorithm. Model selection is bottom-up starting with the time-homogeneous exponential hazard model given by $q_{rs}(t) = \exp\big(\beta_{rs.0}\big)$, for $(r,s) \in \{(1,2),(1,5),(2,1),(2,3),(2,5),(3,2),(3,4),(3,5),(4,3),(4,5)\}$. This intercept-only model with 10 parameters has AIC = 5016. Convergence of the scoring algorithm was reached after 12 iterations, using starting values $\beta_{rs.0} = -3$ for all the parameters.

The age scale is transformed by subtracting 31 years, which results in 1 being the minimal age in the sample. There is limited information on backward transitions. Mortality information is also limited because only 10.5% of the individuals end up in the dead state during follow-up. For this reason, the model is extended by adding parameters for progressive transitions only, and by imposing parameter constraints.

Using the piecewise-constant approximation, a Gompertz model is fitted with restrictions on the parameters for the effect of age. The grid for the piecewise-constant approximation is defined by individually observed follow-up times. The model is given by

$$q_{rs}(t) = \exp\big(\beta_{rs.0} + \xi_{rs} t\big), \tag{2}$$

where $\xi_{21} = \xi_{32} = \xi_{43} = 0$ and $\xi_{15} = \xi_{25} = \xi_{35} = \xi_{45}$. This model has 14 parameters, and needs 16 scoring iterations when using starting values $\beta_{rs.0} = -3$ and $\xi_{rs.0} = 0$ for all the relevant (r,s)-combinations. The model

has AIC = 4909. Covariate information is added for those transitions that represent a decline in cognitive function. For this, model (2) is extended by

$$q_{rs}(t) = \exp\left(\beta_{rs.0} + \xi_{rs}t + \beta_{rs.1}sex + \beta_{rs.2}education\right), \tag{3}$$

where *sex* is 0/1 for women/men, and *education* represents highest educational qualification with 1 for NVQ2/GCE O-Level equivalent or higher, and 0 otherwise. In the UK, O-Levels are qualifications of formal education up to the age of 16 years. For the transitions into the dead state, the constraints on the coefficients for *sex* are $\beta_{15.1} = \beta_{25.1} = \beta_{35.1} = \beta_{45.1}$, and for *education* we set $\beta_{r5.2} = 0$ for $r = 1, 2, 3, 4$. This model has AIC = 4831. Variants of model (3) were investigated which include both Gompertz and Weibull hazard formulations. To summarise, according to the AIC the best model (AIC 4830) is the one with Gompertz hazards for the forward transitions between the living states and Weibull hazards for death. This model is given by

$$\begin{aligned}
q_{rs}(t) &= \exp\left(\beta_{rs.0} + \xi_{rs}t + \beta_{rs.1}sex + \beta_{rs.2}education\right) \\
&\qquad \text{for } (r,s) \in \{(1,2),(2,3),(3,4)\} \\
q_{rs}(t) &= \exp\left(\beta_{rs.0}\right) \\
&\qquad \text{for } (r,s) \in \{(2,1),(3,2),(4,3)\} \\
q_{rs}(t) &= \tau_D t^{(\tau_D - 1)} \exp\left(\beta_{rs.0} + \beta_{D.1}sex\right), \\
&\qquad \text{for } (r,s) \in \{(1,5),(2,5),(3,5),(4,5)\},
\end{aligned} \tag{4}$$

Most of the point estimates are according to expectation. For example, the effect of getting older is associated with decline of cognitive function $\widehat{\xi}_{12}, \widehat{\xi}_{23}, \widehat{\xi}_{34} > 0$. For transitions $1 \to 2$ and $2 \to 3$ more education is associated with a lower risk of moving.

Long-term transition probabilities

Covariance of a function of model parameters can be estimated by using the multivariate delta method, or by using simulation. An important example of such a function is the matrix with the transition probabilities for a specified time interval. Let $\widehat{\mathbf{V}}_{\boldsymbol{\theta}}$ denote the estimated covariance matrix of the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$.
Of interest is the estimation of $\mathbf{P}(t_1, t_2)$ for arbitrary t_1 and $t_2 > t_1$. Let the grid for the piecewise-constant approximation be defined by $u_{j+1} = u_j + h$ for $j = 1, ..., J$ such that $u_1 = t_1$ and $u_J = t_2$. Using the multivariate delta method, the variance of estimated $\mathbf{P}(t_1, t_2)$ is given by

$$\left(\left.\frac{\partial \mathbf{P}(t_1,t_2)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right)^{\top} \widehat{\mathbf{V}}_{\boldsymbol{\theta}} \left(\left.\frac{\partial \mathbf{P}(t_1,t_2)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right).$$

where $\mathbf{P}(t_1, t_2) = \mathbf{P}(u_1, u_2) \times \cdots \times \mathbf{P}(u_{J-1}, u_J)$. The chain rule can be used to derive $\partial \mathbf{P}(t_1, t_2)/\partial \boldsymbol{\theta}$ using the derivatives for the constituent parts.

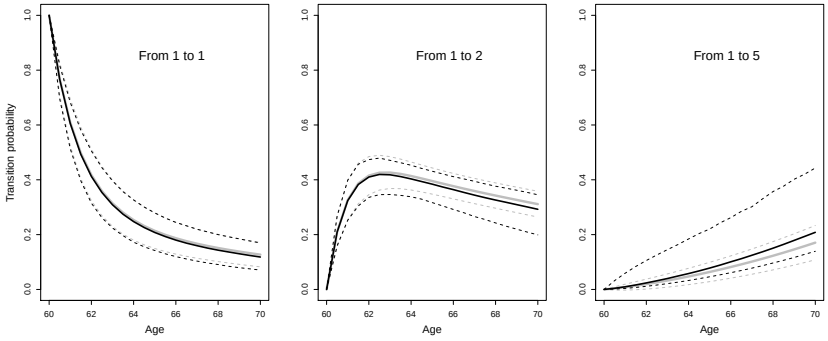


FIGURE 2. Estimated ten-year transition probabilities for men aged 60 with higher level of education, and in state 1 at baseline. Grey lines for delta method, black for simulation method. (Dashed lines for 95% confidence band.)

An alternative to the delta method is to use simulation. In that case, a parameter vector $\boldsymbol{\theta}^{(b)}$ is draw from $N(\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{V}}_{\boldsymbol{\theta}})$, for $b = 1, ..., B$, and for each sampled $\boldsymbol{\theta}^{(b)}$, $\mathbf{P}(t_1, t_2)$ is calculated. Summary statistics such as mean and covariance can be derived easily from the B realisations of $\mathbf{P}(t_1, t_2)$.
Ten-year transition probabilities are estimated for men in state 1, aged 60 with the higher level of education. The grid is defined by $h = 1/2$ years. The estimation is shown in Figure 2 for both the delta method and the simulation method ($B = 1000$). For long-term prediction, the difference between the methods is most striking when 95% confidence bands are compared for the probability of dying at a certain age.
Because transition probabilities are restricted to $[0, 1]$, using simulation is recommended. The delta method does not take the restriction into account and this has a substantial knock-on effect for long-term prediction.
Figure 2 concurs with the expectations. For example, given the progressive trend of the process, it is to be expected that probability of being in state 2 increases in the first years, but then decreases in the later years as being in states 3, 4, and 5 becomes more likely due to increased age.

References

Jackson C.H. (2011). Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, 38, 8.

Kalbfleisch, J. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption, *Journal of the American Statistical Association*, 80, 863–871.

Modeling Cherenkov Telescope images for Variable Construction in Classification

Tobias Voigt[1], Roland Fried[1]

[1] TU Dortmund University, Germany

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

Abstract: We model telescope images by fitting bivariate distributions. The newly constructed variables improve on the currently used Hillas parameters, which are based on elliptic fits, in terms of classification into signal and noise events.

Keywords: Classification; Cherenkov; Astronomy; Image Modelling

1 Introduction

In very high energy gamma astronomy, Cherenkov telescopes record images of so-called air showers, induced by highly energetic cosmic particles. A classification has to be done to separate the signal observations from the dominating background consisting of many irrelevant particles. It is known that the signal to background ratio in this problem is about 1:1000 (Weekes, 2003), so that a very high specificity of the classification is crucial to get rather clean signal samples. For the classification, variables constructed from the recorded images are used. These variables are based on moment analysis parameters and are called Hillas parameters (Hillas, 1985).
Our interest is whether the classification through the currently used Hillas variables can be improved by the construction of new variables. To answer this question, we connect the Hillas variables to statistical modelling. We observe that not all information known about signal observations is used by Hillas variables and extend them by new variables, based on fitting bivariate distributions to the recorded images and using distance measures for densities.

2 Hillas variables

The idea of Hillas variables is to fit an ellipse to a shower image and use the parameters of the ellipse as variables for classification. The first two Hillas

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

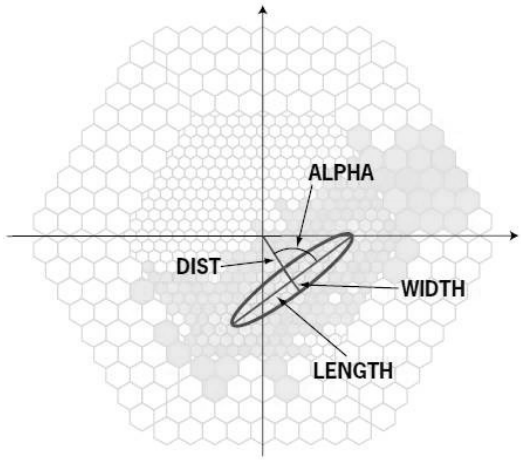


FIGURE 1. Some Hillas variables from fitting an ellipse.
(source: http://ihp-lx.ethz.ch/Stamet/magic/parameters.html)

variables are the center coordinates of the ellipse, also known as the center of gravity, `CoGX` and `CoGY`. Then there are the `Width` and `Length` of the fitted ellipse, that is the lengths of the two semiaxes, where the shorter one is always considered as `Width`. Another Hillas parameter directly calculated from the fitted ellipse is $\texttt{Area} = \pi \cdot \texttt{Width} \cdot \texttt{Length}$, the area of the fitted ellipse. The remaining two ellipse parameters are angles. The first is δ, the angle between the longer semiaxis of the ellipse and the x-axis of the camera. The second is α, which describes the angle between the longer semiaxis of the ellipse and the line between the center of the camera and the center of the ellipse. Besides, there are seven further Hillas variables mainly based on pixel brightnesses and ratios between them. Some Hillas variables and the underlying MAGIC telescope camera (MAGIC collaboration, 2014) can be seen in Figure 1.

3 Gaussian Fit

Fitting an ellipse as done for the Hillas variables corresponds to fitting a contour line of a bivariate elliptic distribution. The most popular example of such a distribution is a bivariate Gaussian, which can be fitted by minimizing a χ^2-distance between the observed and the expected frequencies in the hexagonal telescope image. This allows to include information not only on the boundaries of the ellipse, but also on the values in the interior. The relation between the parameters of the ellipse and the fitted Gaussian is as follows: Let $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^{2x2}$ be the parameters of the fitted bivariate Gaussian distribution. The fitted distribution is thus $\mathcal{N}_2(\mu, \Sigma)$. We look at the spectral decomposition of Σ:

$$\Sigma = U \Delta U^T$$

where

- $\Delta = \text{diag}\,(\lambda_1, \lambda_2)$ is a diagonal matrix with the sorted eigenvalues of Σ: $\lambda_1 > \lambda_2$.
- $U = (u_1, u_2)$ is an orthogonal matrix containing the corresponding eigenvectors $u_1 \in \mathbb{R}^2$ and $u_2 \in \mathbb{R}^2$.
- U^T denotes the transposition of the matrix U.

With this notation, we can describe the relationship between the fitted Gaussian and the other Hillas variables: $\mu = \begin{pmatrix} \texttt{CoGX} \\ \texttt{CoGY} \end{pmatrix}$, $\lambda_1 = c \cdot \texttt{Length}$, $\lambda_2 = c \cdot \texttt{Width}$, where c is some constant, $\texttt{Area} = \pi\lambda_1\lambda_2$, δ is the angle between the x-axis and u_1 and α is the angle between μ and u_1.
Note that the Hillas variables contain the full information about the parameters of the fitted Gaussian distribution. However, it is known that the signal has a more regular shape than the background. For example, signal is known to be unimodal, while background can have several peaks. This information cannot be evaluated by fitting a single ellipse, but can be used by fitting a distribution, for example through the use of goodness of fit measures. Fitting a Gaussian instead of an ellipse allows to incorporate information on the shape of signal images. The left hand side of Figure 2 shows a sample image on the underlying FACT telescope camera (The FACT collaboration, 2014) and a fitted bivariate Gaussian. It can be expected that a bivariate Gaussian fits better to signal images than to background images, because of the more regular shape of signal images. We thus use distance measures between the observed image and the fitted Gaussian distribution in our classification, as these tend to take smaller values for signal events. As distance measures we use the χ^2-distance (e.g. Greenwood & Nikulin, 1996)

$$Q_n = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i},$$

the Kullback-Leibler divergence (Kullback & Leibler, 1951)

$$D_k(P, Q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

and the Hellinger distance (Nikulin, 2001)

$$D_h(P, Q) = 1 - \sum_x \sqrt{p(x)q(x)}.$$

4 Including Information on Skewness or Alignment

Fitting Gaussian distributions for constructing variables for classification can possibly be further improved. Signal images seem to be roughly elliptical, but not exactly. It is known that signal showers are skewed in one

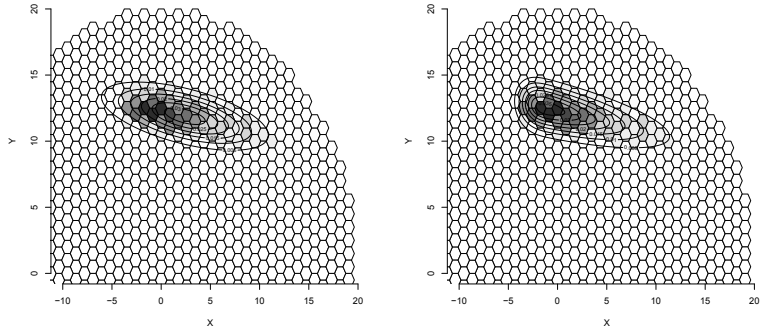


FIGURE 2. Contour lines of a normal (left) and skew-normal (right) distribution fitted to a random shower image. Only a part of the camera is shown.

direction (de Naurois & Rolland, 2009). Because of this, an elliptical fit or fitting a Gaussian distribution can neither describe signal observations nor background observations very well. As the aim of our analysis is to find differences between signal and background, it makes sense to fit a distribution to the images which describes signal images well. Fitting a skewed distribution to the images, instead of a Gaussian, seems therefore to be mandatory.

In a second step, we extend the idea of fitting a Gaussian distribution by fitting a distribution which fits better to the information we have about signal images, especially that signal images are known to be skewed. The family of bivariate skew-normal distributions (Azzalini & Dalla Valle, 1996) reflects the skewness of signal events. The right hand side of Figure 2 shows a sample image on the underlying FACT telescope camera (FACT collaboration, 2014) with a fitted bivariate skew-normal distribution.

Another extension is to incorporate the alignment of signal events to the origin in the modelling. As we know that signal events are aligned to the source (here the camera center), we force the fitted distribution to be aligned, too. In the case of the bivariate Gaussian, we fit the distribution under the constraint that at least one of the eigenvectors u_1 or u_2 has the same direction as μ, that is, one of the semiaxes of the elliptical contour lines is aligned to the camera center.

After fitting a distribution in this way, we can apply the same distance measures as described above.

5 Application to FACT Data

We apply the variable construction described above to simulated data from the FACT telescope. This allows us to assess the quality of the classification with and without the new features.

We fit a bivariate Gaussian distribution, a Skew-normal distribution and a

what we call aligned normal distribution, which includes information about the alignment of signal observations, to the shower. We then evaluate the distance between the observed image and the fitted distributions by using the three distance measures described above.
It can be seen in Table 1 that using the new features leads to significant improvements in the classification errors.
Investigations on the Importance of each variable using the Mean Gini Decrease measured by a random forest showed that especially the aligned skew-normal together with the Kullback-Leibler divergence led to good improvements of the classification.

TABLE 1. The ratio of falsely classified signal and background with standard deviations when using only Hillas variables and combined with the newly constructed variables on a test sample with signal to background ratio of 1:1000.

	Signal		Background	
	Error	sd	Error	sd
Hillas	0.281	0.005	0.0220	0.0002
Hillas+new	0.263	0.006	0.0098	0.0001

6 Conclusion & Outlook

In this paper, we have established a connection between Hillas variables and bivariate Gaussian distributions. We have seen that Hillas variables alone cannot include all information available about signal events and have extended the idea of fitting an ellipse to fitting a bivariate distribution. Although the information on a Gaussian fit is fully given by the Hillas variables, the Gaussian fit allows construction of some additional variables based on distance measures for densities. The additional variables lead to significant improvements in the classification of signal and background events, which is important for further analysis of the signal events. Ongoing work addresses the inclusion of information on both alignment and skewness at the same time by fitting a skew-normal distribution with fixed alignment. Also, variable selection methods might work well in this context, as some of the newly constructed variables and also some of the Hillas variables may be redundant or unimportant for the classification.

Acknowledgments: The work on this paper has been supported by Deut–sche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 “Providing Information by Resource-Constrained Analysis”, project C3 (`http://www.sfb876.tu-dortmund.de`). We gratefully acknowledge the FACT collaboration for supplying us with the test data sets and the ITMC at TU Dortmund for providing computer resources on LiDO.

References

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, **83**, p. 4

The FACT collaboration (2014). The FACT telescope web pages, , URL: http://isdc.unige.ch/fact/

Greenwood, P.E. and Nikulin, M.S. (1996). A Guide to Chi-Squared Testing. In: *Wiley Series in Probability and Statistics*, Wiley

Hillas, A.M. (1985). Cherenkov Light Images of EAS Produced by Primary Gamma. In: *Proceedings of the 19th International Cosmic Ray Conference ICRC*, San Diego, **3**, p. 445

Kullback, S. and Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, **22**, pp. 79 – 86.

The MAGIC collaboration (2014). The MAGIC telescope web pages, URL: https://magic.mpp.mpg.de/home/

de Naurois, M. and Rolland, L. (1996). A high performance Likelihood reconstruction of γ-rays for imaging atmospheric Cherenkov telescopes., *Astroparticle Physics*.

Nikulin, M.S. (2001). Hellinger Distance. In: *Hazewinkel, M.: Encycolpedia of Mathematics*, Springer

Weekes, T. (2003). *Very High Energy Gamma-Ray Astronomy*. Institute of Physics Publishing, Bristol/Philadelphia

Spatial Splines for Generalized Additive Models

Matthieu Wilhelm [1], Laura M. Sangalli[2]

[1] Institut de Statistique, Université de Neuchâtel, Switzerland
[2] Dipartimento di Matematica, Politecnico di Milano, Italy

E-mail for correspondence: matthieu.wilhelm@unine.ch

Abstract: We introduce a novel method for the analysis of spatially distributed data from an exponential family distribution, able to efficiently deal with data occurring over irregularly shaped domains. The proposed generalized additive framework can handle all distributions within the exponential family, including binomial, Poisson and gamma outcomes, hence leading to a very broad applicability of the model. Specifically, we maximize a penalized log-likelihood function where the roughness penalty term involves a suitable differential operator of the spatial field over the domain of interest. Space-varying covariate information can also be included in the model in a semiparametric setting. The proposed model exploits advanced scientific computing techniques and specifically makes use of the Finite Element Method, that provide a basis for piecewise polynomial surfaces.

Keywords: Generalized additive models, spatial regression, finite elements, penalized regression

1 The model and estimation problem

We develop a model that deals with spatially distributed realizations having a distribution within the exponential family. Consider a bounded regular domain $\Omega \in \mathbb{R}^2$ and spatial locations $\mathbf{p}_1, \ldots, \mathbf{p}_n$ scattered over Ω. Let y_i be the variable of interest observed at $\mathbf{p}_i$ with an associated q-vector of covariates $\mathbf{x}_i^t$. Assume $y_1, \ldots, y_n$ have a distribution within the exponential family with canonical parameter $\theta_i = g(\mu_i)$, where $\mu_i = \mathbf{E}[y_i]$ and g is the canonical link function associated with the distribution of interest (hence, the canonical and natural parameter coincide in this case). Assume the following semiparametric generalized linear model:

$$\theta_i = g(\mu_i) = \mathbf{x}_i^t \beta + f(\mathbf{p}_i)$$

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

where $\beta \in \mathbb{R}^q$ contains regression coefficients and the real valued smooth function f, defined over the domain Ω, accounts for the spatial structure of the phenomenon.
We propose to estimate the regression coefficients β and the spatial field f by maximizing the following penalized log-likelihood function (see Wilhelm (2013)):

$$\mathcal{L}_p(\beta, f) = \mathcal{L}(\beta, f) - \frac{1}{2}\ \lambda \int_\Omega (\Delta f)^2 \tag{1}$$

where $\mathcal{L}$ denotes the log-likelihood of the considered distribution. Since the Laplacian of f, Δf, is a measure of local curvature, the second term in (1) penalizes the roughness of the estimated spatial field. In the special case where the considered distribution is the Gaussian, this estimation problem is equivalent to penalized least square error problem considered in Sangalli *et al.* (2013). For distributions other than normal, in order to maximize the penalized log-likelihood (1), we develop a functional generalization of the Penalized Iterative Reweighed Least Squares (PIRLS) algorithm (see e.g., Wood, 2006). The choice of the smoothing parameter can be performed using the GCV criterion (see, e.g., Wahba, 1990).
Likewise in Ramsay (2002) and Sangalli *et al.* (2013), the function f is approximated using a basis expansion provided by finite elements. This makes the model computationally highly efficient and allows to comply with complex domains and prescribed boundary conditions on f. The good performances of the proposed model are illustrated via simulation studies. We apply our model to the estimation of the crime intensity in the city of Portland, Oregon, USA.

Acknowledgments: L. Sangalli acknowledges funding by MIUR Ministero dell'Istruzione dell'Università e della Ricerca, *FIRB Futuro in Ricerca* research project "Advanced statistical and numerical methods for the analysis of high dimensional functional data in life sciences and engineering" (see http://mox.polimi.it/users/sangalli/firbSNAPLE.html), and by the program Dote Ricercatore Politecnico di Milano - Regione Lombardia, research project "Functional data analysis for life sciences".

References

Ramsay, T.O. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society, Series B*, **64**, 307 – 319.

Sangalli, L.M., Ramsay, J.O. and Ramsay, T.O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society, Series B*, **75**, 681 – 703.

Wahba G (1990). Splines models for observational data. In: *SIAM-SIMS Conference Series. Society for Industrial and Applied Mathematics*, Philadelphia.

Wilhelm, M. (2013). Generalized Spatial Regression with Differential Penalization *Master thesis, EPFL*

Wood, S.N. (2006). *Generalized additive models: an introduction with application in R.* New-York: Springer

The Misuse of The Vuong Test For Non-Nested Models to Test for Zero-Inflation

Paul Wilson[1]

[1] School of Mathematics and Computer Science, University of Wolverhampton, United Kingdom, WV1 1LY

E-mail for correspondence: pauljwilson@wlv.ac.uk

Abstract: The Vuong test for non-nested models is being widely misused as a test of zero-inflation. We show that such use of is erroneous and incorrectly assumes that the distribution of the log-likelihood ratios of zero-inflated models versus their non-zero-inflated counterparts is normal. We see that this stems from a mis-understanding of what is meant by the term "non-nested model", and investigate other approaches for determining zero-inflation

Keywords: Non-Nested Models; Zero-Inflation; Vuong Test

1 The Vuong Test for Strictly Non-Nested Models

Zero-inflated models are those based upon mixtures of a zero and a count distribution $f(y;\Theta)$:

$$f(y;\Theta) = \gamma + (1-\gamma)f(0;\Theta) \quad y = 0; \quad (1-\gamma)f(y;\Theta) \quad y = 1,2,3,\ldots \quad (1)$$

The "Vuong Test for Non-Nested Models" test was introduced by Vuong (1989), as a test for "strictly non-nested models". In slightly simplified form, it states that under the null hypothesis that two non-nested models F_θ and G_γ fit equally well, i.e. that the expected value of their log-likelihood ratio equals zero, then under H_0 the asymptotic distribution of the log-likelihood ratio statistic, LR, is normal. In particular, (under H_0):

$$LR_n(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\omega}_n\sqrt{n} \longrightarrow N(0,1) \quad (2)$$

where ω denotes the variance of LR_n and n the sample size. Vuong (1989) also presents tests for nested and overlapping models, and shows that, given certain conditions, their log-likelihood ratios are related to χ^2 distributions. Due to the simplicity of its calculation, the test has become popular among statistical practitioners in various disciplines and is implementable in Stata, and the R-package *pscl* (Jackman, 2012).

This paper was published as a part of the proceedings of the 29th International Workshop on Statistical Modelling, Georg-August-Universität Göttingen, 14–18 July 2014.

2 The Misuse of the Vuong Test

The Vuong test for strictly non-nested models is being widely misused as a test of zero-inflation. For example the help page associated with the *vuong* command in *pscl* states: "*The Vuong non-nested test is based on a comparison of the predicted probabilities of two models that do not nest. Examples include comparisons of zero-inflated count models with their non-zero-inflated analogs (e.g., zero-inflated Poisson versus ordinary Poisson, or zero-inflated negative-binomial versus ordinary negative-binomial).*" Desmarais and Hardin (2013) state that:"*researchers commonly use the Vuong test (Vuong 1989) to determine whether the zero-inflated model fits the data statistically significantly better than count regression with a single equation*" and cite *ten* references to publications that have used the Vuong test for this purpose. That this is an incorrect use of the Vuong test for non-nested models is clearly illustrated by Figure 1. The left histogram illustrates the observed distribution of the log-likelihoods obtained when a one-covariate zero-inflated Poisson (ZIP) model and a Poisson model are fitted to $100,000$ samples of size $n = 100$. Clearly the distribution is non-normal. The $97.5th$ percentiles of the observed distribution is 3.45. The right hand histogram is produced using exactly the same software code that produced the left hand histogram, but here the data is simulated from data that is Poisson distributed on variables x_1, x_2, x_3 and x_4 where each x_i is uniformly distributed, and the "competing" models are on $x_1 + x_2$ and $x_3 + x_4$ respectively, and hence are strictly non-nested according to the definition of Vuong (1989). (Here the observed $97.5th$ percentile is 1.969).

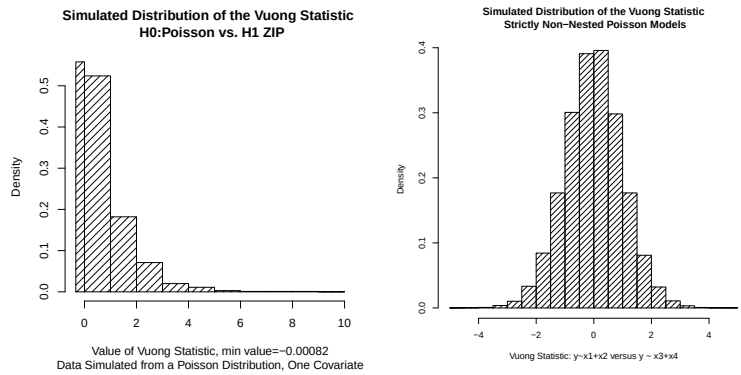


FIGURE 1. Distributions of the Log-likelihood Ratios of ZIP versus Poisson and Strictly Non-Nested Models

Desmarais and Hardin (2013) extensively discuss AIC and BIC type adjustments to the distribution of the log-likelihood ratios, and present evidence that this improves the power of the Vuong test; it should be noted that these adjustments are, for any given comparison of models, constants, and hence only effect the mean of the distribution, not its shape.

2.1 The Cause of The Confusion

The misuse of the test stems from misunderstanding of what is meant by the term "non-nested model". This term, along with "nested model" abounds in statistical literature. As is the case with many frequently used terms their meanings are approximately understood by many, but precisely understood by few. Clarke (2001) observes that "*Defining the concept of 'non-nested' precisely is not an easy task. Definitions are often imprecise and uncomplicated or precise and complicated.*" A statement that applies equally well to nested models. (Indeed one category is not the complement of the other, there exists a third, "inbetween" category of what Vuong (1989) refers to as *overlapping models*.) Simple definitions of nested model include that of Davison (2003): "*Two models are said to be nested if one reduces to the other when certain parameters are fixed.*" Vuong (1989) defines a model G_γ to be nested in a model F_θ by: "G_γ *is nested in* F_θ *if and only if* $G_\gamma \subset F_\theta$.

The distribution of the log-likelihood ratios of non-nested models being normal is dependent however on six assumptions presented elsewhere in Vuong (1989), these refer to various topological and measure theoretic properties. In particular whilst the standard formulation of the probability distribution function (see equation (1)) of a zero inflated Poisson distribution with zero-inflation parameter γ, where $0 \leq 0 \leq 1$, reduces to that of a Poisson distribution when $\gamma = 0$, $-2\times LR$ fails to be χ^2 distributed as $\gamma = 0$ is at the boundary of the parameter space, failing to meet Vuong's prerequisite that it should be interior to the parameter space, this in turn results in non-normality of the sampling distribution of the zero-inflation parameter; as Vuong's subsequent theoretical development of the distributions of log-likelihood ratios (not only of non-nested models, but of nested and overlapping models also), depends upon normality of the sampling distribution of the model parameters, clearly his theory is not applicable.

2.2 Models Fitted Using Link Functions

Zero-inflated models are usually fitted using a logit link to model the expected proportion of perfect zeros. Whilst it is true that the logistic function: $\frac{\exp(t)}{1+\exp(t)} \neq 0$ for all $t \in \mathbb{R}$, and hence in some sense this formulation of the ZIP and Poisson are non-nested, $\lim_{t\to-\infty} \dfrac{\exp(t)}{1+\exp(t)} = 0$, thus this formulation of the zero-inflated model fails to meet Vuong's prerequisite that the parameter space is a compact subset of $\mathbb{R}^p$, and, similar to the scenario presented in the previous section the sampling distribution of the zero-inflation parameter, and hence the distribution of the log-likelihood ratios, is non-normal. Similar statements hold if probit or complementary log-log links are used. It is worth noting that there is confusion in the literature about models being nested if one reduces to the other if certain parameters are fixed, many authors apparently taking this to mean fixed *at zero*. For example, Desmarais and Harden (2013) state: "the count regression f is

not nested in the zero-inflated model, because the model does not reduce to f (the count model) when $\gamma = 0$, in which case the probability of a 0 is inflated by 0.50", apparently alluding to the fact that the value of the logistic function $= 0.5$ when $t = 0$.

2.3 The Null Hypothesis of Vuong's Test

Consistent with a test of zero-inflation, the simulated distribution of the Vuong statistic presented in the left hand side of Figure 1 is derived by resampling from non-zero-inflated data. As stated in Section 1 the null hypothesis of Vuong's test for non-nested models is that the expected value of their log-likelihood ratios equals zero, this implies that under the null hypothesis both models are "equally far away" from the data that is being modelled. If we temporarily ignore the issue of whether zero-inflated models and their non-zero-inflated counterparts are non-nested or otherwise, and consider them non-nested, to appropriately simulate the distribution of the log-likelihood ratios it would be necessary to resample from data that was somehow equidistant from zero-inflated and non-zero-inflated data, it is difficult to envisage the nature of such data. More importantly, non-rejection of the null hypothesis of Vuong's test for non-nested models, where the (supposedly) non-nested models are, say, the zero-inflated Poisson and standard Poisson model would mean that there is no evidence to conclude that either model fits the data better than the other, not that there is no evidence to support zero-inflation, and its rejection simply implies that either the zero-inflated Poisson model fits the data better than the Poisson model, or vice-versa, not that zero-inflation is present or absent.

3 Other Approaches

Distributional Methods: Early research by the author indicates that if negative values of the log-likelihood ratio that are very close to zero are considered as zeros, then the distribution of ZIP versus Poisson log-likelihood ratios, where the zero inflation parameter is only modelled by an intercept, is a mixture of a point mass at zero and a χ^2_1 distribution, the weighting of the mixture being dependent on the number of covariates. If the zero-inflation parameter is modelled, a mixture of a zero point mass and some other distribution still occurs; whilst the nature of this other distribution is yet to be determined, a weighted mixture of χ^2 distributions is certainly a candidate, this being consistent with Vuong's theory of overlapping models, but further research is necessary. Note that when the value of γ is allowed to be both positive or negative fitted values of γ do not "pile up" close to zero, and the distribution of zero-modification parameter is normal and hence a non-zero-inflated model is *nested* in its zero-inflated counterpart, and hence a Vuong test for nested models could be used as a test of zero-inflation/deflation. Wilson (2010) presents a method of adapting zero-deflated data so that zero-modified models may be fitted via standard

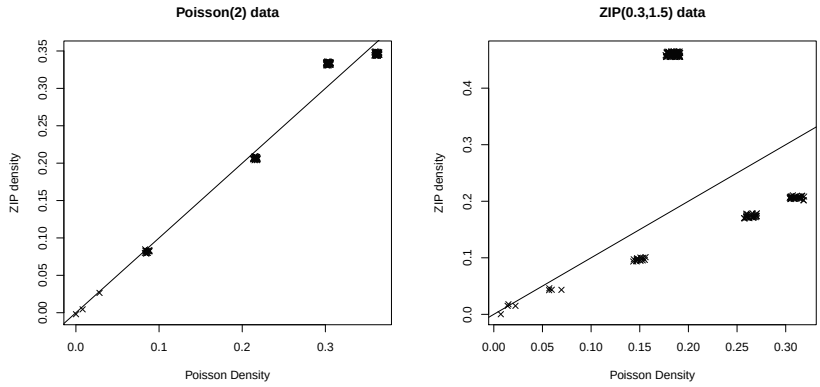


FIGURE 2. Fitted Probabilities under both models plotted against each other.

software for fitting zero-inflated models. The lack of software for fitting zero-modified models is due to the fact that the standard link-functions employed to fit zero-inflated models are incompatible with zero-deflation. Dietz and Böhning (2000) proposed a link function that allowed for zero-deflation, and more recently Todem, Hsu and Kim (2012) have proposed a score test that incorporates a link function that allows for both zero-inflation and deflation.

Graphical Methods: Many statistical tests for determining normality (or otherwise) of data exist, practitioners nearly invariably rather use a "normal QQ plot" to assess normality. A parallel here would be to plot the individual fitted probabilities (contributions to the likelihood) of the observed data under the zero-inflated and the non-zero inflated model against each other, if the points lie approximately along the line $x = y$, then zero-inflation is not indicated. Examples are shown in Figure 2, the left-hand diagram where 150 data have been simulated from a Poisson(1) distribution, and the lower diagram where 150 data have been simulated from zero-inflated Poisson data with Poisson mean 1.5 and zero-inflation parameter 0.3. "Jitter" has been applied in both diagrams. We see that in the top diagram the points fall approximately along the line $x = y$, consistent with lack of zero-inflation, whereas in the lower diagram, where the zero-inflated model is appropriate, this is not the case. Another approach would be to plot contributions to the log likelihood under both models against each other.

4 Conclusion

It is beyond doubt that the widespread practice of using Vuong's test for non-nested models as a test of zero-inflation is erroneous. The misuse is rooted in a misunderstanding of what is meant by the term "non-nested model". The derivation of the distribution of the log-likelihood ra-

tios of zero-inflated versus non-zero inflated models is not straightforward, and possible alternative approaches are to develop tests based upon zero-*modified* models where the zero-"inflation" parameter may be negative, or to develop graphical methods.

References

Clarke, K. (2001). Testing nonnested models of international relations:re-evaluating realism. *American Journal of Political Science*, **45**, 724 – 744.

Davison, A. (2003). *Statistical Models*. Cambridge University Press.

Desmarais, B.A., and Harden, J.J. (2013). Testing for zero inflation in count models: Bias correction for the Vuong test. *The Stata Journal*, **13**, 810 – 835.

Dietz D. and Böhning D. (2000) On the estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and data Analysis* **34**, 441 – 459.

Jackman, S. (2012). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. Stanford University. R package version 1.04.4.
URL http://pscl.stanford.edu/

Todem, D., Hsu, W and Kim, K (2012) On the Efficiency of Score Tests for Homogeneity in Two-Component Parametric Models for Discrete Data. *Biometrics* **68** 975 - 982.

Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307 – 333.

Wilson, P. (2010) Zero Augmentation: A method for fitting zero-modified count models that allow both zero-inflation and deflation. In Bowman, A. ed. *Proceedings of the 25th International Workshop on Statistical Modelling*, University of Glasgow, 575 - 580

Index

Abegaz, Fentaw, 225
Aerts, M, 235

Baker, Peter, 207
Bakka, Haakon, 51
Barrientos, Andrés, 15
Bender, Andreas, 57
Bink, Marco, 117
Brockhaus, Sarah, 63
Bruyneel, Luk, 185

Cadarso-Suárez, Carmen, 111, 145, 167
Carmada, Carlo, 69
Cederbaum, Jona, 75
Chebon, Sammy, 81
Chiann, Chang, 231
Chiogna, Monica, 99
Claeskens, Gerda , 273
Crespo–Cuaresma, Jesus, 163
Currie, Iain, 87
Cysneiros, Audrey, 93

de Bastiani, Fernanda, 93
de Falguerolles, Antoine, 3
de Smedt, Ann, 81
de Sousa, Bruno, 111
Diop, Aliou, 239
Djordjilovic, Vera, 99
Donoghoe, Mark, 105
Duarte, Elisa, 111
Dunlop, Lindsay, 157
Dupuy, Jean-François, 239

Eilers, Paul, 117, 133, 331
Etxeberria, J, 123

Faes, C, 235
Faes, Christel, 81, 145, 291
Fasola, Salvatore, 127
Feldkircher, Martin, 163
Frasso, Gianluca, 133
Fried, Roland, 353
Friede, Tim, 303

García-Zattera, María José, 173
Gertheiss, Jan, 139, 219
Geys, Helena, 81
Goicoa, T, 123
Grün, Bettina, 163
Greven, Sonja, 63, 75, 325
Guédon, Yann, 269
Gude, Francisco, 145, 167
Guler, Ipek, 145

Harbering, Jonas, 201
Hartl, Wolfgang, 57
Hawe, David, 341
Haynes, Michele, 207
Heiner, Karl , 151
Held, Leonhard, 213
Heller, Gillian, 157
Hessel, Engel, 139
Heumann, Christian, 297
Hinde, John , 151
Hofmarcher, Paul, 163
Hoole, Phil, 75
Hothorn, Torsten, 63
Humer, Stefan, 163

Jahfari, Sara , 273
Jara, Alejandro, 15, 173

Küchenhoff, Helmut, 57
Klein, Nadja, 167
Kneib, Thomas, 111, 167, 201, 243, 337
Koeble, Renate, 179

Komárek, Arnošt, 173

Lamboni, Matieyendou, 179
Leão, D, 307
Leip, Adrian, 179
Lesaffre, Emmanuel, 185, 307
Li, Baoyue, 185
Lovison, Gianfranco, 189

Möller, Annette, 219
Maier, Verena, 139
Manisera, Marica, 195
Manitz, Juliane, 201
Marquart, Louise, 207
Marschner, Ian, 105
Massa, Maria Sofia, 99
Meyer, Sebastian, 213
Militino, A, 123
Mohammadi, Abdolreza, 225
Molenberghs, Geert, 291
Montoril, Michel, 231
Morettin, Pedro, 231
Muggeo, Vito M.R, 127
Mutambanengwe, Chenjerai, 235

Ndao, Pathé, 239

O'Sullivan, Finbarr, 341
Oelker, Margret, 243
Ogden, Helen, 249

Pan, Jianxin, 255
Pan, Yi, 255
Pauger, Daniela, 261
Peyhardi, Jean, 269
Pircalabelu, Eugen , 273
Pouplier, Marianne, 75
Puig, Pedro, 285
Pößnecker, Wolfgang, 279

Röver, Christian, 303
Rabe-Hesketh, Sophia, 27
Rakhmawati, Trias, 291
Ramzan, Shahla, 297
Rigby, Robert, 93
Rodrigues, Vitor, 111
Romualdi, Chiara, 99
Russo, Cibele, 307

Sangalli, Laura, 359
Scarrott, Carl, 313
Schöbel, Anita, 201
Schauberger, Gunther, 319
Scheipl, Fabian, 57, 63, 325
Schindler, Christian, 189
Schmidt, Marie, 201
Schnabel, Sabine, 331
Skrondal, Anders, 27
Sobotka, Fabian, 243, 337
Staicu, Ana-Maria, 139, 325
Sweeney, James, 341

Trottier, Catherine, 269
Tutz, Gerhard, 29, 219, 279, 297, 319

Ugarte, M, 123

Van den Hout, Ardo, 347
Verbeke, Geert, 291
Voigt, Tobias, 353

Wagner, Helga, 261
Waldorp, Lourens , 273
Wilhelm, Matthieu, 359
Willemsen, S, 307
Wilson, Paul, 363
Wit, Ernst, 225
Wood, Simon N., 43

Zuccolotto, Paola, 195

29th IWSM 2014 Sponsors

We are very grateful to the following organisations for sponsoring 29th IWSM 2014.

- Georg-August-University, Göttingen
- Toyota Motor Corporation
- Leonard N. Stern School of Business,
  New York University
- Springer Publisher
- Statistical Modelling Society
- RTG 1644 "Scaling Problems in Statistics"