UNIVERSITÀ DEGLI STUDI DI TRIESTE

IWSM
International Workshop on Statistical Modelling

Proceedings of the 36th International Workshop on Statistical Modelling

July 18-22, 2022 - Trieste, Italy

Editors
Nicola Torelli
Ruggero Bellio
Vito Muggeo

Proceedings of the 36th International Workshop on Statistical Modelling

July 18-22, 2022 - Trieste, Italy

Editors
Nicola Torelli
Ruggero Bellio
Vito Muggeo

EUT EDIZIONI UNIVERSITÀ DI TRIESTE

International Workshop on Statistical Modelling (36°. 2022. Trieste)

Proceedings of the 36th International Workshop on Statistical Modelling : July 18-22, 2022 Trieste, Italy / Nicola Torelli, Ruggero Bellio, Vito Muggeo (editors). – Trieste : EUT Edizioni Università di Trieste, 2022. – 1 risorsa online : PDF (629 p. : ill.)

ISBN: 978-88-5511-309-0

Autori:

Torelli, Nicola
Bellio, Ruggero
Muggeo, Vito

Soggetti:

1. Statistica – Congressi. 2. Modelli econometrici – Congressi
330.015195 = Statistica matematica

Editors:

NICOLA TORELLI
University of Trieste, DEAMS

RUGGERO BELLIO
University of Udine, DIES

VITO MUGGEO
University of Palermo, SEAS

ISBN 978-88-5511-309-0 (online)

EUT Edizioni Università di Trieste
via Weiss 21 – 34128 Trieste
http://eut.units.it
https://www.facebook.com/EUTEdizioniUniversitaTrieste

Scientific Programme Committee

Serena Arima
University of Salento (Italy)

Ruggero Bellio
University of Udine (Italy)

Carlo Giovanni Camarda
Institut National d'études démographiques (France)

Pierpaolo De Blasi
University of Turin (Italy)

Maria Luz Durban
Universidad Carlos III de Madrid (Spain)

Jochen Einbeck
University of Durham (UK)

Jutta Gampe
Max Planck Institute for Demographic Research (Germany)

Goeran Kauermann
Ludwig Maximilians Universität (Germany)

Antonio Hermes Marques da Silva Júnior
Federal University of Rio Grande do Norte (Brasil)

Vito Muggeo
University of Palermo (Italy)

María Xosé Rodríguez-Álvarez
BCAM - Basque Center for Applied Mathematics & IKERBASQUE, Basque Foundation for Science (Spain)

Sabine Schnabel
Wageningen University & Research (Netherlands)

Nicola Torelli (Chair)
University of Trieste (Italy)

Helga Wagner
Johannes Kepler University Linz (Austria)

Local Organizing Committee

Michela Battauz
University of Udine, DIES

Ruggero Bellio
University of Udine, DIES

Domenico De Stefano
University of Trieste, DiSPeS

Gioia Di Credico
University of Trieste, DEAMS

Leonardo Egidi
University of Trieste, DEAMS

Giovanni Fonseca
University of Udine, DIES

Vincenzo Gioia
University of Udine, DIES, and University of Trieste, DEAMS

Luca Grassetti
University of Udine, DIES

Giovanni Millo
University of Trieste, DEAMS

Vito Muggeo
University of Palermo, SEAS

Gabriella Schoier
University of Trieste, DEAMS

Roberta Pappadà
University of Trieste, DEAMS

Francesco Pauli
University of Trieste, DEAMS

Laura Rizzi
University of Udine, DIES

Marco Stefanucci
University of Rome - La Sapienza, DSS

Nicola Torelli (Chair)
University of Trieste, DEAMS

Susanna Zaccarin
University of Trieste, DEAMS

Preface

We are very pleased to host the 36th International Workshop on Statistical Modelling (IWSM). This edition of the workshop is the fir t one held in presence after a two-year hiatus due to the COVID-19 pandemic, and we are delighted to meet in person all the attendees in Trieste.

This edition is going to be quite lively, with 60 oral presentations and 53 posters, covering a vast variety of topics. The numbers of both the original submissions and the final papers are probably higher than we were expecting. Indeed, due to the uncertainty related to the pandemic trend of this past winter, it was not easy to predict such good numbers! Luckily, we received a large number of excellent papers, and it was challenging for the Scientific Committee to make a selection for the oral talks. However, regardless of their final status, all the presentations stand out for their noteworthy contribution to statistical modelling, and we hope they can catch your interest. As usual, the extended abstracts of the papers are collected in the IWSM proceedings, but unlike the previous workshops, this year the proceedings will be not printed on paper: the IWSM goes *light*!

As customary for the IWSM, we have a thrilling group of plenary talks, covering various areas of statistics, and we greatly thank the invited speakers Anthony Davison, Héléne Jacqmin-Gadda, Ioannis Kosmidis, Claudia Czado and Antonio Canale. A special thank goes to Ioannis Ntzoufras and Leonardo Egidi for the short course *Statistical Modelling of Football Data.*

The workshop proudly maintains its almost unique feature of scheduling one plenary session for the whole week. This choice has always contributed to the stimulating atmosphere of the conference, combined with its informal character, encouraging the exchange of ideas and cross-fertilization among different areas

As a distinguished tradition of the workshop, student participation has been strongly encouraged. This IWSM edition is particularly successful in this respect, as testified by the large number of students included in the program. We award three students for the best paper, the best oral presentation, and the best poster respectively. Furthermore, two student travel grants have been kindly provided by the Statistical Modelling Society.

Finally, the organization of this conference is the result of all the hard work of the Scientific Committee and the Local Organizing Committee. We eagerly thank all of you for your exceptional effo ts to bring IWSM back on track.

Welcome to Trieste, and enjoy the conference.

Nicola, Ruggero and Vito
Trieste, Udine and Palermo, June 2022

Contents

Part I

1ϕ Antonio Canale
Bayesian dimensionality reduction

2ϕ Claudia Czado
Vine copula based stress testing

3c Anthony Davison
How long could a human live?

4л Hélène Jacqmin-Gadda, Léonie Courcoul, Antoine Barbieri, Hugues De Courson, Christophe Tzourio
Is blood pressure variability a risk factor for cerebro-vascular event? Joint modelling approaches

пс Ioannis Kosmidis
Improved estimation through additive adjustment of estimating functions

Part II

53 Timo Adam, Richard Glennie, Théo Michelot
State-switching varying-coefficient stochastic dii erential equations

58 María Alonso-Pena, Rosa M. Crujeiras
Modelling zebrafish escape strategies with circular modal regression

64 Umut Altay, John Paige, Andrea Riebler, Geir-Arne Fuglstad
Spatial Modelling with Covariates for Survey Data with Positional Uncertainty

69 Mario Angelelli, Christian Catalano
A quantile regression ranking for cyber-risk assessment

74 David Aristei, Silvia Bacci, Manuela Gallo, Maria Iannario
Multidimensional latent mixed models for don't know responses in panel data concerning *Financial knowledge*

80 Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto, Carla Rampichini
Evaluating the school ei ect using multilevel models: adjusting for pre-test or using gain scores?

85 Clara Bertinelli Salucci, Azzeddine Bakdi, Ingrid Kristine Glad, Erik Vanem, Riccardo De Bin
A novel semi-supervised learning approach for maritime lithium-ion battery monitoring

91 Nicolas Bianco, Mauro Bernardi, Daniele Bianchi
Bernoulli-Gaussian model for dynamic sparsity in time varying parameter regression

97 Guillermo Briseño Sanchez, Andreas Groll
Bivariate mixed binary-survival additive regression modelling

103 Silvia Brosolo, Alessandra R. Brazzale, Giovanna Menardi
Locating γ-Ray Sources on the Celestial Sphere via Mixture Models

107 Paul T. Brown, Chaitanya Joshi, Deane Searle, Stephen Joe
Spatial anisotropic modelling of the repeat/near repeat victimisation phenomenon

113 Viviana Carcaiso, Leonardo Grilli
Analysis of count data by quantile regression coefficient modelling: student's gained credits after online teaching

117 Manuel Carlan, Thomas Kneib
Bayesian Discrete Conditional Transformation Models

123 Angela Carollo, Hein Putter, Paul Eilers, Jutta Gampe
Competing risks models with two time scales

129 Roberto Colombi, Sabrina Giordano, Maria Kateri
A Responding Attitude Component in Hidden Markov Models

134 Enrico A. Colosimo, Emilly M. Lima, Maria C.P. Nunes
Alternative Approaches to Dynamic Predictions: An Application on a Cohort of Patients with Chagas Disease

138 Ilse Cuevas Andrade, Ardo van den Hout, Nora Pashayan
Multi-state models as an alternative to joint models; estimating treatment ei ects using longitudinal data

144 Daniele Cuntrera, Vito M.R. Muggeo, Luigi Augugliaro
Variable Selection with Quasi-Unbiased Estimation: the CDF Penalty

150 Fernanda De Bastiani, Mikis D. Stasinopoulos, Robert A. Rigby, Thomas Kneib
Exploring relationships using non-linear correlation coefficients

156 Hortense Doms, Philippe Lambert, Catherine Legrand
Flexible joint model for time-to-event and non-Gaussian longitudinal outcomes

162 Paul H.C. Eilers
Smoothing on Networks with P-splines

167 Laura **F**reijeiro-González, Manuel **F**ebrero-Bande, Wenceslao González-Manteiga
Detection of relevant covariates involved in casual rentals in the Capital bikeshare program of Washington, D.C., by means of new nonparametric specification tests for additive concurrent models formulation

172 Vincenzo Gioia, Matteo Fasiolo, Ruggero Bellio
A comparison of unconstrained parameterisations for additive mean and covariance matrix modelling

178 Caterina Gregorio, Giulia Barbati, Francesca Ieva
Personalised ei ect of discontinuing treatment in Heart Failure patients through multi-state modeling

183 Oswaldo Gressani, Niel Hens, Christel Faes
The power of Laplacian-P-splines for inference in epidemiological and survival models

187 Colin Griesbach, Andreas Mayr, Elisabeth Bergherr
Variable Selection and Allocation in Joint Models via Gradient Boosting Techniques

193 Katharina Hechinger, Xiao Xiang Zhu, Göran Kauermann
Categorizing the World into Local Climate Zones - Towards Quantifying Labelling Uncertainty for Machine Learning Models

199 Pavel Hernández-Amaro, María Durbán, María Carmen Aguilera-Morillo, Cristobal Esteban Gonzalez, Inma Arostegui
Functional regression on variable domains: a penalized approach

205 James Jackson, Robin Mitra, Brian Francis, Iain Dove
Using saturated models for data synthesis

211 Michele Lambardi di San Miniato, Ruggero Bellio, Luca Grassetti, Paolo Vidoni
Robust regression and adaptive filterin

217 Philippe Lambert, Michaela Kreyenfeld
Laplace approximation for penalty selection in double additive cure survival model with exogenous time-varying covariates

222 Timo Lohrmann, Anders Kvellestad, Riccardo De Bin
A modified dividing local Gaussian processes algorithm for theoretical particle physics applications

228 Maritza Márquez, Cristian Meza, Dae-Jin Lee, Rolando de la Cruz
Estimation and classification in semiparametric nonlinear mixed models using P-Splines and the SAEM algorithm

234 Andrea Mascaretti, Antonio Canale
Bayesian Mixtures of Envelope Models

238 Andrew McInerney, Kevin Burke
Statistical Information-Criteria-Based Neural Network Input and Hidden Node Selection

242 Rouven Michels, Marius Ötting, Roland Langrock
A varying coefficient state-space model for investigating betting behaviour within in-play markets

247 Alise Danielle Midtfjord, Riccardo De Bin, Arne Bang Huseby
A boosting model for survival analysis with dependent censoring

253 Thomas Minotto, Ingrid Hobæk Haff, Geir Kjetil Sandv
Detecting statistical interactions in immune receptor datasets

258 David Moriña, Amanda Fernández-Fontelo, Alejandra Cabaña, Argimiro Arratia, Pedro Puig
Estimated Covid-19 burden in Spain: ARCH underreported non-stationary time series

264 Pouyan Nejadi, Davood Roshan, John Newell
Movement Pattern Discovery With Applications In Elite Soccer

269 Meadhbh O'Neill, Kevin Burke
Automatic Variable Selection in Distributional Regression Models using a Smooth Information Criterion

274 John Paige, Geir-Arne Fuglstad, Andrea Riebler, Jon Wakefield
Aggregating from Point to Areal Prevalences: A complete Population Model

280 Ioannis Papageorgiou, Ioannis Kontoyiannis
Bayesian mixture models for time series based on context trees

284 Jennifer Pohle, Johannes Signer, Ulrike Schlägel
Markov-switching step selection analysis

289 Anja Rappl, Thomas Kneib, Stefan Lang, Elisabeth Bergherr
Spatial Joint Models through Bayesian Structured Additive Joint Modelling for Longitudinal and Time-to-Event Data

294 Dayasri Ravi, Andreas Groll, Tamara Schikowski
Non-linear modelling of systolic and diastolic blood pressures via environmental factors

300 Davood Roshan, John Ferguson, Charles R. Pedlar, John Newell
Univariate and Multivariate Adaptive Reference Ranges for Longitudinal Monitoring

305 Abdul Salam, Marco Grzegorczyk
Learning the structure of the mTOR protein signalling pathway

311 Gunther Schauberger, Luana Fiengo Tanaka, Moritz Berger
Tree-based Modeling of Confounding Ei ects in Matched Case-Control Studies

317 Reto Stauffer, Moritz N. Lang, Achim Zeilei
Graphical Assessment of Probabilistic Precipitation Forecasts

321 Philipp Sterzinger, Ioannis Kosmidis
Maximum softly-penalized likelihood for mixed ei ects logistic regression

327 Mattia Stival, Mauro Bernardi, Manuela Cattelan, Petros Dellaportas
Longitudinal clustering of athletes' careers under informative missing data patterns

333 Lawrence Tray, Ioannis Kontoyiannis
The Feature-First Block Model

337 Jan Vávra, Arnošt Komárek
GLMM Based Clustering of Multivariate Mixed Type Longitudinal Data

343 Massimo Ventrucci, Garritt Page
Adjusting for spatial confounding using eigendecomposed CAR models

348 Veronica Vinciotti, Ernst Wit
Hierarchical graphical modelling of microbiome interactions in related environments

354 Alexandra Welsh, Deborah A. Costain, Andrew C. Titman
Estimating quality-adjusted life-years: Assessing the bias induced for a terminal decline quality of life model

360 Boyao Zhang, Colin Griesbach, Elisabeth Bergherr
Bayesian Boosting for Simultaneous Estimation and Selection of Fixed and Random Ei ects in High-Dimensional Mixed Models

366 Patrick Zietkiewicz, Ioannis Kosmidis
Mean and median bias reduction in generalized linear models with large data set

Part III

373 Adetola Adedamola Adediran, Jack Noonan, Robin Mitra, Stefanie Biedermann
Comparing recovery sample designs to test for the presence of MNAR

379 Giuseppe Alfonzetti, Ruggero Bellio
Reliable generalized linear latent variable models estimation via simulated maximum likelihood

385 Rosario Barone, Andrea Tancredi
Bayesian mixtures of discretely observed multi-state models

390 Michela Battauz, Paolo Vidoni
Boosting for variance components in mixed models

394 Laurens Bogaardt, Anoukh van Giessen, Susan Picavet, Hendriek C. Boshuizen
A Model of Individual BMI Trajectories

400 Cristian Castiglione, Mauro Bernardi
Probabilistic load forecasting via dynamic quantile regression

406 Annalisa Cerquetti
On the first two size-biased picks from the normalized Inverse Gaussian prior

410 Daniele Cuntrera, Vito M.R. Muggeo
Adaptive P-splines via L_1-type penalty in generalized additive models

414 Nicoletta D'Angelo, David Payares, Giada Adelfio, Jorge Mate
Hawkes processes on networks for crime data

418 Willian L. de Oliveira, María Durbán, Carlos A.R. Diniz
Bernoulli-exponential semiparametric regression model

423 Claudia Di Caterina, Davide Ferrari
Sparse composite likelihood selection

427 John Ferguson, Alberto Alvarez, Catriona Reddin, Robert Murphy, Martin O'Donnell
Quirks with joint hierarchical models: Examples involving global relationships between sodium and potassium intake, GDP and healthy life expectancy

433 Edoardo Filippi-Mazzola, Federica Bianchi, Ernst Wit
Causes of the decrease of patent similarities from 1976 to 2021

437 Giovanni Fonseca, Federica Giummolè, Paolo Vidoni

Probabilistic prediction: aims and solutions

442 Alexander Gerharz, Carmen Ruff, Andreas Groll, Andreas D. Mei
Model Selection for Predicting Readmissions with Health Insurance Data

448 Niamh Graham, Natalia Bochkina, Damian Mole
Estimating periodicity in disease dynamics

452 Alba Halliday
Identification of Possible COVID-19 Pandemic Impact on Scottish National Therapeutic Indicators

458 Luke Hardcastle, Samuel Livingstone, Claire Black, Federico Ricciardi, Gianluca Baio
A Bayesian hierarchical model for improving exercise rehabilitation in mechanically ventilated ICU patients

462 Sargon Hasso, Kenan M Matawie
Experimental Results and Comparisons of Semantically Enriched Query Alternatives in Information Retrieval Models

467 Aisouda Hoshiyar, Jan Gertheiss
Regularization and Model Selection for Item-on-Item Regression

472 Caizhu Huang, Nicola Sartori
Directional test for the comparison of moderate-dimensional normal mean vectors

478 Aliaksandr Hubin, Riccardo De Bin
On a genetically modified mode jumping MCMC approach for multivariate fractional polynomials

484 Aoife K Hurley, James Sweeney
Joint Species Spatial Modelling of Deer Count Data: A Simulation Study

489 Ivana Malá, Adam Čabla
Comparison of models of the unemployment duration in the Czech Republic

495 Rui Martins, Lisete Sousa, Iúri Correia, Inês Farias, Ivone Figueiredo
Modelling the population dynamics of the Blackspot Seabream (Pagellus Bogaraveo) on the Portuguese coast

499 Owen McGrath, Kevin Burke
Binomial Confidence Intervals for Rare Events: Importance of Defining Margin of Error Relative to Magnitude of Proportion

503 Laura McQuaid, Shirin Moghaddam, Kevin Burke
Penalized Power-Generalized Weibull Distributional Regression

508 Lethani Ndwandwe, J.S. Allison, M. Smuts, I.J.H. Visagie
On new classes of tests for the Pareto distribution based on the empirical characteristic function

513 Lizzie Neumann, Jan Gertheiss
Covariate-adjusted Association of Sensor Outputs

517 Thobeka Nombebe, Leonard Santana, James S. Allison, Jaco Visagie
Investigating dii erent parameter estimation techniques for the Lomax distribution

521 Thiago P. Oliveira, Jana Obšteter, Ivan Pocrnic, Gregor Gorjanc
A method for partitioning trends in genetic mean and variance

527 Hyebin Park, Juyong Hong
A combination technique for unbalanced binary classification using artificial neural networ

531 Jeong-Soo Park, Thanawan Prahadchai
Fast computation for performance and independence weighting of climate multi-models

536 Diana M. Pérez-Valencia, María Xosé Rodríguez-Álvarez, Martin P. Boer, Fred A. van Eeuwijk
Performance of spatio-temporal hierarchical P-spline models using simulated data

Christian Pfeifer, Sabine Danzinger, Christian Singer
'Predicted conversion rate' (PCR) of neoadjuvant treatment in relation to the (y)pN stage of HER2–positive breast cancer cases

Thanawan Prahadchai
Analysis of maximum precipitation in Thailand using ensemble of non-stationary extreme value models

Thanawan Prahadchai, Jeong-Soo Park
Projecting Extreme Precipitation in the Philippines using the CMIP6 Multi-model Ensemble

Ilaria Prosdocimi
Extending Generalized Additive Models for extreme value modeling: a software review

Juan M. Rodríguez-Díaz
Review of covariance structures of multiresponse and multisubject models from the point of view of Optimal Design of Experiments

Cristian Roner, Claudia Di Caterina, Davide Ferrari
Robust zero-inflated interval regression for cyber security survey data

Abdul Salam, Marco Grzegorczyk
A new refined non-homogeneous Bayesian network with globally coupled interaction parameters

Yire Shin, Jeong-Soo Park
Statistical Downscaling of Air Temperature using Variational Autoencoded Regression

577 Shubhangi Sikaria, Rituparna Sen
Option pricing using Hawkes Process

582 Bernhard Spangl, Matthias Medl, Johannes Tintner
MD-dating of pinewood using FTIR-spectroscopy and statistical learning algorithms like random forests and CNNs

586 Federica Stolf, Antonio Canale
Bayesian spatial modeling of extreme precipitation

591 Jasper ten Dam, A.J. Rodenburg, K. Katona, A. van Giessen
A Model for Alcohol Consumption Trajectories

597 Vávra, Arnošt Komárek
Model based clustering of households from the EU-SILC database

603 Hiu Ching Yip, Gianluca Mastrantonio, Enrico Bibbona, Marco Gamba, Daria Valente
Nearest Neighbours Gaussian Process Model for Time-Frequency Data: An Application in Bio-acoustic Analysis

608 Daniele Zago, Antonio Canale, Marco Stefanucci
Bayesian multiscale mixtures of multivariate Gaussian kernels for density estimation

612 Yingjuan Zhang, Jochen Einbeck
Simultaneous linear dimension reduction and clustering with flexible variance matrice

618 Dafne Zorzetto, Falco J. Bargagli-Stoffi, Antonio Canale, Francesca Dominici
Dependent Dirichlet Mixture Processes for Causal Inference

624 Lore Zumeta-Olaskoaga, Andreas Bender, Helmut Küchenhoff Dae-Jin Lee
Modelling the recurrence of injuries in football players using piece-wise exponential additive mixed models

Part I

Bayesian dimensionality reduction

Antonio Canale[1]

[1] Università degli Studi di Padova, Italy

E-mail for correspondence: antonio.canale@unipd.it

Abstract: Modern applications generate highly-complex multidimensional data whose analysis is inherently cursed by the notorious curse of dimensionality. To face this phenomenon, it is popular to assume that the data actually lie in a lower dimensional subspace. Principal component analysis, for example, is an ubiquitous standard technique building upon such assumption. In this paper I will review some recent contributions that exploit this idea, under a Bayesian context. I will focus on some of the fundamentals problems in statistics, and specifically on density estimation and factor analysis, regression, and clustering.

Keywords: Clustering; Curse of dimensionality; Envelope Models; Factor analysis; Multivariate regression.

1 Introduction

In many modern applications, it is routine to collect high-dimensional data $y_i = (y_{i1}, \ldots, y_{ip})$ for $i = 1, \ldots, n$, with p (dimension of the data) being larger than the sample size n. For example, in modern biomedical studies we may have data consisting of a variety of high-dimensional biomarkers for each single patient in the study. The advances in technology allow, at least conceptually, to increase p to arbitrarily high values by including multiple types of data including omics data, medical imaging, monitor information, etc.

A common problem in high dimensional statistics, which is exacerbated in these contexts, is the so called *curse of dimensionality* (Bellman, 1961). Under the assumption that the data concentrate near a low-dimensional subspace, *dimensionality reduction* techniques, mapping each y_i in a $d \ll p$ space, become the standard tools to combat the curse of dimensionality. For example, principal component analysis and factor analysis are everlasting successfull tools that work under the assumption of linear mapping. A side benefit of such decompositions is the possibility of interpreting the

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

transformed variables (or latent factors) gaining not only on the side of parsimony in data modeling but also in a better understanding of the principal features driving the phenomena under study.
In this paper I will review some recent contributions that specify these ideas, in a Bayesian context. In Section 2 I will describe a class of infinite factorization models which exploits the Bayesian learning mechanism in two respects: through a shrinkage prior to induce a penalty on the magnitude of the latent dimension and through a structured sparsity prior to emphasize the latent factor interpretability. In Section 3 I will discuss high-dimensional model-based clustering, showing an underrated aspect of the curse of dimensionality in these settings. To solve this issue I will present a first solution based once again on a latent factor representation showing how to turn the curse of dimensionality into a blessing. Finally, in Section 4, I will discuss issues in multivariate regression where both the response variable and the regressors are high-dimensional vectors. Specifically, I will briefly present some preliminary work on Bayesian envelope models and their extension in the context of Bayesian nonparametric mixtures. The paper ends with a brief discussion of the open points of the presented approaches.

2 Factor Analysis

Factorization models are routinely used in psychometrics, biology, marketing, and finance both as dimensionality reduction tool and as descriptive tools linking the p-dimensional numbers of observed variables into a smaller number of underlying variables which may represent some interpretable trait for the phenomena under study.
To formalize, consider the following simple Gaussian factor model for y_i,

$$y_i = \Lambda\eta_i + \epsilon_i, \quad \epsilon_i \sim N(0, \Sigma), \tag{1}$$

with Λ a $p \times k$ loadings matrix, η_i a k dimensional factor, and ϵ_i a p-dimensional independent noise. In this context, the dimensionality reduction is achieved choosing $k \ll p$.
Although there is a rich literature on choosing k in factorization models, selection of k is far from being solved. We focus here on a recent approach based on over-fitted factorizations, which include more than enough components with shrinkage priors adaptively removing unnecessary ones by shrinking their coefficients close to zero (Bhattacharya and Dunson, 2011; Legramanti et al., 2020)
Although these approaches are widely used in many applications, they lack of consideration of possible structured sparsity. In many applications, indeed, the p variables may be associated to known chracteristics in the form of *meta covariate*. For example, the p variables may correspond to different genes in genomic applications which may be associated to known gene

pathways. Current methodologies, however, focused on priors for Λ that are exchangeable within columns but clearly genes are not exchangeable. The idea of using meta covariates is consistent with the group-lasso of Yuan and Lin (2006) adopted in structured penalized regression. Motivated by this, in the next section I review a recent approach for Bayesian structured sparsity in factor loadings introduced by Schiavon et al. (2022).

2.1 Structured Increasing Shrinkage Priors

Following Schiavon et al. (2022), a structured sparsity structure on Λ is induced through a hierarchical shrinkage prior. Specifically

$$\lambda_{jh} \mid \theta_{jh} \sim N(0, \theta_{jh}), \quad \theta_{jh} = \tau_0 \, \gamma_h \phi_{jh}, \tag{2}$$

where the local ϕ_{jh}, column-specific γ_h, and global τ_0 scales are assigned suitable independent distributions on $[0, \infty)$.
Differently from most of the existing literature on shrinkage priors, Schiavon et al. (2022) define a non-exchangeable structure that includes *meta covariates* x informing the sparsity structure of Λ. In many applications, meta covariates provide information to distinguish the p different variables as opposed to traditional covariates that serve to distinguish the n subjects. Letting x denote a $p \times q$ matrix of such meta covariates, the distribution for ϕ_j is defined not depending on the index h and such that

$$E(\phi_{jh} \mid \beta_h) = g(x_j^T \beta_h), \quad \beta_h = (\beta_{1h}, \ldots, \beta_{qh})^T, \tag{3}$$

where $x_j = (x_{j1}, \ldots, x_{jq})^T$ denotes the jth row vector of x, and β_h are coefficients controlling the impact of the meta covariates on shrinkage of the factor loadings in the hth column of Λ.
Schiavon et al. (2022) focus, as default choice, on the following specification:

$$\begin{aligned} &\tau_0 = 1, \quad \gamma_h = \vartheta_h \rho_h, \quad \phi_{jh} \mid \beta_h \sim \text{Ber}\{\text{logit}^{-1}(x_j^{\text{T}} \beta_h) \, c_p\}, \\ &\vartheta_h^{-1} \sim \text{Ga}(a_\theta, b_\theta), \quad a_\theta > 1, \quad \rho_h = \text{Ber}\,(1 - \pi_h)\,, \quad \beta_h \sim N_q(0, \sigma_\beta^2 I_q), \end{aligned} \tag{4}$$

and assume for $\pi_h = \text{pr}(\gamma_h = 0)$ the recent cumulative stick-breaking process of Legramanti et al. (2020), i.e.

$$\pi_h = \sum_{l=1}^{h} w_l, \quad w_l = v_l \prod_{m=1}^{l-1} (1 - v_m), \quad v_m \sim \text{Be}(1, \alpha),$$

with $\text{Be}(a, b)$ indicating the beta distribution with mean $a/(a+b)$, such that $\pi_{h+1} > \pi_h$ is guaranteed for any $h = 1, \ldots, \infty$ and $\lim_{h \to \infty} \pi_h = 1$ almost surely. The prior specification is completed assuming $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with $\sigma_j^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$ for $j = 1, \ldots, p$.

2.2 Application to Bird Species Occurence Dataset

Schiavon et al. (2022) illustrate the performance of such structured sparsity prior in an ecological application. The data consist of an $n \times p$ binary matrix denoting occurrence of $p = 50$ species in $n = 137$ sampling areas in Finland. The model described in previsous section is generalized for these binary data through a multivariate probit regression model and specialized for the specific application as follows:

$$y_{ij} = \mathbb{I}(z_{ij} > 0), \quad z_{ij} = w_i^T \mu_j + \epsilon_{ij}, \quad \epsilon_i \sim N_p(0, \Lambda\Lambda^T + I_p), \tag{5}$$

where w_i^T is the ith row of a matrix of location-specific covariate matrix w, $z_{ij} \in \mathbb{R}$ is a latent continuous variable underlying y_{ij}, and the residual vector ϵ_i is modelled as multivariate normal, with dependence across species characterized via the factor analytic term $\Lambda\Lambda^T$. The coefficients μ_j characterize impact of the environmental covariates w on the species occurrence probabilities. Meta covariate matrix is available and includes species traits as mean body mass, migratory strategy, and a 7-level superfamily indicator. Such exogenous information is used to model Λ as discussed in previous section.

To appreciate the informative value of the induced sparsity patterns, Figure 1 reports the estimated Λ and meta covariate coefficients β obtained by Schiavon et al. (2022). For details on the point estimation procedure refer to Section 3.3 therein.

The loadings matrix is quite sparse, indicating that each latent factor impacts a small group of species. Positive sign of the loadings means that high levels of the corresponding factors increase the probability of observing birds from those species. Lower elements of β, represented with light cells on the right panel, induce higher shrinkage on the corresponding group of birds. To facilitate interpretation, the rows of Λ are ordered according to the most relevant species traits in terms of shrinkage, which are migration strategy and body mass. The first factor impacts mainly the species characterized by short distance or resident migratory strategies and a larger body mass. The strongly negative value of β suggests heavier species of birds tend to have loadings close to zero for the second factor. This is also true for the third factor, which also does not impact short-distance migrants.

3 Clustering

Bayesian clustering is typically based on mixture models of the form:

$$y_i \sim f, \quad f(y) = \sum_{h=1}^{k} \pi_h \mathcal{K}(y; \theta_h), \tag{6}$$

where k is the number of components, $\pi = (\pi_1, \ldots, \pi_k)^T$ are probability weights, $\mathcal{K}(y; \theta_h)$ is the density of the data within component h, and the

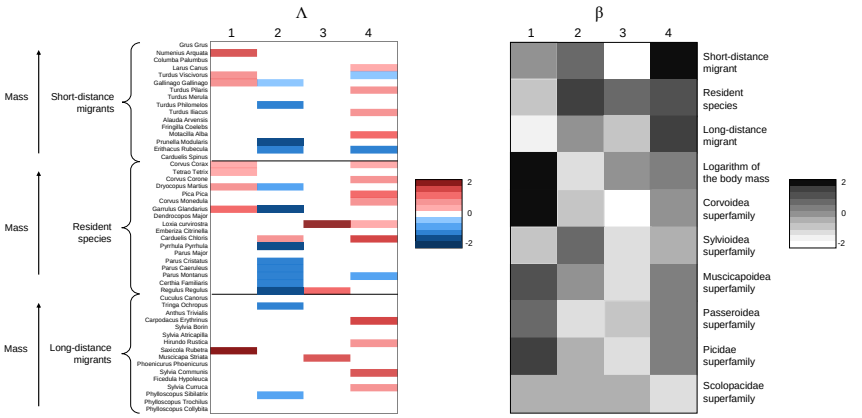


FIGURE 1. Posterior point estimate of Λ and β. The rows of the left matrix refer to the 50 birds species considered; the rows of the right matrix refer to the ten species traits considered. Light coloured cells of β induce shrinkage on corresponding cells of Λ.

number of clusters in data corresponds to the number of occupied components $k_n \leq k$. When p is large and $y_i \in \mathbb{R}^p$, a typical approach chooses $\mathcal{K}(y; \theta_h)$ as a multivariate Gaussian density with a constrained and parsimonious covariance. To avoid to pre-specify k *a priori*, one can exploit a mixture of finite mixture model (Miller and Harrison, 2018; Frühwirth-Schnatter et al., 2021) or let $k = \infty$ as done in Bayesian nonparametrics. When p is very large, however, several problems arise. Celeux et al. (2018) show that the posterior of k_n can concentrate on large values, often the posterior mode of k_n is even equal to n so that each subject is assigned to its own singleton cluster. The authors also conjectured that this aberrant behavior is mainly due to slow mixing of Markov chain Monte Carlo (MCMC) samplers. Chandra et al. (2022), instead, show that this behavior is not just related to the MCMC. They show that for $p \to \infty$ and n fixed the posterior distribution on the space of partitions, under different assumptions on the kernel and regardless of the true data generating model, degenerates to two degenerate clustering. In particular, depending on the choice of kernel and base measure, the posterior may assign probability one to either $k_n = 1$ or $k_n = n$. For a formal description of the problem see Theorem 1 in Chandra et al. (2022).

3.1 Latent Factor Mixture for Bayesian Clustering

To overcome these problems Chandra et al. (2022) propose a general class of *LA*tent factor *M*ixture models for *B*ayesian clustering (Lamb) defined as

$$y_i \sim f(y_i; \eta_i, \psi), \quad \eta_i \sim \sum_{h=1}^{\infty} \pi_h \mathcal{K}(\eta_i; \theta_h), \tag{7}$$

where $\eta_i = (\eta_{i1}, \ldots, \eta_{id})^T$ are d-dimensional latent variables, $f(\cdot; \eta_i, \psi)$ is the density of the observed data conditional on the latent variables and measurement parameters ψ and $\mathcal{K}(\cdot; \theta)$ is a d-dimensional kernel density. As default example, consider

$$y_i \sim N_p(\Lambda\eta_i, \Sigma), \quad \eta_i \sim \sum_{h=1}^{\infty} \pi_h N_d(\mu_h, \Delta_h), \tag{8}$$

where $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is a $p \times p$ diagonal matrix, and Λ is a $p \times d$ matrix of factor loadings. The key idea is to incorporate all the cluster-specific parameters at the latent data level instead of the observed data level to favor parsimony. The proposed solution have some similarities with the models presented by Galimberti et al. (2009), Baek et al. (2010), and Montanari and Viroli (2010). These contributions, starting from different motivations, proposed similar latent factor mixture models as (8) albeit with additional constraints.

Notably, the curse of dimensionality described in Theorem 1 of Chandra et al. (2022) is not present in this model specification. Indeed d here is fixed and the clustering is performed on this d dimensional latent space. Thus, also considering the limit for $p \to \infty$, the posterior distribution on the clustering is not affected by the aberrant behavior described in Theorem 1. Moreover, Chandra et al. (2022) introduce the notion of *oracle clustering* that is the probability distribution on the space of partition that one should obtain with a perfect knowledge of the latent factor η in (8). Under the Lamb model the limit $p \to \infty$ turns to be a bless of dimensionality as discussed in Theorem 2 of Chandra et al. (2022).

3.2 Application to ScRNASeq Cell Line Dataset

Chandra et al. (2022) illustrate the performance of Lamb in a genomic application. Specifically they analyze a single cell RNA-seq dataset containing 630 cells from 7 cell lines and p = 7,666 genes. The known cell types are used as benchmark to assess performance in clustering.

Figure 2 reports the obtained clustering in a UMAP projection (McInnes et al., 2018). The proposed Lamb achieves an adjusted Rand index of 0.977 with a 95% credible interval equal to $[0.900, 0.985]$. The posterior probability of having between 11 and 13 clusters is 0.98. This suggests that the posterior distribution is highly concentrated, which is consistent with the simulations presented in Section 5 of Chandra et al. (2022).

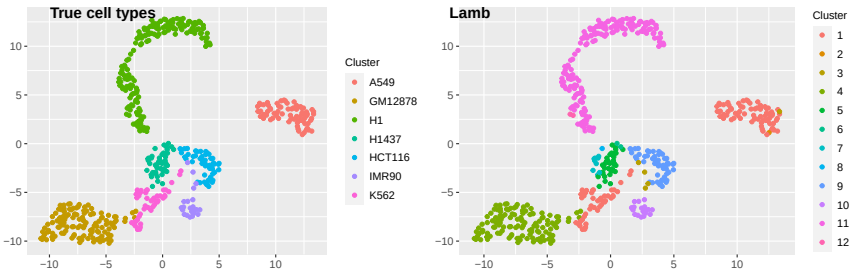


FIGURE 2. UMAP plots of the cell line dataset: Clusterings corresponding to the true cell-types and Lamb estimate

4 Multivariate Regression

In multivariate regression, we model the dependence of a r-dimensional response y from a p-dimensional vector of covariates X. Specifically

$$y_i = \mu + \beta X_i + \varepsilon_i, \quad i = 1, \dots, n, \tag{9}$$

where clearly β is a $r \times p$ matrix of regression coefficients and ε is a random noise with variance Σ. In modern applications both r and p may be large. Envelope models (Cook et al., 2010; Cook and Zhang, 2015) are a class of models that express β as a product of two matrices of smaller dimensions, consistently with what is done in factorization models as discussed in the previous section.
The rationale behind envelope models is that not all linear combinations of responses are influenced by all predictors. This intuition is formalized assuming that there exist two matrices Γ and Γ_0 such that $O = [\Gamma, \Gamma_0]$ is orthogonal and

1. $\Gamma_0^T y | X \sim \Gamma_0^T y$
2. $\Gamma^T y \perp \Gamma_0^T y | X$

These conditions induce that $\text{span}(\beta) \subseteq \text{span}(\Gamma)$ and $\Sigma = \Sigma_1 + \Sigma_2 = P_\Gamma \Sigma P_\Gamma + Q_\Gamma \Sigma Q_\Gamma$, where $P_{(\cdot)}$ is the orthogonal projector operation on a space and $Q_{(\cdot)} = I - P_{(\cdot)}$ is the projection on the orthogonal space. These nicely imply that $\beta = \Gamma\eta$ where Γ is $r \times u$ and η is $u \times p$.
This modelling approach proves to be effective in a variety of situations, but suffers from a major drawback. Specifically, it is assumed that this conjecture is the same for each observation in the sample whilst we may observe a sample made of different subpopulations. For these reasons Mascaretti and Canale (2022) propose a Bayesian nonparametric extension of envelope models through nonparametric mixture models.

4.1 Nonparametric Mixture of Envelops

Instead of assuming

$$y_i \sim N(\mu + \Gamma\eta X_i, \Gamma\Omega\Gamma^T + \Gamma_0, \Omega_0\Gamma_0),$$

it is trivial to extend the successful Dirichlet process (DP) (Ferguson, 1973) and dependent Dirichlet process (DDP) (MacEachern, 2000; Quintana et al., 2022) mixture framework to envelope models. Specifically assume that y_i has density f and f has the following mixture specification

$$f\left(y|X\right) = \sum_{h=1}^{+\infty} \omega_h \mathcal{N}\left(y; \mu_h + \Gamma_h\eta_h X, \Gamma_h\Omega_h\Gamma_h^T + \Gamma_{0h}, \Omega_{0h}\Gamma_{0h}^T\right).$$

Each mixture component is parametrized by $\theta_h = (\mu_h, \eta_h, \Gamma_h, \Gamma_{0h}, \Omega_h, \Omega_{0h})$. Mascaretti and Canale (2022) adopt a modification of the prior proposed by Khare et al. (2017) as base measure for θ_h and standard stick-breaking prior for the sequence of weights.
In a brief simulation study the authors show a consistent estimation of different groups and related parameters. See Mascaretti and Canale (2022) for details.

5 Discusssion

Dimensionality reduction techniques in modern statistics are of paramount practical importance. While in this paper I discussed how this idea turns out to be successful in three different settings, many open problems deserve further methodological improvements or theoretical investigations. I conclude this paper with a short discussion on the importance of having data-driven criteria to choose the size of the latent space, limiting the attention to the three cases discussed here.
The factor model presented in Section 2 exploits a shrinkage prior, thus leading to a coherent Bayesian selection of the latent dimension k. The same does not hold for the two constructions discussed in Sections 3-4. The assumption $d \ll p$ in the latent factor model for Bayesian clustering of Section 3 is crucial in solving the curse of dimensionality. Chandra et al. (2022) consider d fixed. A conservative choice for d is sufficient to solve the theoretical pitfall of Bayesian model-based clustering but the model could be improved by studying a data-driven method to choose d. Similarly, in the Bayesian mixture of envelope models of Section 4, the dimension u is fixed consistently with the frequentist literature that either predetermine or select it by means of AIC or BIC scores. Such procedures clearly hinder uncertainty quantification and a fully Bayesian method to choose u, is currently under investigation.

Acknowledgments: I want to thank the IWSM2022 organizers and all the collaborators involved in the works reviewed in this short paper: Lorenzo Schiavon, David Dunson, Noirrit K. Chandra, and Andrea Mascaretti.

References

Baek, J., McLachlan, G. J., and Flack, L. K. (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**, 1298 – 1309.

Bellman, R. E. (1961). *Adaptive control processes: a guided tour*. Princeton University Press.

Bhattacharya, A., and Dunson, D.B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98**, 291 – 306.

Celeux, G., Kamary, K., Malsiner-Walli, G., Marin, J.-M., and Robert, C. P. (2018). Computational solutions for Bayesian inference in mixture models. In S. Frühwirth-Schnatter, G. Celeux, and C. P. Robert, editors, *Handbook of Mixture Analysis*, chapter 5, pages 77–100. CRC Press.

Chandra, N.K., Canale, A., Dunson, D.B. (2022). Escaping the curse of dimensionality in Bayesian model based clustering *arXiv*, 2006.02700.

Cook, R.D., Bing, L., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, **60**, 927 – 960.

Cook, R.D., and Zhang, X. (2015). Foundations for Envelope Models and Methods. *Journal of the American Statistical Association*, **110**, 599 – 611.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209 – 230.

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2020). Dynamic mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, **16**, 4, 1279 – 1307

Galimberti, G., Montanari, A., and Viroli, C. (2009). Penalized factor mixture analysis for variable selection in clustered data. *Computational Statistics & Data Analysis*, **53**, 4301 – 4310.

Khare, K., Pal, S., and Su, Z. (2017). A Bayesian approach for envelope models. *The Annals of Statistics*, **45**, 196 – 222.

Legramanti, S., Durante, D., and Dunson, D.B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika* **107**, 745 – 752.

MacEachern, S.N. (2000). Dependent Dirichlet processes. Technical report. Department of Statistics, The Ohio State University

Mascaretti, A. and Canale, A. (2022). Bayesian Mixtures of Envelope Models. *Proceedings of the 36th International Workshop on Statistical Modelling*, Trieste, Italy, 18-22 July 2022.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*,**3**, 861.

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the numbe of components. *Journal of the American Statistical Association*, **113**, 340 – 356.

Montanari, A. and Viroli, C. (2010). Heteroscedastic factor mixture analysis. *Statistical Modelling*, **10**, 441 – 460.

Quintana, F.A., Müller, P., Jara, A., MacEachern, S.N. (2022). The Dependent Dirichlet Process and Related Models. *Statistical Science*, **37**, 24 – 41.

Schiavon, L., Canale, A., and Dunson, D.B. (2022). Generalized infinite factorization models. *Biometrika*, in press.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B* **68**, 49 – 67.

Vine copula based stress testing

Claudia Czado[1],

[1] Department of Mathematics and Munich Data Science Institute, Technical University of Munich, Garching, Germany

E-mail for correspondence: cczado@ma.tum.de

Abstract: In this paper we show how copulas can be applied to construct stress tests. In particular we study three different types of stress senarios of increasing complexities. The use of copulas allows for easy quantification of stress levels by setting extreme levels for the copula values of the stressor institutions/instruments. We utilize the class of multivariate vine copulas, which are able to allow for symmetric and asymmetric (tail) behavior of different pairs of variables in a single model. We show that applying the D-vine regression approach of Kraus and Czado (2017) allows for a non simulation based assessment in one scenario type, while simulation is needed for the other two scenario types.

Keywords: Dependence; Vine Copula; Stress testing.

1 Introduction

Stress tests are commonly applied in finance. For a survey and recent developments see Pliszka (2021). They are designed to determine the ability of a financial instrument or institution to deal with economic crisis situations. This is often facilitated by building statistical models in which appropriate crisis senarios can be analyzed. For this we denote by $X_1, \ldots, X_d$ the institutions to be stressed and $S_1, \ldots, S_m$ the instruments/institutions to be used as stressor. We will consider the following situations:

- **Senario 1:** Effect of multiple stressors on a single institution

$$X_j | S_1 = s_1, \ldots S_m = s_m \text{ for } j = 1, \ldots d$$

- **Senario 2:** Effect of single stressor on a set of institutions

$$(X_1, \ldots, X_d) | S_k = s_k \text{ for } k = 1, \ldots m$$

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

- **Senario 3:** Effect of multiple stressors on a set of institutions

$$(X_1, \dots X_d)|S_1 = s_1, \dots S_m = s_m \text{ for } j = 1, \dots d, k = 1, \dots m$$

The different scenarios are illustrated in Figure 1.

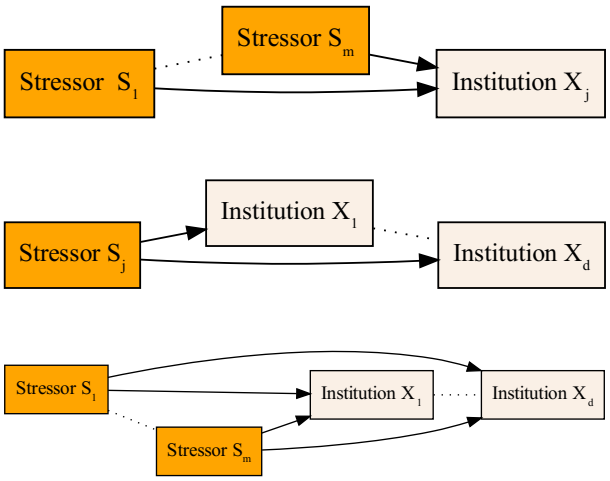


FIGURE 1. Stress test scenarios studied (top: Senario 1, middle: Senario 2, bottom: Senario 3)

For the analysis of these scenarios we need to choose a multivariate statistical model to model the dependence among the components. It should allow for estimation and simulation. Further we need a metric on which to quantify the stress effects and a choice of an appropriate level for the stressor(s).
To model the dependence we follow in general a copula based approach (Sklar, 1959) and in particular a vine copula based approach. The class of multivariate vine copulas is much more flexible than the class of elliptical copulas such as the Gaussian or Student t copulas or the class of Archimedean copulas. It can accommodate in one model symmetry and asymmetry and dependence in the extremes measured by tail dependence. Multivariate vine copulas are copulas built out of bivariate copulas called pair copulas. A pair copula construction (PCC) is possible through conditioning. Joe (1996) gave a first example. In particular it allows for separate specification of the marginal distributions and a set of unconditional and conditional pair copula families and their parameters. Many PCC's are feasible. Bedford and Cooke (2002) introduced a graphical structure of linked trees, called the vine tree structure, to organize them. Gaussian vines were analyzed in Kurowicka and Cooke (2006), while maximum likelihood and sequential estimation for Non Gaussian ones started with aasczado.

More precisely, the joint vine density in d dimensions is the product of marginal densities together with a product of $\binom{d}{2}$ pair copulas identified by the vine tree structure. Here the vine tree structure consists of d trees, where the nodes in the next tree are specified by the edges in the tree before. Edges are allowed when the edges in the previous tree are linked by a single node of that tree (proximity condition). The first tree has nodes 1 to d. The arguments of the pair copulas can involve conditional distribution functions, which can be determined recursively using only the pair copulas identified by the vine tree structure. A greedy sequential algorithm to select the vine tree structure, its associated pair copula families and parameters was developed by Dißmann et al. (2013). The most general class are called regular vines (R-vines), while if the all trees in the vine tree sequence are path trees we speak of D-vines. The associated vine tree structure is completely determined by the order of nodes in the first path tree. The class of C-vines have star form for all trees and are determined by the specification of the centers of all star trees.
A step by step introduction to vine copulas can be found in Czado (2019) and more general results in Joe (2014). See also `vine-copula.org` and Czado and Nagler (2022) for software packages and recent developments.

2 Vine copula based modeling

To facilitate a vine copula based inference we work with three scales: the original scale, the copula scale achieved by applying the marginal distribution functions to each component and normalized margin scale, where the copula data is transformed for each component to achieve standard normal margins. The resulting copula data has then the marginal effects removed. The normalized margin scale can be used to construct pairwise contour plots. If the shape is different than an elliptical shape then a Gaussian pair copula is not warranted.
Estimation is done in a two step approach. First the margins are estimated and then the estimated marginal distribution function is used to form the (pseudo) copula data. The copula data is then used in a second step to select and estimate the vine copula structure (Dißmann et al., 2013). For this way of proceeding we need at least i.i.d. data samples.
For multivariate time series data we need therefore a filtering step to remove the serial dependence. In particular appropriate univariate time series models to each component of the time series are fitted and used to form standardized residuals, which are then approximately i.i.d. The fitted innovation distribution function is then used to transform the standardized residuals to the copula scale. This copula data is then ultilized to assess the dependence among the components. This filtering step is illustrated in Figure 2.
For some of the stress scenarios we propose a simulation based approach

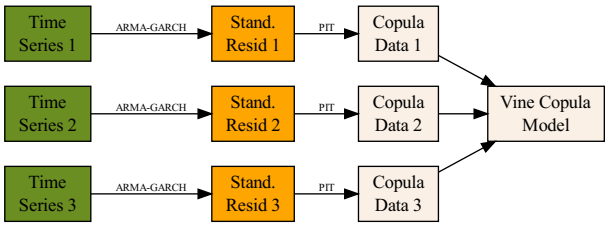


FIGURE 2. Filtering to remove serial dependence for multivariate time series

based on fitted vine copulas. For this we need vine simulation algorithms as discussed for example in Chapter 6 of Czado (2019).

3 Vine based stress testing in Senario 1

In this senario we assume that we have time series log returns $(s_{t1}, \ldots, s_{tm})$ for m stressors $S_{t1}, \ldots, S_{tm}$ at time $t = 1, \ldots, T$ available. The serial dependence of each components is removed by forming standardized residuals $(r^S_{t1}, \ldots, r^S_{tm})$. The standardized residual values r^S_{tk} are an approximate i.i.d sample for the random variable R^S_k with distribution given by the innovation distribution F^S_k for $k = 1, \ldots, m$. This gives rise to corresponding random copula variable $U^S_k = F^S_k(R^S_k)$. To identify a stress level for the k th stressor we set the copula variable U^S_k to a high value such as $U^S_k = u^S_k = .95$ for $k = 1, \ldots m$. Similarly we define the random standardized variable $V_j = G_j(R_j)$, where G_j is the innovation distribution for $X_{tj}, t = 1, \ldots, T$ representing the log returns of the institution to be stressed. To see the effect of the stress levels $u^S_k = .95$ for $k = 1, \ldots m$ for the stressors on the jth institution we need the conditional distribution of V_j given $(U^S_1 = .95, \ldots, U^S_m = .95)$. This conditional distribution is modeled by the D-vine regression model of Kraus and Czado (2017). Using this approach the conditional distribution function of V_j given $(U^S_1 = .95, \ldots, U^S_m = .95)$ and its associated conditional quantile function is analytically available without the need of integration and can be simply evaluated. In Kraus and Czado (2017) the joint distribution of $(V_j, U^S_1, \ldots, U^S_m)$ is modeled as a D-vine copula, where V_j is the first node and the order of the variables $U^S_1, \ldots, U^S_m$ are chosen sequentially in a forward manner improving the conditional copula likelihood until adding a further copula stressor U^S_k does not increase the conditional log likelihood. This approach was applied to 38 senior CDS spreads of 18 banks and 20 (re)-insurers using data from Jan. 4 until Oct. 25, 2011 in Kraus and Czado (2017). Here eight systemic EU banks served as stressor institutions and the effect on each of the remaining institutions were considered separately. On average only 3.7 of the eight stressor banks were selected and Deutsche Bank was selected most often. Most of the selected pair copulas were Stu-

dent t copulas with high association and low degree of freedom. For more details on the results consult Kraus and Czado (2017).

4 Vine based stress testing in Senario 2

In this scenario the impact of a large copula data value of single stressor S on the copula values of institutions $(X_1, \dots X_d)$ is jointly studied. For C-vines the joint distribution given a single conditioning value is analytically available, but marginal quantiles cannot easily computed. Therefore simulation in contrast to the situation of Scenario 1 is needed.
So for copula stressor variable U^S set $u^S = .95$ and sample from

$$(V_1, \dots, V_d)|U^S = u^S.$$

Here V_j is the copula variable corresponding to the j th institution to be stressed. Brechmann et al. (2013) develop the necessary conditional sampling algorithms and use the same data as in Kraus and Czado (2017). In particular the sampling is repeated $R = 10000$ times for each company giving $\widetilde{v}_{\ell,j|u^S},\ j \in \{1, ..., 38\} \setminus u^S,\ \ell = 1, ..., R$. These simulated values are then used to assess the impact of the stressor on a group of institutions. Here the effect on groups formed by sectors (banks/insurances) and regions (EU, US and Asia/Pacific) are investigated. Their major results were

- The stress effect of a member on the other members of the sector highest compared to other sectors in US and EU.
- The stress effect of US banks on US insurance, EU banks and EU insurance is similar.
- The stress effect of EU banks on EU insurances is higher compared to the one on US companies.
- Stress of EU insurance on EU banks higher than on US companies
- Stress of US insurance on US banks and EU companies similar.

More details can be found in Brechmann et al. (2013).

5 Vine based stress testing in Scenario 3

This is the most complex scenario. In this case vine copula models can still be used, however only certain conditional distribution of the chosen vine copula are explicitly available without the need of integration. Therefore Markov Chain Monte Carlo (MCMC) are needed to facilitate the necessary integration. Here also the full class of R-vines can used. This approach has been applied in Kähm (2014) to the same data set as discussed before and his simulation results showed good performance in 10 dimensions.

6 Conclusions and outlook

We have shown how vine copula based modelling is used to evaluate three different types of stress scenarios. This approach allows for simple quantification of the stress level using the copula scale. Especially the Scenario 1 can is computationally cheap since it does not require simulation and thus can be applied involving large numbers of stressors.
Recently vine copulas have also been applied to construct stress tests for portfolios in Sommer (2022). Here Value at risk (VaR) and expected shortfall (EF) are used as risk measures for the portfolio. It utilizes also the approach of Kraus and Czado (2017), but assesses the effect of a single stressor on the VaR or EF of the portfolio using one day ahead forecasts. In this future we like to conduct larger case studies and to develop fast software tools to allow for larger number of stressors and institutions to be stressed.

References

Bedford, T. and R. M. Cooke (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics 30*(4), 1031–1068.

Brechmann, E., K. Hendrich, and C. Czado (2013). Conditional copula simulation for systemic risk stress testing. *Insurance: Mathematics and Economics 53(3)*, 829–839.

Czado, C. (2019). *Analyzing dependent data with vine copulas*. Lecture Notes in Statistics, Springer.

Czado, C. and T. Nagler (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application 9*, 453–477.

Dißmann, J., E. Brechmann, C. Czado, and D. Kurowicka (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics and Data Analysis 52*(1), 52–59.

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. In L. Rüschendorf and B. Schweizer and M. D. Taylor (Ed.), *Distributions with Fixed Marginals and Related Topics*.

Joe, H. (2014). *Dependence modeling with copulas*. London: Chapman &Hall/ CRC.

Kraus, D. and C. Czado (2017). D-vine copula based quantile regression. *Computational Statistics & Data Analysis 110*, 1–18.

Kurowicka, D. and R. Cooke (2006). *Uncertainty analysis with high dimensional dependence modelling*. Chichester: Wiley.

Kähm, O. (2014, Aug). Assessing system relevance of financial institutions using pair-copula constructions for modeling. Master thesis, Technische Universität München.

Pliszka, K. (2021). System-wide and banks' internal stress tests: Regulatory requirements and literature review. *Deutsche Bundesbank Discussion Paper 19*, 1–41.

Sklar, A. (1959). Fonctions dé repartition á n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris 8*, 229–231.

Sommer, P. E. M. (2022, Apr). Estimation and backtesting of the expected shortfall and value at risk using vine copulas - an unconditional and conditional rolling window approach. Masterarbeit, Technische Universität München, Garching b. München.

How long could a human live?

Anthony Davison[1]

[1] Institute of Mathematics, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

E-mail for correspondence: Anthony.Davison@epfl.ch

Abstract: There is long-standing and widespread interest in understanding if there is any limit to human lifetimes. Apart from its intrinsic interest, changes in survival in old age have implications for the sustainability of social security systems. Recent analyses of data on the oldest human lifespans have led to competing claims about survival and to some controversy, due in part to inappropriate use of statistical methods. One central question is whether the endpoint of the underlying lifetime distribution is finite. This talk will discuss the particularities associated with such data, outline correct ways of handling them and present suitable models and methods for their analysis. We illustrate the ideas through analysis of data on semi-supercentenarian lifetimes, which suggests that any upper limit to human lifetimes lies well beyond the highest lifetime yet reliably recorded, with lower limits to 95% confidence intervals around 130 years, and maximum likelihood estimates well above 130 years.

Keywords: Human lifespan; sampling frame; statistics of extremes; survival data

1 Introduction

The existence or not of an upper limit to human lifetimes is of perennial public and scientific interest. Most people hope for a long and happy life, capped by an old age crowned with wisdom and free of worries. The roles of diet, exercise and social contact in achieving these are regularly trumpeted in the media, and societies worldwide are adapting to the presence of a larger active elderly population. Medical and social advances, as well as general population growth, have increased the numbers of centenarians (people aged over 100 years), of semi-supercentenarians (who die between the ages of 105 and 110 years) and supercentenarians (who live to at least 110 years). The oldest person for whom reliable documentation is available, the Frenchwoman Jeanne Calment, died in 1997 aged 122 years and 164

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

days, and despite the increase in the number of well-documented centenarians she remains the only person who can plausibly be said to have lived past her 120th birthday. For some this suggests that there must be a hard upper bound on human lifetimes, and the general increase in health simply means that more and more people will live well closer and closer to this limit, but that it cannot be breached without a major biological breakthrough. Others attribute this event to the play of chance and are willing to invest large sums in attempting to 'stop ageing'. For example, the Palo Alto Longevity Prize aims to reward innovations that could restore the body's homeostatic capacity and thus extend human lifetimes.
The talk summarised by this paper discusses the statistical evidence on this topic, which is plagued by three main issues: data, models, and extrapolation. It is based on work performed jointly with Léo Belzile (HEC Montréal) Jutta Gampe (Max Planck Institute for Demographic Research, Rostock), Holger Rootzén and Dmitrii Zholud (Chalmers University of Technology and University of Gothenburg), detailed in Belzile et al. (2021, 2022).

2 Data

Much of the data available on individuals dying at extreme ages is essentially anecdotal. Sites such as that of the Gerontology Research Group (`grg.org`) or `gerontology.fandom.com/` contain information about many individuals, but there appears to be no well-defined sampling scheme for finding these persons, and this makes statistical analysis impossible. Moreover the kudos given to the very old can give incentives to exaggerate longevity, so it is important to check ages carefully, especially as any extrapolation towards a possible upper bound is likely to depend strongly on the few largest lifetimes. Those who live to great ages were by definition born long ago, and validating their life-course (using official birth, baptismal, marriage and similar records) can be difficult and time-consuming even in countries with long-standing and reliable systems of public records. Fortunately a systematic effort to assemble and validate data from several countries has led to the creation of the International Database on Longevity (IDL, `www.supercentenarians.org`), which has been set up by demographers of various countries to help shed light on living to great age. The data are regularly updated (so some care is needed in making comparisons between analyses at different times) and freely available. At the time of writing the IDL contains 1161 supercentenarians and around 18,000 semi-supercentenarians; the former all systematically validated using national records, and the latter validated by sampling. An important part of the database is the meta-data, which explains the sampling schemes used; these have important implications for subsequent inferences.
The sampling scheme used for the IDL is to retain all the individuals dying above a given age threshold (110 years for supercentenarians) between two

given dates. This implies that the sample is truncated, because individuals who do not reach the threshold or who do, but die outside the given dates, do not appear in the database. A further complication is that the increasing numbers of the very old means that the rate of entry into the sample is rapidly changing over time. The effects of truncation and of the increased population size have been ignored by some authors, leading to biased estimation of the lifetime distribution.
Other large databases are also available, though they are typically unvalidated and may have slightly different sampling schemes.

3 Modelling and extrapolation

The goal of analysis is inference on the putative upper limit to the lifetime distribution. We call this limit the 'human lifespan' and refer to individual life-lengths as 'lifetimes'. The fact that an infinite lifespan may be compatible with the finiteness of every individual lifetime is not obvious to those with a shaky grasp of probability theory, and some care is needed when presenting results.
The distribution of human lifetimes is of long-standing actuarial interest and a variety of models have been proposed for the hazard function, also known in the present context as the 'force of mortality'. The Gompertz–Makeham hazard at age t is of the form $\lambda_0+\lambda_1 \exp(t/\sigma)$, where $\lambda_0, \lambda_1\sigma > 0$, and can be interpreted as stemming from a constant hazard of accidents at any age plus an exponential hazard due to aging. Although this hazard function increases without limit, the corresponding lifespan is infinite; again a common misconception is that an unlimited hazard must lead to a finite lifespan. Such a distribution cannot resolve whether the lifespan is finite.
A more satisfactory approach is the use of statistical extreme-value theory and in particular the generalized Pareto distribution, which can be regarded as the canonical model for exceedances over a high threshold u, say. This distribution has hazard function $\max(\sigma + \xi t, 0)$ for $t > 0$, positive scale parameter σ, and real shape parameter ξ. If the latter is negative, then the lifespan ι equals $u-\sigma/\xi$, and if $\xi \geq 0$, then ι is infinite. There are subtleties in the application of this model, which provides an asymptotic approximate to the exceedance distribution, valid as $u \to \iota$, and the stability of its fit should be carefully assessed. Fortunately this domain of statistics is now well-developed and numerous techniques are available for data analysis that takes into account truncation and censoring if necessary.

4 Results

Figure 1 summarises the evidence for the human lifespan based on suitable fits of the generalized Pareto distribution to various datasets. With the exception of French men, whose hazard of dying is significantly higher than

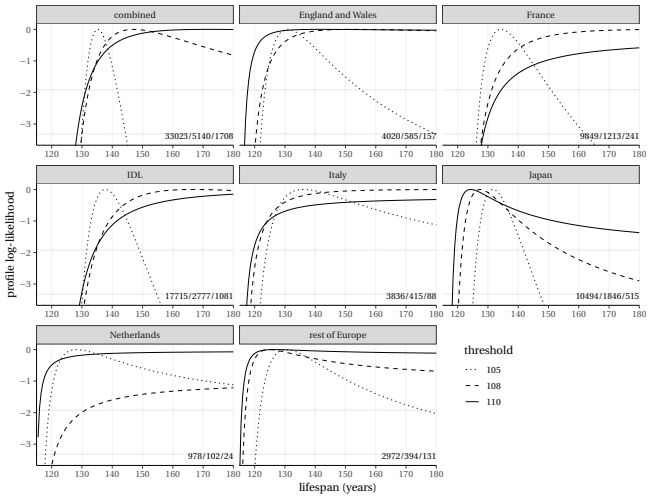


FIGURE 1. Profile log likelihoods for the human lifespan based on various datasets using the generalized Pareto model with thresholds at 105, 108 and 110 years. The numbers of exceedances of the thresholds are given at the bottom right of each panel. Adapted from Belzile et al. (2022).

that of their female counterparts, there seem to be no national or gender differences (though women outnumber mean 10:1 at great age). The general summary in the top left panel shows that the maximum likelihood estimate for the lifespan is around 135 for threshold 105 years with approximate 95% confidence interval $(130, 140)$, but as the threshold is increased the maximum likelihood estimate increases; the upper limit to the confidence interval is infinte for higher thresholds. The simplest model above 108 years is exponential and corresponds to a probability of dying in any given year of 0.5, conditional on survival to that point. Similar patterns appear for most of the other datasets, and we conclude that the statistical evidence suggests that even if the human lifespan is not infinite, it is unlikely to be approached in the near future without a major medical advance — under this model, even if a million people reached age 110, we would expect just one to reach the age of 130.

References

Belzile, L., Davison, A. C., Rootzén, H. and Zholud, D. (2021). Human mortality at extreme age. *Royal Society Open Science*, **8**: 202097.

Belzile, L., Davison, A. C., Gampe, J., Rootzén, H. and Zholud, D. (2022). Is there a cap on longevity? A statistical review. *Annual Review of Statistics and its Application*, **9**, 21–45.

Is blood pressure variability a risk factor for cerebro-vascular event? Joint modelling approaches

Jacqmin-Gadda Hélène[1], Courcoul Léonie[1], Barbieri Antoine[1], De Courson Hugues[1], Tzourio Christophe[1]

[1] Bordeaux Population Health Inserm Research Center, Univ. Bordeaux, France

E-mail for correspondence: helene.jacqmin-gadda@inserm.fr

Abstract: High blood pressure is a well known risk factor for cardio-vascular events, stroke and dementia. Several studies have also suggested that high variability of blood pressure could be a risk factor for cerebro-vascular events independent from the mean blood pressure. Besides, the medical literature is full of hypotheses regarding a possible link between the variability of risk factors or markers and the onset of clinical events. For instance, the emotional instability could be associated with the risk of psychiatric events while the variability of glycemia could impact the prognosis of diabetes patients.
However, studying the relationship between the variability of a factor and an event risk raises methodological challenges, especially due to the fact that the measure of the variability requires repeated measures over a period of time. Joint modelling of the repeated measures of the marker over time and the event risk is recommended to avoid biases due either to sample selection or sample attrition. In this presentation, we will discuss both joint models including a location-scale mixed submodel and bivariate joint models considering repeated measures of the empirical variance and the mean of blood pressure as two correlated markers.

Keywords: Joint models, Longitudinal data, Risk prediction, Variance modelling.

1 Introduction

Stroke is the leading cause of acquired physical disability in adults and the second leading cause of death. It is thus important to identify risk factors for stroke in order to implement prevention programs. High blood pressure is a well known risk factors for stroke and, more recently, a large variability

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

of blood pressure has also been found associated with an increased risk of stroke (Shimbo et al, 2012) and dementia (Alpérovitch et al, 2014) independently of the mean level of blood pressure. However, these studies are based on standard Cox models including the standard deviation of blood pressure as an explanatory variable and thus are exposed to potentially important shortcomings (De Courson et al, 2021). When the standard deviation is estimated using all the measures collected during all the follow-up, biases may arise because measurements after the current time (and in some analyses after the event time) are used to predict the event. When the standard deviation is computed on an initial period of time excluding subjects who had the event during this period, a selection bias and a loss of power are possible. In addition, the measurement error and the unequal number of measures for each individual are not handled with these approaches.

Rigorous statistical approaches would therefore be useful for studying if the variability over time of exposures or biomarkers is associated with the onset of clinical events. For instance, clinicians are interested in the predictive ability of emotional instability for psychiatric events or in the impact the variability of glycemia on the prognosis of diabetes patients. As many clinical and epidemiological studies include frequent repeated measures of exposure or biomakers, often using IoT tools, the data required for such fine analyses are now available.

Joint modelling of longitudinal data and time to events is a key approach to study time-varying risk factors as determinant of health events and to develop dynamic individual prediction models for the events based on repeated measures of markers. They combine a mixed model for the repeated measures of the time-varying variable and a time-to-event model. Functions of the random effects from the mixed model are included as explanatory variables in the time-to-event model to account for the association between the two outcomes. Joint models account for the measurement error of the risk factor, avoid imputation of the last observed risk factor value for all the event times, and are suitable for endogenous variables (variables that can be impacted by the onset of the event at an earlier time such as biological markers). Although most joint models assume that the event risk depends only on the mean individual trajectory of the marker (through the current value or the current slope for instance), Barrett et al (2019) recently proposed a joint model allowing the event risk to depend on the intra-subject variability of the marker.

In this work, we extend the Barrett et al's location-scale joint model to allow more flexible dependence structure between the event and the markers and handle competing risks. We used this new model to investigate the relation between blood pressure variability and the risk of stroke recurrence accounting for the competing risk of death from another cause in a large international clinical trial. Then this approach is applied to a sample of patients hospitalized in an intensive care unit after subarachnoid hemorrhage (SAH) to predict the risk of vasospasm (one of the main complications

of SAH) from intensive blood pressure measures. To explore the change over time in the variability of blood pressure and its impact on the risk of vasospasm, we consider also a bivariate joint model considering repeated measures of the individual mean of the blood pressure and of its individual standard-deviation computed on 6-hours windows as two longitudinal markers.

2 Location-scale joint models

2.1 Model formulation

Let us denote $Y_i = (Y_{i1}, ..., Y_{ij}, ..., Y_{in_i})$ the vector of repeated measures of the blood pressure for subject i $(i = 1, ..., N)$ at times $(t_{ij}, j = 1, ...n_i)$. Assuming two two competing events, the observed event time, denoted T_i, is the minimum between the time of onset of the competing events and the time of censoring while the failure indicator is $\delta_i \in \{0, 1, 2\}$ with $\delta_i = 0$ in case of censoring and $\delta_i = k \in \{1, 2\}$ if event k is observed. The model can include vectors of possibly time-dependent covariates X_{ij} and Z_{ij} in addition to t_{ij}.
The location-scale joint model for competing risks is a shared random-effects joint model including a subject-specific random effect for the residual variance of the blood pressure. The submodel for the longitudinal marker is defined by :

$$Y_{ij} = Y_i(t_{ij}) = \tilde{Y}_i(t_{ij}) + \epsilon_{ij} = X_{ij}^{\mathrm{T}}\beta + Z_{ij}^{\mathrm{T}}b_i + \epsilon_{ij} \tag{1}$$

with

$$b_i \sim \mathcal{N}(0, \Sigma), \epsilon_{ij} \sim \mathcal{N}(0, \sigma_i^2),$$
$$\log(\sigma_i) \sim \mathcal{N}(\mu_\sigma, \tau_\sigma^2) \quad \text{and} \quad b_i \perp \sigma_i.$$

The cause-specific models for the competing events are defined for the event $k \in \{1, 2\}$ by :

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp\left(W_i^{\mathrm{T}}\gamma_k + \alpha_{1k}\tilde{y}_i(t) + \alpha_{2k}\tilde{y}_i'(t) + \alpha_{\sigma k}\sigma_i\right), \tag{2}$$

with $\lambda_{0k}(t)$ the baseline risk functions and W_i a vector of time-fixed covariates, while $\tilde{y}_i(t)$ and $\tilde{y}_i'(t)$ are the expected value and the derivative of the individual blood pressure trajectory at time t. We consider either standard parametric baseline risk function (Weibull for instance) or flexible functions defined using B-splines.

2.2 Estimation method

The model is estimated by maximising the marginal likelihood with numerical integration over the random effects. Taking advantage of the conditional

independence between the marker and the events given the random effects, the individual contribution to the likelihood may be written as :

$$L_i(Y_i, T_i, \delta_i) = \int f(Y_i|b_i, \sigma_i) \exp(-\Lambda_1(T_i|b_i, \sigma_i) - \Lambda_2(T_i|b_i, \sigma_i)) \prod_{k=1}^{2} \lambda_k(T_i|b_i, \sigma_i)^{\mathbb{1}_{(\delta_i=k)}} f(b_i) f(\sigma_i) db_i d\sigma_i \quad (3)$$

where $\Lambda_k(t|b_i, \sigma_i)$ is the conditional cumulative risk for event k and $f(Y_i|b_i, \sigma_i)$ is the product of n_i conditional univariate Gaussian distributions. Delayed entry can be handled by dividing the likelihood by the probability to be free of event at entry in the cohort.
The integral over the random effects are computed by a Quasi Monte Carlo approach and the cumulative risks function are approximated by the Gauss-Konrod quadrature with 15 points. The optimisation algorithm is the Marquardt-Levenberg algorithm, a robust variant of the Newton-Raphson algorithm, with stringent convergence criteria. The estimation algorithm is implemented in a R-package that will be made available on GitHub.

3 Application to blood pressure and cerebro-vascular events

3.1 Blood pressure and recurrence of stroke

This first analysis aims at studying the impact of intra-individual variability of blood pressure on the risk of stroke recurrence and death from other causes in a large sample of subjects with a history of stroke. Data comes from the randomized, double-blind, international controlled trial PROGRESS (Perindopril Protection against Stroke Study). We analysed data from the 3032 patients included in the placebo arm. They were followed for at least 4 years with 5 visits in the first year and 2 annual visits for the next 4.5 years with blood pressure measures collected at each visit. During the follow-up, 406 strokes and 213 deaths without strokes were observed.
We estimated a joint model defined by (1) and (2) adjusting each submodel on age at baseline, sex and ethnicity (Asian or Non-Asian). The baseline risk function and the change over time of the blood pressure were defined using a splines basis.
Estimates supported the hypothesis of an heterogeneous intra-subject variability ($\hat{\tau^2} = 0.36, SE(\hat{\tau^2}) = 0.008$) significantly associated with the risk of death ($\hat{\alpha}_{\sigma 2} = 0.071, SE(\hat{\alpha}_{\sigma 2}) = 0.027$) but not with the risk of recurrent stroke ($\hat{\alpha}_{\sigma 1} = -0.023, SE(\hat{\alpha}_{\sigma 1}) = 0.023$). Conversely, the current value of blood pressure was not associated with the risk of death

($\hat{\alpha}_{12} = 0.006, SE(\hat{\alpha}_{12}) = 0.006$) but high current value of blood pressure significantly increased the risk of recurrent stroke ($\hat{\alpha}_{11} = 0.024, SE(\hat{\alpha}_{11}) = 0.004$). The slope of the blood-pressure trajectory was not associated with the events and was deleted from the model.

3.2 Blood pressure and risk of vasospasm after SAH in ICU

The objective of this analysis was to evaluate the association of blood pressure variability with the risk of vasospasm and death after a SAH. The database included 201 patients hospitalized in the neurologic ICU of the Bordeaux University Hospital for the management of a SAH. Data were collected from June 2018 to June 2019 inclusive. For each patient, blood pressure was measured every hour until discharge or for a maximum of 14 days after entry into the ICU; 46 vasospasms and 10 deaths were observed over 14 days. The longitudinal sub model and the time-to-event submodel were adjusted for age, sex, elapsed time between HSA and entry in ICU, weight, a severity score and history of hypertension before inclusion.
Although the outcomes, the time range and the patients status were completely different, we found qualitatively similar results. A high current value of blood pressure was associated with an increased risk of vasospasm ($\hat{\alpha}_{11} = 0.042, SE(\hat{\alpha}_{11}) = 0.019$) while a large intrasubject variability increased the risk of death($\hat{\alpha}_{\sigma 1} = 0.66, SE(\hat{\alpha}_{\sigma 1}) = 0.21$).
A limit of the above analyses was that the intra-subject variance of blood pressure was assumed to be constant over time while clinicians hypothesize that changes in blood pressure variability could be predictive of vasospasm or death in a near future. As the location-scale joint model does not allow flexible modeling of the time trend of the intra-subject variance, we performed an exploratory analysis of this hypothesis using a standard joint model for two longitudinal markers and competing events. We described the change over time of the empirical individual mean and empirical individual standard-deviation of blood pressure computed on non-overlapping 6-hours time intervals using a bivariate mixed model with flexible time-trend and tested in a joint model if the risk of vasospasm and death was associated with the current mean and standard-deviation values. The model was estimated with the R-package JMBayes2. In the presentation, we will contrast results of the two approaches.

4 Conclusion

The proposed location-scale joint models makes possible to study the variability over time of any exposure or biomarkers as a risk factor for clinical events. This model includes a flexible modelling of the time trend for both the marker and the event risk as well as the dependence structure between the outcomes. A free R-package will be made available soon. Future work includes flexible modeling of time-dependent intra-subject variance.

Acknowledgments: This work was funded by the French National Agency for Research, grant ANR-21-CE36-0013-01 for project JMECR

References

Alpérovitch, A., Blachier, M., Soumaré, A. et al. (2014), Blood pressure variability and risk of dementia in an elderly cohort, the Three-City Study. *Alzheimer's & Dementia*, **10**, S330 – S337.

Barrett, J.K., Huille, R., Parker, R et al. (2019) Estimating the association between blood pressure variability and cardiovascular disease: An application using the ARIC study. *Statistics in Medicine*,**38**,1855 – 1868.

de Courson, H., Ferrer, L., Barbieri, A. et al. (2021) Impact of model choice when studying the relationship between blood pressure variability and risk of stroke recurrence. *Hypertension*,**78**,1520 – 1526.

Shimbo, D., Newman, J. D., Aragaki, A. K.et al. (2012). Association between annual visit-to-visit blood pressure variability and stroke in postmenopausal women: data from the Women's Health Initiative. *Hypertension*, **60**, 625-630.

Improved estimation through additive adjustment of estimating functions

Ioannis Kosmidis[1]

[1] University of Warwick, Coventry, CV4 7AL, United Kingdom

E-mail for correspondence: ioannis.kosmidis@warwick.ac.uk

Abstract: This short paper focuses on an integrated, concise overview of methods that improve mean and median bias through the additive adjustment of estimating functions. A unified theoretical and algorithmic framework for the task is presented, along with recent developments in improving the bias of general M-estimators, and an analysis of the equivariance properties of the improved estimators under parameter transformation.

Keywords: mean bias; median bias; penalized likelihood

1 Adjusted estimating functions

1.1 Maximum likelihood estimation

Let $\ell(\theta)$ be the log-likelihood about a parameter vector θ with $\theta \in \Re^v$. If the model is appropriate, then under fairly general regularity conditions, the maximum likelihood (ML) estimator $\hat{\theta} = \arg\max \ell(\theta)$ has $\hat{\theta} \xrightarrow{p} \theta$, and $\{i(\theta)\}^{1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I_v)$, where I_v is the $v \times v$ identity matrix, and $i(\theta) = \mathrm{Var}_\theta(\nabla\ell(\theta))$ is the expected information matrix. Hence, the ML estimator is consistent and asymptotically has zero mean and median bias, and attains the Cramér-Rao lower bound $i(\theta)$ for the variance of unbiased estimators. More refined results can be obtained using stochastic Taylor expansions, and include that the mean bias is $\mathrm{E}_\theta(\hat{\theta} - \theta) = O(n^{-1})$, where n is a measure of information about θ, often equal to the sample size, the components $\hat{\theta}$ have median bias $P(\hat{\theta}_j \leq \theta_j) = 1/2 + O(N^{-1/2})$, and $\hat{\theta}$ has variance $\mathrm{Var}(\hat{\theta}) = i(\theta) + O(n^{-1})$.
These asymptotically optimal properties have rendered ML as one of the most popular estimation methods in statistical modelling. Nevertheless, for

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

finite samples, there can be substantial deviations from the expected behaviour. There have been numerous attempts to improve the finite-sample properties of the ML estimator. Prominent approaches in this direction define a new estimator $\bar{\theta}$ to be such that $S(\bar{\theta}) + A(\bar{\theta}) = 0_v$, for an adjustment $A(\theta) = O_p(1)$ as $n \to \infty$, where $S(\theta) = \nabla\ell(\theta)$ and 0_v is a v-vector of zeros. The key idea is then to derive the form of $A(\theta)$ that results in estimators with asymptotically smaller mean (Firth, 1993; Kosmidis and Lunardon, 2020) or median bias (Kenne Pagui et al., 2017), in the sense that $\mathrm{E}_\theta(\bar{\theta} - \theta) = o(n^{-1})$ and $P(\hat{\theta}_j \leq \theta_j) = 1/2 + O(N^{-3/2})$, respectively.

1.2 Estimation using a quasi-Newton iteration

A general iterative procedure for computing the estimates $\bar{\theta}$ is a generalization of the quasi-Newton iteration in Kosmidis and Firth (2010), where the value of $\bar{\theta}$ at the $(m-1)$th iteration is updated to

$$\bar{\theta}^{(m)} := \bar{\theta}^{(m-1)} + \{j(\bar{\theta}^{(m-1)})\}^{-1} S(\theta^{(m-1)}) + \{j(\bar{\theta}^{(m-1)})\}^{-1} A(\theta^{(m-1)}) \,.$$

If the iteration converges, it must hold that $S(\bar{\theta}^{(\infty)}) + A(\bar{\theta}^{(\infty)}) = 0_v$, hence the iteration has the correct stationary point. More specialized algorithms can be developed for particular model classes and adjustment functions; see for example Kosmidis and Firth (2021) for an algorithm that proceeds by repeated maximum likelihood fits on adjusted versions of the binomial counts and totals when $A(\theta)$ is the gradient of the Jeffreys invariant prior.

1.3 Inference

The attractiveness of additively adjusted likelihood equations approaches is that $\bar{\theta}$ has the same asymptotic distribution as the ML estimator generally does, and, as a result, is asymptotically efficient. The distribution of $\bar{\theta}$ for finite samples can be approximated by $N(\theta, \{i(\theta)\}^{-1})$. This result is due to the adjustment $A(\theta)$ being of order $O_p(1)$ as $n \to \infty$, and hence, dominated by $\nabla\ell(\theta)$, which is $O_p(n^{1/2})$, as information increases. The implication is that standard errors for $\bar{\theta}$ can be computed exactly as for the ML estimator, using the square roots of the diagonal elements of the inverse of $i(\theta)$ or $j(\theta) = -\nabla\nabla^{\mathrm{T}}\ell(\theta)$ at the estimates. Furthermore, first-order inferences, like standard Wald tests and Wald-type confidence intervals and regions are constructed in a plugin fashion, by replacing the ML estimates with the value of $\bar{\theta}$ in the usual procedures in standard software.

2 Mean bias reduction

Firth (1993) shows that an estimator θ^* with mean bias $E_\theta(\theta^* - \theta) = O(n^{-2})$, which is asymptotically smaller than the bias of $\hat{\theta}$, results for

$$A_t(\theta) = \frac{1}{2}\mathrm{trace}\left[i(\theta)^{-1}\left\{P_t(\theta) + Q_t(\theta)\right\}\right] \quad (t = 1, \ldots, v)\,, \tag{1}$$

where $P_t(\theta) = \mathrm{E}_\theta(S(\theta)S(\theta)^\mathrm{T}S_t(\theta))$ and $Q_t(\theta) = -\mathrm{E}_\theta(j(\theta)S_t(\theta))$. Mean bias reduction (mBR) has been found to result in estimates away from the boundary of the parameter space in a range of categorical data models. See, for example, Kosmidis and Firth (2009, Section 6) for row-column association models; Bull et al. (2002) and Kosmidis et al. (2020, Section 6) for baseline category models; and Kosmidis (2014) for cumulative link models. If θ is the canonical parameter of a full exponential family, like in binomial and multinomial logistic regression, then $j(\theta) = i(\theta)$ and $j(\theta)$ does not depend on the stochastic part of the model. Hence, $Q_t(\theta) = 0_{v\times v}$, where $0_{v\times v}$ is a $v \times v$ matrix of zeros, and some algebra gives that the solution of the mean bias-reducing adjusted score equations is equivalent to the maximization of the penalized log-likelihood

$$\ell(\theta) + \frac{1}{2}\log\det\{i(\theta)\}\,, \tag{2}$$

where the penalty is the logarithm of the Jeffreys invariant prior. Recent work by Kosmidis and Firth (2021) considers the impact of penalized likelihoods like (2) in the estimation of many well-used binomial-response generalized linear models, including logistic, probit, complementary log-log, and cauchit regression. Among other results, Kosmidis and Firth (2021) prove that maximizing the likelihood after penalizing it by arbitrary positive powers of the Jeffreys prior always results in finite estimates, and derive the shrinkage directions implied by the penalty.

Kosmidis and Lunardon (2020) have proposed an alternative and more general bias reduction approach that not only applies to general M-estimation problems (where $S(\theta) = \sum_{i=1}^k S^i(\theta)$ is a general estimating function formed by independent contributions $S^1(\theta), \ldots, S^k(\theta)$), but also does not require computing expectations of products of log-likelihood derivatives (like $P_t(\theta)$, $Q_t(\theta)$, and $i(\theta)$) under the model, which can be a daunting task even for simple models. In particular, it can be shown that a reduced-bias M-estimator (RBM-estimator) θ^{**} with mean bias $E_\theta(\theta^{**} - \theta) = O(n^{-3/2})$ results if

$$A_t(\theta) = -\mathrm{trace}\left\{j(\theta)^{-1}d_t(\theta)\right\} - \frac{1}{2}\mathrm{trace}\left[j(\theta)^{-1}e(\theta)\left\{j(\theta)^{-1}\right\}^\mathrm{T}u_t(\theta)\right]\,, \tag{3}$$

where $u_t(\theta) = \sum_{i=1}^k \nabla\nabla^\mathrm{T}S_t^i(\theta)$, $j(\theta)$ has rth row $-\sum_{i=1}^k \nabla S_r^i(\theta)$, $[e(\theta)]_{rt} = \sum_{i=1}^k S_r^i(\theta)S_t^i(\theta)$, and $[u_r(\theta)]_{st} = \sum_{i=1}^k\{\partial S_r^i(\theta)/\partial\theta_s\}S_t^i(\theta)$.

The bias-reduction method defined by (3) is a major generalization over past bias-reduction methods whose applicability is limited to either cases where the log-likelihood function of a correctly-specified model is required (e.g., Firth, 1993), or samples from a correctly-specified model can be simulated (e.g., Gourieroux, 1993 for indirect inference, and and Kuk, 1995 for iterative bootstrap). Also, importantly, this generalization comes with less implementation requirements because (3) requires only the estimating function contributions and the first two derivatives of those.

RBM-estimators can directly be computed and used in settings that involve realizations of k independent random vectors with dependent components. Examples of such settings are the generalized estimating equations in Liang and Zeger (1986) for estimating marginal regression parameters for correlated responses, and composite likelihood methods Varin et al. (2011). What makes RBM estimation appealing is the fact that if $S(\theta)$ is the gradient of an objective function $\ell(\theta)$, then $S(\theta)+A(\theta)$ is formally the gradient of $\ell(\theta) - \text{trace}\left\{(j(\theta))^{-1}e(\theta)\right\}/2$, and the RB$M$ estimates can be computed by maximizing the latter expression. Hence, unlike other approaches, RBM estimation always has a penalized likelihood interpretation.

3 Median bias reduction

Kenne Pagui et al. (2017) show that an estimator $\theta^\dagger$ with $P(\theta_t^\dagger \leq \theta_t) = 1/2 + O(N^{-3/2})$, which is asymptotically closer to 1/2 than the median bias of $\hat{\theta}$, results for

$$A(\theta) = \frac{1}{2}\text{trace}\left[i(\theta)^{-1}\left\{P_t(\theta) + Q_t(\theta)\right\}\right] - i(\theta)F(\theta)\,. \tag{4}$$

In the above expression, $F_t(\theta) = [i(\theta)^{-1}]_t^{\mathrm{T}}\tilde{F}_t(\theta)$, with

$$\tilde{F}_{tu}(\theta) = \text{trace}\left[\tilde{i}_u(\theta)\left\{\frac{1}{3}P_t(\theta) + \frac{1}{2}Q_t(\theta)\right\}\right] \quad (t = 1, \ldots, g)\,,$$

and $\tilde{i}_u(\theta) = [i(\theta)^{-1}]_u[i(\theta)^{-1}]_u^{\mathrm{T}}/[i(\theta)^{-1}]_{uu}$ $(u = 1, \ldots, v)$, where A_u and A_{tu} denote the uth column and (t,u)th element of a matrix A. When $j(\theta) = i(\theta)$, expression (4) simplifies in a similar manner as expression (1) does. In fact, for one-parameter models ($v = 1$) that are exponential families in canonical parameterization, it can be shown that median bias reduction (mdBR) is formally equivalent to the maximization of $\ell(\theta) + \log\det\{i(\theta)\}/6$ (see, Kenne Pagui et al., 2017, Section 2.1. However, mdBR has no penalized likelihood interpretation for $v > 1$.
Furthermore, to date there has been no empirical adjustment like (3) that delivers mdBR in general M-estimation problems.

4 Bias reduction and parameter transformation

The ML estimator is equivariant in the sense that the ML estimator of $g(\theta)$ is exactly $g(\hat{\theta})$ for any one-to-one transformation $g(\cdot)$. Hence, there is no need to maximize the log-likelihood about $g(\theta)$ if the ML estimator of θ has already been computed. In contrast, the mBR and mdBR estimators are equivariant only for specific transformations $g(\cdot)$.
The mBR estimator is equivariant under linear transformations of the parameters; the mBR estimator of $C\theta$ for a known matrix C is exactly $C\theta^*$.

This is useful in regression problems with categorical covariates where parameter contrasts are usually of interest. The same equivariance property holds for the RBM-estimator but not for the mdBR estimator.

The mdBR estimator of $(g_1(\theta_1), \ldots, g_v(\theta_v))^{\mathrm{T}}$ is $(g_1(\theta_1^\dagger), \ldots, g_v(\theta_v^\dagger))^{\mathrm{T}}$ for any set of one-to-one functions $g_1(\cdot), \ldots, g_v(\cdot)$. Hence, and unlike mBR, the mdBR estimator is equivariant under component-wise transformations.

Di Caterina and Kosmidis (2019) present a simple way to derive the mean bias of $h(\bar{\theta})$ for any three-times differentiable function $h : C \to D$, with $C \subset \Re^p$ and $D \subset \Re$, when $\bar{\theta}$ is an estimator with $o(n^{-1})$ bias. The estimator $h(\bar{\theta})$ of $\zeta = h(\theta)$ has mean bias

$$\mathrm{E}(h(\bar{\theta}) - h(\theta)) = \frac{1}{2}\mathrm{trace}\left\{i(\theta)^{-1}\nabla\nabla^{\mathrm{T}}h(\theta)\right\} + O(n^{-2}), \qquad (5)$$

where $\nabla\nabla^{\mathrm{T}}h(\theta)$ is the hessian of $h(\cdot)$ at θ. Note that for linear transformations, $\nabla\nabla^{\mathrm{T}}h(\theta) = 0_{v\times v}$, and hence $E(h(\bar{\theta}) - h(\theta)) = O(N^{-2})$, which confirms the earlier discussion that the mBR and RBM estimators being exactly equivariant for linear transformations of the parameters. The first term in the right-hand side of (5) can be evaluated at $\bar{\theta}$ and be used to derive reduced-bias estimators based only on $\bar{\theta}$, $i(\bar{\theta})$ or $j(\bar{\theta})$, and $\nabla\nabla^{\mathrm{T}}h(\bar{\theta})$. Obvious such estimators are $h(\bar{\theta}) - \mathrm{trace}\left\{i(\bar{\theta})^{-1}\nabla\nabla^{\mathrm{T}}h(\bar{\theta})\right\}/2$. and $h(\bar{\theta}) - \mathrm{trace}\left\{j(\bar{\theta})^{-1}\nabla\nabla^{\mathrm{T}}h(\bar{\theta})\right\}/2$.

For example, consider the case of estimation of the odds-ratios $\exp(\beta_j)$ in a logistic regression model with linear predictor $\eta_i = x^{\mathrm{T}}\beta$, for a covariate vector x. Expression (5) gives that the odds-ratio at the mBR estimator has $\mathrm{E}(\exp(\beta_j^*)) = \exp(\beta_j)\left[1 + v_{jj}(\theta)/2\right] + O(n^{-2})$, where $v_{jj}(\theta) = [i(\theta)^{-1}]_{jj}$. Hence, two estimators of $\zeta_j = \exp(\beta_j)$ with $O(n^{-2})$ bias are

$$\zeta_j^{(1)} = \exp(\beta_j^*)\left[1 - \frac{1}{2}v_{jj}(\theta^*)\right] \quad \text{and} \quad \zeta_j^{(2)} = \frac{\exp(\beta_j^*)}{1 + v_{jj}(\theta^*)/2},$$

arising from subtracting an estimate of the bias at $\theta := \theta^*$ from $\exp(\beta_j^*)$, and dividing $\exp(\beta_j^*)$ by the correction factor $1 + v_{jj}(\theta^*)/2$, respectively. The estimator $\zeta_j^{(2)}$ for the odds-ratio ζ_j has the advantage of being always positive, while $\zeta_j^{(1)}$ takes negative values if $v_{jj}(\theta^*) > 2$. The approximation $\exp\{v_{jj}(\theta)/2\} \approx 1 + v_{jj}(\theta)/2$ for small $v_{jj}(\theta)$ can be used to show that the mBR estimator $\zeta_j^{(2)}$ closely relates to the reduced-bias estimator $\exp\{\beta_j^* - v_{jj}(\theta^*)/2\}$ derived in Lyles et al. (2012).

The discussion in Section 1.3 implies that estimated standard errors for reduced-bias estimators of transformed parameters can be computed using the delta method, as for the ML estimator.

References

Bull, S. B., C. Mak, and C. M. T. Greenwood (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis* **39**, 57 – 74.

Di Caterina, C. and I. Kosmidis (2019). Location-adjusted Wald statistics for scalar parameters. *Computational Statistics & Data Analysis*, *138*, 126 – 142.

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

Gourieroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics*, **8**, 85 – 118.

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

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

Kosmidis, I. and D. Firth (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, **96**, 793 – 804.

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

Kosmidis, I. and D. Firth (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, **108**, 71 – 82.

Kosmidis, I. and N. Lunardon (2020). Empirical bias-reducing adjustments to estimating functions. *arXiv preprint arXiv:2001.03786*.

Kosmidis, I., E. C. Kenne Pagui, and N. Sartori (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing*, **30**, 43 – 59.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 395 – 407.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13 – 22.

Lyles, R. H., Y. Guo, and S. Greenland (2012). Reducing bias and mean squared error associated with regression-based odds ratio estimators. *Journal of Statistical Planning and Inference*, **142**, 3235 – 3241.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5 – 42.

Part II

State-switching varying-coefficient stochastic differential equations

Timo Adam[1], Richard Glennie[1], and Théo Michelot[1]

[1] University of St Andrews, St Andrews, UK

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

Abstract: Varying-coefficient stochastic differential equations (SDEs) are a useful tool to uncover mechanistic relationships underlying time series. By modelling the parameters of the process of interest as smooth functions of covariates, they provide an extension of basic SDEs to capture detailed, non-stationary features of the data-generating process. In practice, these parameters often vary at multiple time scales, which we illustrate using dive data collected on beaked whales: while their posture in the water within a single dive can be described by varying-coefficient SDEs, different types of dive have different dynamics. In this paper, we propose *state-switching* varying-coefficient SDEs as a novel class of statistical models that accounts for disparate patterns *between* dives while simultaneously allowing us to make inference on the underlying behavioural processes that occur *within* dives. This enables us to draw a multi-scale picture of the whales' diving behaviour.

Keywords: Hidden Markov models; Smoothing splines; Stochastic differential equations; Temporal resolution; Time series modelling.

1 Introduction

Stochastic differential equations (SDEs) with covariate-dependent coefficients constitute a popular class of statistical models for time series. In recent years, they have been applied to model the movement of elephants, the body condition of seals, and the diving behaviour of whales (Michelot *et al.*, 2021), to name but a few examples. However, in practice, the relationship between the parameters of the process of interest and covariates can be subject to state-switching over time (Leos-Barajas *et al.*, 2017; Adam *et al.*, 2019), which cannot readily be accommodated in the existing approach. To overcome this problem, we propose a *state-switching* extension of varying-coefficient SDEs. The suggested approach is applied to dive data

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

collected on beaked whales (a subset of which was used in Michelot *et al.*, 2021), where we analyse the dynamics of their postural angles (pitch and roll) through different dive phases (*within*-dive scale) and how these vary across dives (*between*-dives scale).

2 Methods

2.1 Model formulation and dependence structure

State-switching varying-coefficient SDEs comprise two stochastic processes that are connected with each other:

- a hidden state process, $\{B_d\}_{d=1,\dots,D} \in \{1,\dots,N\}$, where d is a dive index, D is the number of dives, and N is the number of states;
- an observed state-dependent process, $\{Y_{d,t}\}_{d=1,\dots,D,t=1,\dots,T}$, where the index t denotes the t-th observation within the d-th dive.

The hidden state process is modelled by a discrete-time, N-state Markov chain with initial distribution $\boldsymbol{\delta} = (\delta_i)$, $\delta_i = [B_1 = i]$, and transition probability matrix (t.p.m.) $\boldsymbol{\Gamma} = (\gamma_{i,j})$, $\gamma_{i,j} = [B_{d+1} = j | B_d = i]$. While the observed state-dependent process can be any Wiener process, we consider the specific case of Brownian motion with t-distributed noise, which is determined by its drift (the average change of the process over an infinitesimal small time interval) and diffusion (its variability). For each of the two dive variables, the drift, $r_{d,t}^{(b_d)}$, and diffusion, $s_{d,t}^{(b_d)}$, are modelled as

$$\begin{aligned} r_{d,t}^{(b_d)} &= \zeta_d + f_r^{(b_d)}(x_{d,t}),\ \zeta_d \sim N\big(\mu_\zeta, (\sigma_\zeta)^2\big); \\ \log\big(s_{d,t}^{(b_d)}\big) &= \xi_d + f_s^{(b_d)}(x_{d,t}),\ \xi_d \sim N\big(\mu_\xi, (\sigma_\xi)^2\big), \end{aligned}$$

where $x_{d,t}$ is a covariate and $f_r^{(b_d)}$ and $f_s^{(b_d)}$ are basis-penalty smooths,

$$f_\theta^{(b_d)}(x_{d,t}) = \sum_{k=1}^{m} \beta_{\theta,k}^{(b_d)} \psi_{\theta,k}(x_{d,t}),$$

where $\beta_{\theta,k}^{(b_d)}$ is the state-dependent basis coefficient associated with the k-th basis function, $\psi_{\theta,k}(x_{d,t})$, and m denotes the number of basis functions considered (Eilers and Marx, 1996). Incorporating other basis functions, such as individual-specific random effects, is straightforward.

2.2 Likelihood evaluation and model fitting

As the transition density of the model outlined in Section 2.1 is Markovian, the likelihood of all observations within the d-th dive, $\mathbf{y}_d = (y_{d,1},\dots,y_{d,T})$,

being generated by the j-th varying-coefficient SDE is given by

$$[\mathbf{y}_d|B_d = j] = \prod_{t=2}^{T}[Y_{d,t} = y_{d,t}|y_{d,t-1}, r_{d,t}^{(j)}, s_{d,t}^{(j)}], \tag{1}$$

where $j \in \{1, \ldots, N\}$. The transition density in (1) can be written as

$$[Y_{d,t} = y_{d,t}|y_{d,t-1}, r_{d,t}^{(j)}, s_{d,t}^{(j)}] = f\left(\frac{y_{d,t+1} - y_{d,t} - r_{d,t}^{(j)}\Delta_{d,t}}{s_{d,t}^{(j)}\sqrt{\Delta_{d,t}}}\right) \cdot \frac{1}{s_{d,t}^{(j)}\sqrt{\Delta_{d,t}}},$$

where f is the density of a Student's t-distribution and $1/(s_{d,t}^{(j)}\sqrt{\Delta_{d,t}})$ is the Jacobian of the transformation from $Y_{d,t}$ to the increment $D_{d,t}$ (Michelot *et al.*, 2021). By exploiting the Markovian dependence structure, the likelihood of the full model can be evaluated using the forward algorithm,

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \boldsymbol{\delta}\mathbf{P}(\mathbf{y}_1)\prod_{d=2}^{D}\boldsymbol{\Gamma}\mathbf{P}(\mathbf{y}_d)\mathbf{1}, \tag{2}$$

where $\mathbf{P}(\mathbf{y}_d) = \text{diag}([\mathbf{y}_d|B_d = 1], \ldots, [\mathbf{y}_d|B_d = N])$ and $\mathbf{1}$ denotes a column vector of ones (Zucchini *et al.*, 2016). To avoid overfitting, we add a roughness penalty term to (2) and maximise the penalised log-likelihood

$$l_p(\boldsymbol{\theta}|\mathbf{y}) = \log\big(\mathcal{L}(\boldsymbol{\theta}|\mathbf{y})\big) - \sum_{i=1}^{2}\sum_{j=1}^{N}\lambda_i^{(j)}\boldsymbol{\beta}_i^{(j)\top}\mathbf{S}_i\boldsymbol{\beta}_i^{(j)},$$

where $\lambda_i^{(j)}$ is a smoothing parameter, for the i-th smooth term, $\boldsymbol{\beta}_i^{(j)}$ is a vector of basis coefficients, and $\mathbf{S}_i$ is the smoothing matrix associated with the chosen penalty (Michelot *et al.*, 2021). The smoothing parameters can be estimated by optimising the Laplace-approximated marginal likelihood; see Wood *et al.* (2017) for details. Model fitting is conducted using Template Model Builder (Kristensen *et al.*, 2016).

3 Results

The estimated state-dependent drift and diffusion in the whales' postural angles (pitch and roll) as functions of the proportion of time through dive, along with 95% confidence intervals, are displayed in Figure 1. When state 1 (2) is active, the drift in pitch decreases (increases) through the dive, where the diffusion is smallest (largest) in the middle of the dive, and increases (decreases) towards its start and end. The diffusion in roll is much larger in state 2 than in state 1, where the decrease in the middle of the dive (state 1) indicates gliding behaviour, which most often occurred during shallow dives. The increased diffusion exhibited in state 2, in contrast, suggests

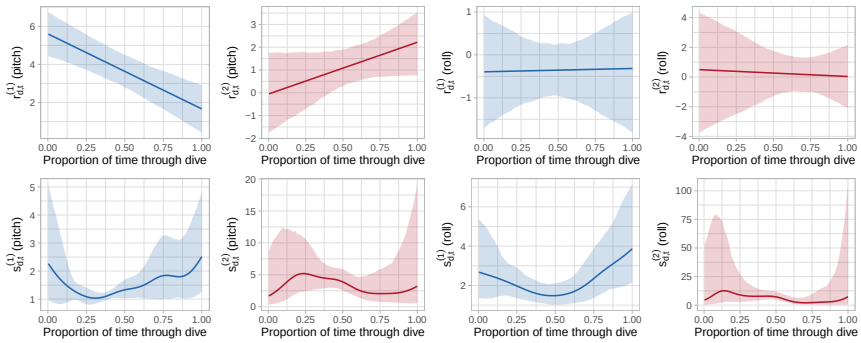


FIGURE 1. Estimated state-dependent drift (top) and diffusion (bottom) in the whales' postural angles (pitch and roll) as functions of the proportion of time through dive, along with 95% confidence intervals. Blue refers to state 1 (gliding behaviour), while state 2 (foraging behaviour) is coloured in red.

continued stroking, which can be linked to foraging behaviour and most often occurred during deep dives. The t.p.m.s that determine the switches between these two dive types were estimated as

$$\hat{\mathbf{\Gamma}}_{\text{Pitch}} = \begin{pmatrix} 0.832 & 0.168 \\ 0.433 & 0.567 \end{pmatrix}, \quad \hat{\mathbf{\Gamma}}_{\text{Roll}} = \begin{pmatrix} 0.835 & 0.165 \\ 0.764 & 0.236 \end{pmatrix},$$

which imply the stationary distributions $(0.720, 0.280)$ and $(0.822, 0.178)$, indicating that, according to both models, most dives were generated in state 1. Furthermore, the high (low) persistence in state 1 (2) suggests that a foraging dive, which is characterised by a high energy consumption, is likely to be followed by multiple, less energy-consuming gliding dives.

4 Discussion

The proposed modelling framework can be extended in various ways. To gain more detailed insights into the interaction of whales with their environment, covariates, such as exposure to underwater noise, can be incorporated into the hidden state process (do they alter their behaviour only at the *within*-dive scale, e.g., by returning to the surface, or also at the *between*-dives scale, e.g., by exhibiting different dive types after being exposed to underwater noise?). More generally, multivariate models, or models with other stochastic processes beyond Brownian motion (e.g., Ornstein-Uhlenbeck processes) can be used, the investigation of which provides a promising avenue for future research.

Acknowledgments: The beaked whale data were collected as part of the SOCAL–BRS project, primarily funded by the US Navy's Chief of

Naval Operations Environmental Readiness Division and subsequently by the US Navy's Living Marine Resources Program. Additional support for environmental sampling and logistics was also provided by the office of Naval Research, Marine Mammal Program. All research activities for that study were authorised and conducted under US National Marine Fisheries Service permit 14534; Channel Islands National Marine Sanctuary permit 2010–004; US Department of Defense Bureau of Medicine and Surgery authorisation; a federal consistency determination by the California Coastal Commission; and numerous institutional animal care and use committee authorisations.

References

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

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**(2), 89 – 121.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., and Bell, B. (2016). TMB: automatic differentiation and Laplace approximation. *Journal of Statistical Software*, **70**, 1 – 21.

Leos-Barajas, V., Gangloff, E.J., Adam, T., Langrock, R., van Beest, F.M., Nabe-Nielsen, J., and Morales, J. (2017). Multi-scale modeling of animal movement and general behavior data using hidden Markov models with hierarchical structures. *Journal of Agricultural, Biological and Environmental Statistics*, **22**(3), 232 – 248.

Michelot, T., Glennie, R., Harris, C., and Thomas, L. (2021). Varying-coefficient stochastic differential equations with applications in ecology. *Journal of Agricultural, Biological and Environmental Statistics*, **26**(3), 446 – 463.

Wood, S.N., Pya, N., and Säfken, B. (2017). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, **111**(516), 1548 –1563.

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

Modelling zebrafish escape strategies with circular modal regression

María Alonso-Pena[1], Rosa M. Crujeiras[1]

[1] Centre for Mathematical Research and Technology Transfer of Galicia, (CITMAga), Universidade de Santiago de Compostela, Spain

E-mail for correspondence: mariaalonso.pena@usc.es

Abstract: The escape behaviour of larval zebrafish from a chasing predator is analysed by using a fully nonparametric method, which estimates the conditional local modes of a circular response variable. The performance of the method is studied both asymptotically and through simulation experiments. This new methodology allows to flexibly model the preferred escape directions of the fish when startled by a predator.

Keywords: Animal escape; Kernel smoothing; Multimodal regression.

1 Introduction

A large number of works dealing with animal behaviour study the escape responses of different animals. More specifically, understanding escape strategies when animals are chased by potential predators is a highly relevant issue in the biological field (see Domenici et al., 2011). This work focuses on analysing how larval zebrafish escape from predators, from a modal regression perspective. The data were obtained by Nair et al. (2017) in an experiment where a robot disguised as an adult zebrafish moved through an aquarium startling zebrafish larvae. Our variable of interest, Φ, is the direction of escape, while the stimulus direction Θ (the angle in which the larvae perceive the threat) is considered as a covariate. Figure 1 depicts a diagram showing how the variables were measured. Escape directions in $[-\pi, 0)$ are known as contralateral escapes (see Figure 1, left), which is the expected behaviour of the animals. Directions in $[0, \pi)$ indicate ipsilateral escapes (Figure 1, right). Nair et al. (2017) used simple linear regression to model the data (see left panel of Figure 2), obtaining that the angle of stimulus was not a significant covariate. However, the circular (periodic) nature of the variables makes usual regression methods not suitable for a proper analysis. The middle panel of Figure 2 shows the same dataset where the units of Φ are transformed to from $[-\pi, \pi)$ to $[0, 2\pi)$, along with

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

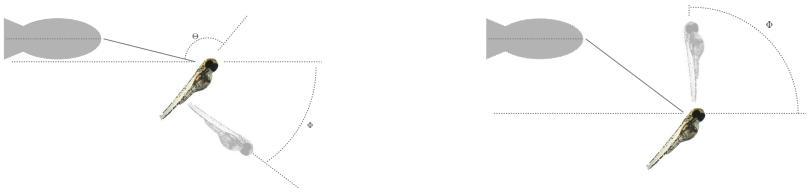


FIGURE 1. Diagram of the zebrafish experiment, with a coloured larvae representing the initial position of the fish, the translucent one indicating the movement of the larvae and the grey shape representing the robot predator.

the regression line estimated after the unit change, and it is clear that the regression line obtained is different from the one in the left panel. In addition, conflictive conclusions are obtained when not taking into account the circular nature of the variables, given that, after the unit transformation, the angle of stimulus is found to be a significant covariate. To overcome this problem, methods specifically tailored for circular data can be used. The continuous line in the right panel of Figure 2 shows the circular nonparametric estimator proposed by Di Marzio et al. (2013). However, although this estimator accounts for the circular nature of the variables, it may lie on regions without data, which is due to the multimodal structure of the data.

We propose a new methodology where the conditional local modes of the escape direction are estimated, instead of the conditional mean targeted by classical methods, while still considering the periodic behaviour of the data. The potential of the estimator is highlighted in the right panel of Figure 2, where the dotted line represents the circular multimodal estimator, which is able to capture the different trends in the data.

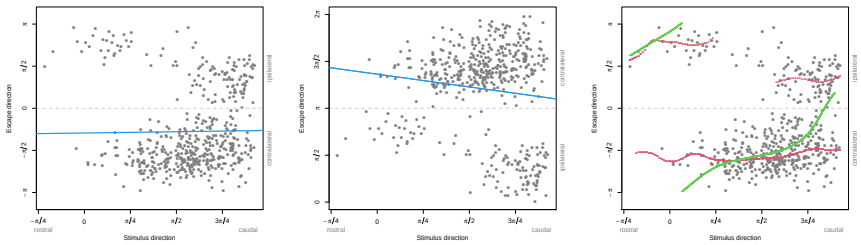


FIGURE 2. Scatter plots of the zebrafish dataset with fitted regression lines (left and middle) and nonparametric mean (right, continuous line) and modal (right, dotted lines) estimators.

2 Circular multimodal regression

The modal regression multifunction is defined as the local maxima of the conditional density function for a given value of the predictor, *i.e.*,

$$M(\theta) = \left\{ \phi : \frac{\partial}{\partial \phi} f(\phi|\theta) = 0, \frac{\partial^2}{\partial \phi^2} f(\phi|\theta) < 0 \right\}, \tag{1}$$

with $f(\phi|\theta)$ being the conditional density of Φ given the value of Θ. Note that $M(\theta)$ is not a function, but a multi-valued function or multifunction. In order to estimate $M(\theta)$, we use an indirect approach: first, $f(\phi|\theta)$ is estimated with a circular kernel regression method; afterwards, the modes of the conditional density are computed iteratively with the so-called circular mean shift algorithm.
More formally, given a bivariate sample $\{(\Theta_i, \Phi_i)\}_{i=1}^n$ from (Θ, Φ), the estimator of the conditional density (see Di Marzio et al., 2016) is given by

$$\hat{f}_{\nu,\kappa}(\phi|\theta) = \left(\sum_{i=1}^{n} K_\nu(\Theta_i - \theta) K_\kappa(\Phi_i - \phi) \right) \Big/ \left(\sum_{i=1}^{n} K_\nu(\Theta_i - \theta) \right),$$

where K_ν and K_κ are circular kernel functions with concentration (smoothing) parameters ν and κ, respectively. The local maxima of the conditional circular kernel density estimator cannot be computed analytically, and therefore a conditional version of the circular mean shift algorithm is employed to compute them numerically. We summarise the algorithm when the kernel associated to the response variable, K_κ, is a von Mises density. For each value in the support of the predictor variable, namely θ, initial values $\phi_1^{(0)}, \ldots, \phi_k^{(0)}$ are selected. Then, for the rth initial angle, a local maximum is computed by iterating until convergence

$$\phi_r^{(l+1)} = \operatorname{atan2}\left[S_\theta\left(\phi_r^{(l)}\right), C_\theta\left(\phi_r^{(l)}\right) \right], \quad l = 0, 1, \ldots$$

where

$$S_\theta\left(\phi_r^{(l)}\right) = \sum_{i=1}^{n} K_\nu(\Theta_i - \theta) \exp\{\kappa \cos(\Phi_i - \phi_r^{(l)})\} \sin \Phi_i$$

and

$$C_\theta\left(\phi_r^{(l)}\right) = \sum_{i=1}^{n} K_\nu(\Theta_i - \theta) \exp\{\kappa \cos(\Phi_i - \phi_r^{(l)})\} \cos \Phi_i.$$

The conditional circular mean shift is, for each value θ, a gradient ascent method on the circumference, and converges to a local maximum of $\hat{f}_{\nu,\kappa}(\phi|\theta)$. The multimodal regression estimator of (2), namely $\widehat{M}(\theta)$, is formed by the collection of estimated local modes for each θ.

In order to measure the quality of the estimator $\widehat{M}(\theta)$, error measures as the Mean Squared Error are not adequate given that the estimator is a multifunction, *i.e.*, for each value of the predictor there might be several different values of the estimated multifunction. Consequently, metrics usually employed in the set estimation context, such as the Hausdorff distance, are considered. It can be proved that, under some regularity conditions, as $n, \nu, \kappa \to \infty$,

$$\widetilde{\text{Haus}}\left[\widehat{M}(\theta), M(\theta)\right] = O\left(\nu^{-1} + \kappa^{-1}\right) + O_P\left(\sqrt{\frac{\kappa^{3/2}\nu^{1/2}}{n}}\right), \qquad (2)$$

where

$$\widetilde{\text{Haus}}\left(A, B\right) = \max\left\{\sup_{x\in A} d(x, B), \sup_{x\in B} d(x, A)\right\}$$

and $d(x, A) = \inf_{z\in A}[1 - \cos(x - z)]$. It is interesting to note that the rate of point-wise convergence of the multimodal regression estimator in (2) coincides with the rate of convergence of the partial derivative of $\hat{f}_{\nu,\kappa}(\phi|\theta)$ with respect to the response variable.

3 Simulation experiments

The finite sample performance of the estimators was investigated through a Monte Carlo study with 500 replicates. We show results obtained when the data was drawn from a multifunction with two different *branches*, given by $M(\theta) = \{3\pi/4\cos\theta - \pi/2, \pi/2 - \cos\theta\}$. Data belonging to each of the two *branches* of the multifunction were simulated as $\Phi_{1i} = 3\pi/4\cos\Theta_{1i} - \pi/2 + \varepsilon_{1i}$, $\Phi_{2i} = \pi/2 - \cos\Theta_{2i} + \varepsilon_{2i}$, where the first subscript denotes the branch number and $\varepsilon_{1i}, \varepsilon_{2i}$ are random errors following a von Mises distribution with zero mean and concentration parameters $\tau = 10, 12, 14$.
To ascertain the quality of the estimators, the modal Integrated Circular Error was computed as $\text{ICE}_m(\widehat{M}) = \int_{-\pi}^{\pi} \widetilde{\text{Haus}}\left[\widehat{M}(\theta), M(\theta)\right] d\theta$. Monte Carlo averages of ICE_m are shown in Table 1 for different sample sizes. The smoothing parameters were selected with a modal cross-validation method. It can be seen that the errors diminish as the sample size increases and, as expected, a higher concentration of the errors leads to smaller values of the average ICE_m.

4 Analysis of the zebrafish data

In order to study the preferred escape directions of the larval zebrafish when startled by the robot predator, the multimodal regression estimator was applied to the zebrafish dataset. The estimated multifunction is depicted on the right panel of Figure 2, as a dotted line. It shows that when

TABLE 1. Monte Carlo averages of ICE_m for the simulated model.

(n_1, n_2)	$\tau = 10$	$\tau = 12$	$\tau = 14$
$(100, 100)$	0.227	0.173	0.150
$(100, 200)$	0.232	0.159	0.120
$(200, 200)$	0.091	0.069	0.059
$(200, 300)$	0.092	0.066	0.049
$(300, 300)$	0.055	0.043	0.033

the robot appeared in the fish's peripherical vision (stimulus directions tending towards $-\pi/4$ or π), there are two estimated branches, one corresponding to a contralateral escape and another one indicating an ipsilateral escape. On the other hand, when the animals see the robot laterally (where their eyes are located), there is just one estimated trend, indicating a contralateral escape. This contrasts with the conclusions obtained with the more classical conditional mean estimator (continuous line in right panel of Figure 2). Note that when the multimodal estimator detects just one escaping response, the estimated preferred directions are very similar to the expected escape directions estimated with the method of Di Marzio et al. (2013).

The multimodal estimator can also be used to perform inference from a modal perspective, such as for the construction of prediction sets. For a prediction level of 0.90, prediction sets for new observations are represented in the right panel of Figure 3, where it is seen that the modal prediction sets are narrower than the prediction bands based on the mean estimator of Di Marzio et al. (2013) (left panel of Figure 3).

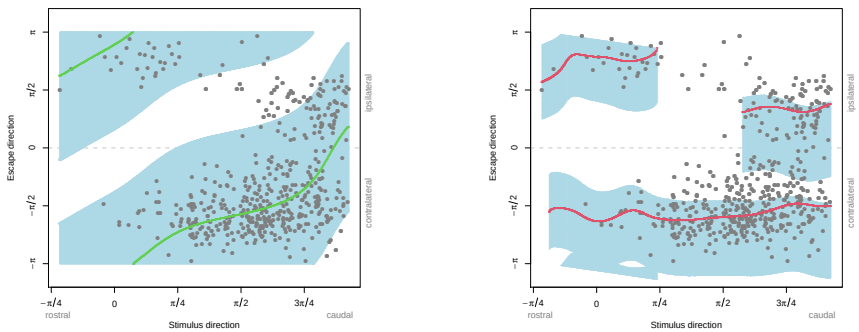


FIGURE 3. Scatter plots of the zebrafish dataset with the mean (left) and modal (right) regression estimators and 90% prediction sets.

Acknowledgments: This work was supported by Project PID2020-11658 7GB-I00 funded by MCIN/AEI/10.13039/501100011033, the Competitive Reference Groups 2021–2024 (ED431C 2021/24) from Xunta de Galicia and grant ED481A-2019/139 from Xunta de Galicia. The authors also acknowledge the Supercomputing Center of Galicia (CESGA) for the computational resources.

References

Di Marzio, M., Panzera, A. and Taylor, C.C. (2013). Non-parametric regression for circular responses. *Scandinavian Journal of Statistics*, **40**, 238 – 255.

Di Marzio, M., Fensore, S., Panzera, A. and Taylor, C.C. (2016). A note on nonparametric estimation of circular conditional densities. *Journal of Statistical Computation and Simulation*, **86**, 2573 -– 2582.

Domenici, P., Blagburn, J.M. and Bacon, J.P. (2011) Animal escapology I: Theoretical issues and emerging trends in escape trajectories. *Journal of Experimental Biology*, **214**, 2463 -– 2473.

Nair et al. (2017). Fish prey change strategy with the direction of the threat. *Proceedings of the Royal Society B*, **284**, 20170393.

Spatial Modelling with Covariates for Survey Data with Positional Uncertainty

Umut Altay[1], John Paige[1], Andrea Riebler[1], Geir-Arne Fuglstad[1]

[1] Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

E-mail for correspondence: umut.altay@ntnu.no

Abstract: The Demographic and Health Surveys (DHS) Program provides GPS coordinates for visited clusters. However, these coordinates are jittered according to a known jittering distribution to ensure privacy of participants. In recent work, we developed a fast method to account for such jittering in a Bayesian hierarchical model. This paper extends this approach to also account for uncertainty that arises when spatial covariates are extracted from rasters based on jittered GPS coordinates. We use this full geostatistical model with both the spatial effect and covariates to estimate vaccination coverage based on the DHS survey in Nigeria 2018 while accounting for jittering. The approach is fast and we find that accounting for jittering on average gives relatively 6.4% lower coefficients of variation for predictions and that the uncertainty about the effect of the covariates increases compared to treating GPS coordinates as correct.

Keywords: Template Model Builder; Spatial Anonymization; Demographic Health Survey; Measles Containing Vaccine.

1 Introduction

Uncertainty in spatial covariates is sometimes induced by uncertainty in the associated spatial locations used to extract them. A common way to address this is to ignore the positional error and fit the geostatistical model using the covariate at the observed location. Another approach is to average the covariate values throughout the raster pixels within the jittering zone of each cluster and then to weigh them with respect to their distances from the corresponding cluster center (Perez-Heydrich et al., 2013). This might fail to capture some of the uncertainty due to assigning same average

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

value to the multiple locations. In this study we estimate the probability of receiving one dose of measles-containing vaccine (MCV1) for 12-23 months old children in Nigeria. We also aim to estimate the associations between the covariates and the response.

2 Data and Methods

We use the Nigeria-2018 Demographic Health Survey (NDHS-2018). The survey consists of 1,389 observed clusters that are scattered over 37 counties. We extract the number of 12-23 months old children who received a dose of MCV1 vaccine, as the response variable for each cluster. In total there are 3,210 children who fit this description. In previous work, we accounted for jittering in survey clusters by creating a number of integration points around each cluster center and integrating out the unknown locations from the joint likelihood function. This gives a mixture of different likelihoods for each cluster.

Local Burden of Disease Vaccine Coverage Collaborators (2020) conducted a study to map routine measles vaccination in low- and middle-income countries. Among the covariates they have used, we included five, all extracted from data rasters: elevation (Elev) from the mean sea level in meters, population density (Pop) (estimated total number of people per grid-cell), travel time (Travel) to the nearest city in minutes, urbanization rate (Urb) and minimum distance (Dist) to the lakes and rivers in degrees. Except for the urbanization rate and the minimum distance to the lakes and rivers, all other covariates are transformed via $ln(1 + x)$, before the computations.

We include five covariates and a spatial random effect to explain the spatial variation in survey responses. The DHS jitters urban clusters up to 2 km. 99% of the rural clusters are jittered up to 5 km, while the remaining 1% are jittered up to 10 km (Burgert et al., 2013), while the new locations are required to remain within the same county. For clusters $c = 1, \ldots, C$, observations follow a binomial model:

$$y_c | r(s^*), n_c \sim Binomial(n_c, r(s^*))$$

where, y_c denotes the total number of children who received an MCV1 dose among the n_c children who are 12-23 months old and living in the corresponding household cluster. Accordingly, we model the risk of receiving one dose of MCV1 vaccine at unknown true locations $\boldsymbol{s}_c^* \in \mathbb{R}^2$ as follows:

$$\text{logit}(r(\boldsymbol{s}^*)) = \mu + \beta^\intercal \mathbf{x}(\boldsymbol{s}^*) + u(\boldsymbol{s}^*),$$

Here, the intercept and design matrix are denoted by μ and X, respectively. $u(\cdot)$ is a Gaussian random field (GRF) with a Matérn covariance function with marginal variance σ^2_{SF}, range ρ, and fixed smoothness $\nu = 1$. We

constructed the approach that accounts for jittering based on numerically integrating out unknown true locations, which is only a 2D integral for each observation. We implement the approach using C++ to utilize autodifferentiation via the TMB. Extending the approach into a full geostatistical model with covariates involves extracting the values of covariates at each integration point and include them in the computation process via the design matrix. Detailed explanation about the model and the application on TMB can be found in (Altay et al., 2022).

3 Results and Discussion

We fitted the above model in two different ways: by assuming that the observed locations are the true locations, and by applying the new method that accounts for jittering. Figure 1 below shows the predicted posterior expectations for the probabilities of receiving an MCV1 vaccine dose for the children aged 12-23 months old, together with the corresponding coefficients of variations (CV), when jittering is accounted for. The figures show that the vaccination probability decreases together with the increasing uncertainty towards the north of Nigeria. Figure 2 shows the cross-plots of the predicted posterior expectations for the probabilities and the corresponding coefficient of variation values that are obtained from the standard and jittering accounted models. Left hand side plot shows that the models tend to yield different predictions, where the standard model appears to attenuate the predictions close to 0 or 1 compared to the model that accounts for jittering. It is also visible from the right hand side plot that the CVs are smaller when jittering is accounted for. Table 1 shows that when positional error is accounted for, there is a slight increase in the length of 95% credible intervals of the corresponding estimated model parameters. We found out that on average, we obtain relatively 6.4% lower CVs for the predictions, when we account for jittering. The computation time increased from 1.85 minutes to 11.75 minutes in the new approach, but this is still quite fast compared to the current methods with the prohibitively slow computation times.

4 Conclusion

The new approach provides a fast implementation to account for the positional uncertainty in a geostatistical model including spatial covariates, with a slight increase of uncertainty in covariates. The amount of difference in the parameter estimates and the uncertainties indicate the importance of accounting for jittering in the spatial covariates. In future work, we plan to conduct simulation studies to evaluate model performance under various scenarios. An interesting direction of research would be to compare the presented method against the approach where the covariates are smoothed by averaging them over the jittering zones around each cluster.

TABLE 1. Intervals of covariate values, median parameter estimates and the lengths of corresponding 95% credible intervals that are obtained from the model that accounts for jittering. The medians and lengths in the parantheses belong to the parameter estimates that are obtained from the standard model.

Parameters	Value Interval	Median	Length
Intercept		-0.98 *(-0.76)*	2.76 *(2.62)*
β_{Dist}	(0, 2.24)	0.18 *(0.24)*	0.89 *(0.95)*
β_{Urb}	(0, 100)	-0.012 *(-0.004)*	0.013 *(0.006)*
β_{Travel}	(0.00, 7.04)	-0.03 *(-0.03)*	0.19 *(0.14)*
β_{Elev}	(0.69, 7.99)	0.03 *(0.02)*	0.46 *(0.45)*
β_{Pop}	(0, 7.59)	0.65 *(0.38)*	0.38 *(0.23)*
ρ		155 *(127)*	137 *(112)*
σ^2_{SF}		0.84 *(0.89)*	0.30 *(0.31)*

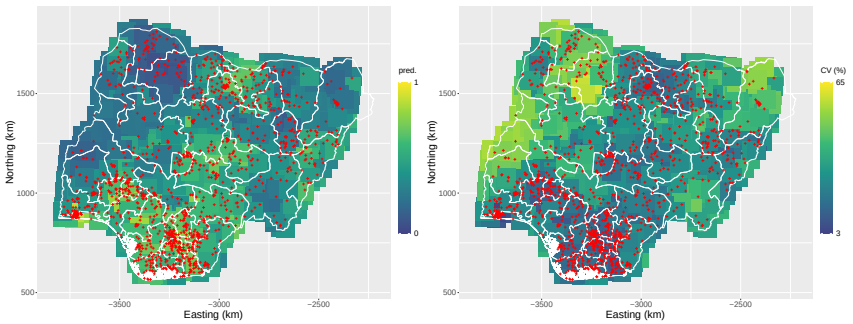


FIGURE 1. Predicted posterior expectations ("pred.") for the probabilities of receiving an MCV1 vaccine dose (left) and the corresponding coefficients of variations (CV) (right) for the model that accounts for jittering. The red points indicate 1,322 DHS survey clusters in Nigeria.

References

Altay, U., Paige, J., Riebler, A. and Fuglstad, G-A. (2022). Accounting for Spatial Anonymization in DHS Household Surveys. *arXiv preprint arXiv:2202.11035*.

Burgert, C. R., Colston, J., Roy, T., Zachary, B. (2013). Geographic displacement procedure and georeferenced datarelease policy for the Demographic and Health Surveys, DHS Spatial Analysis Reports No. 7. *Demographic and Health Surveys Program*.

Local Burden of Disease Vaccine Coverage Collaborators. (2020). Mapping routine measles vaccination in low- and middle-income countries. *Nature*.

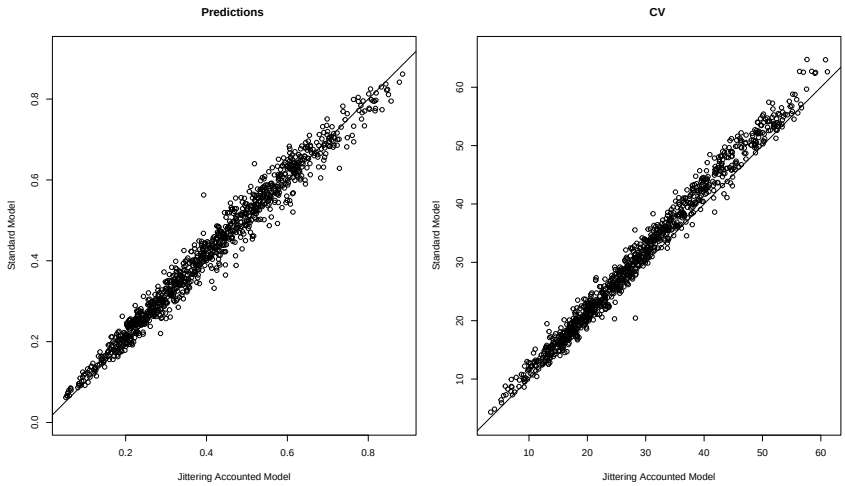


FIGURE 2. Cross-plots of predicted posterior expectations for the probabilities of receiving an MCV1 vaccine dose (left) and the corresponding coefficients of variations (CV) (right) that are obtained from the standard model and the model that accounts for jittering.

Perez-Heydrich, C., Warren, J. L., Burgert, C. R., Emch, M,E. (2013). Guidelines on the Use of DHS GPS Data. Spatial Analysis Reports No. 8. *ICF International*.

A quantile regression ranking for cyber-risk assessment

Mario Angelelli[1,2], Christian Catalano[1,3]

[1] Department of Innovation Engineering, University of Salento, Lecce, IT
[2] CAMPI - Centre of Applied Mathematics and Physics for Innovation, Lecce, IT
[3] CRLab - Cybersecurity Research Lab, Lecce, IT

E-mail for correspondence: mario.angelelli@unisalento.it

Abstract: The ranking of cyber-vulnerabilities based on their severity is a major decision problem in the digital society. This ranking is often based on information related to intrinsic properties of the cyber-vulnerabilities, but contextual factors (diffusion of a vulnerable technology, available resources to exploit a vulnerability) may have an impact on the actual risk attributed to cyber-incidents. This work introduces a quantile regression model as a basis for the ranking of cyber-vulnerability impact dimensions. The prioritisation of these impact dimensions in relation to the quantile level is discussed, with the aim of supporting statistical modelling for informed cyber-risk assessment and threat intelligence.

Keywords: Ranking; Quantile regression; Cyber-risk.

1 Introduction

Cyber-security is becoming a crosscutting issue for digital societies. Cyber-vulnerabilities of devices, networks, systems, or Information and Communication Technologies (ICTs) may have an impact on both organisations and individuals, since system failures and cyber-attacks may compromise or interrupt service supply, or undermine the operational continuity of critical infrastructures. New vulnerabilities are emerging as a consequence of large digital connectivity, and even individual devices or sensors may represent an access point to information systems through escalation procedures. In addition to economic losses and potential impacts on safety, cyber-incidents also relate to the improper use of personal or sensitive data, which leads to privacy concerns.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

This context is leading regulatory bodies and companies to develop and adopt new methods for cyber-risk assessment, aiming at a more informed decision-making. Sources supporting these decisions are heterogeneous and include severity rankings produced by Institutions (in particular, the National Institute of Standards and Technology - NIST and Computer Security Incident Response Teams - CSIRTs), but also reports, expert evaluations, and data from web resources and databases. The decision-maker can choose different (official and unofficial) sources to prioritise her actions, both defensive or offensive, depending on her objectives.
In this contribution, we explore the role of risk level attribution in ranking cyber-vulnerabilities based on a set of observations from different databases. We start from a quantile regression model to infer the conditional τ-level quantile given a set of predictor variables; then, we discuss a criterion to prioritise severity attributes for cyber-vulnerabilities based on the estimated parameters.
This work extends the literature of statistical modelling of cyber-risk, see e.g. Giudici and Raffinetti (2021). We pay special attention to the different prioritisation that a decision-maker can assess to mitigate (for defenders) or exploit (for attackers) a cyber-vulnerability. The statistical modelling is also relevant to connect different information sources, in particular technical proprieties of a cyber-vulnerability (regressors) and actual data quantifying hosts exposed to such a vulnerability.

2 Dataset

The attributes assessing the intrinsic impact of a cyber-vulnerability come from the attack vector in the NIST model (Mell, Scarfone, and Romanosky, 2007). The attack vector is a 6-tuple of categorical variables that evaluate the severity of a vulnerability. The three impact dimensions refer to the potential impact on data Confidentiality (X_{C}), Integrity (X_{I}), and Availability (X_{A}), respectively. Each of these attributes has three modalities: "none", "partial", and "complete". Three additional attributes assess technical characteristics for exploitation: we will discuss X_{AV}, i.e. the Access Vector with modalities "Requires local access", "Local Network accessible", and "Network accessible". The remaining components refer to Access Complexity (X_{AC}), with modalities "high", "medium", and "low", and Authentication requirements (X_{Au}), with modalities "Requires no authentication", "Requires single instance of authentication", and "Requires multiple instances of authentication". These variables represent the regressors in our model.
The envisaged response variables y indicate the actual extent of a cyber-vulnerability: in addition to the intrinsic severity from the attack vector, they might depend on external factors too, e.g. the diffusion of vulnerable technology, or available resources to exploit such weaknesses.

We realised a dataset from observation of hosts exposed to known vulnerabilities: each vulnerability in the dataset is identified by a Common Vulnerabilities and Exposures (CVE) code, is endowed with an associated attack vector; for a selected sample of CVEs, we have scraped the Shodan database (`https://exposure.shodan.io`) to retrieve exposure data from different Countries during the period 1999-2021. The final dataset is composed of $n = 714$ records, which represent the statistical units of our regression model. The response variable is the number N_i of exposed hosts reported in the Shodan database for the i-th CVE, $i \in \{1, \dots, n\}$.

3 Model definition

Setting $Q_\tau := \inf_y\{y : \tau \leq F(y))$ as the τ-th quantile for a random variable y with cumulative distribution function (CDF) F, the regression model to estimate Q_τ based on n observed data and k regressors is

$$Q_\tau(y_i|\mathbf{X}_i, \beta) = \mathbf{X}_i^\intercal \cdot \beta(\tau), \quad i \in \{1, \dots, n\}. \tag{1}$$

The estimated parameters $\beta(\tau)$ depend on the choice of the quantile level τ. They are derived from (1) through the minimisation of

$$\hat{\beta}(\tau) := \text{argmin}_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \varrho_\tau(y_i - \mathbf{X}_i^\intercal \cdot \beta) \tag{2}$$

where

$$\varrho_\tau(u) := u \cdot (\tau - \mathbb{I}(u < 0)) \tag{3}$$

and $\mathbb{I}$ is the characteristic function of a subset of $\mathbb{R}$. We refer to Koenker and Hallock (2001) and references therein for further details.
From the coefficient estimates given by (2), we get a criterion to rank the attributes that describe the severity of a cyber-vulnerability based on the risk-level attribution (the quantile level τ). In the present case study, the attributes correspond to the components of the attack vector in a subset $\mathcal{I} \subseteq \{X_{\text{C}}, \dots, X_{\text{Au}}\}$. Being associated with categorical data, we introduce

$$\Pi := \{(p, \ell) : p \in \{X_{\text{C}}, \dots, X_{\text{Au}}\}, \ell \in \{1, \dots, L_p - 1\}\} \tag{4}$$

where L_p is the number of modalities for the p-th variable; then, we adopt an ANOVA representation considering parameters $\beta_\pi(\tau)$ for indicator variables X_π indexed by $\pi \in \Pi$.
For each specification of the quantile level τ, the set of parameters estimated through (2) can be used to produce a qualitative ranking of the attack vector attributes. Specifically, for each τ we consider a subset $\mathcal{S} \subseteq \Pi$ such that the associated variables X_π ($\pi \in \mathcal{S}$) in the regression model (1)-(2) are statistically significant. Then, we order the modalities associated with these variables based on the corresponding estimates $\hat{\beta}_\pi$. Formally, we define a

partial ranking $\prec_\tau$ on the set Π or, equivalently, on the regressors X_π, $\pi \in \Pi$: given $\pi_1, \pi_2 \in \Pi$, $\pi_1 \prec_\tau \pi_2$ means that the estimated effect $\hat{\beta}_{\pi_1}$ of the variable X_{π_1} on the response variable, compared to a reference severity level, e.g. the base level for the associated component of the attack vector, is significantly smaller than the effect $\hat{\beta}_{\pi_2}$ of X_{π_2}.

4 Preliminary results

The regression model highlights a significant contribution of the exploitability characteristics at given quantile levels. When we consider the attack vector modalities as regressors in (1), we find a positive value for $\hat{\beta}_{\mathrm{AV},1} = 14167$ associated with the "local network accessible" property, with t-statistics 2.868 and p-value 0.426%. In general, the statistics $\max_{\pi\in\Pi}\{\hat{\beta}_\pi(\tau)\}$ informs us on the attribute of the attack vector that, if positive, contributes the most to raise the τ-quantile. Assuming that the (log-)exposure acquired from Shodan relates to the actual tendency to detect a vulnerability, the increase of the exposure coefficient associated with the τ-quantile is equivalently described by a reduced population within that value of the exposure coefficient.
Comparing Table 1 and Table 2, we see a ranking inversion following the change of the quantile level, from $\tau = .5$ to $\tau = .89$: parameters estimated for the "Partial" levels in both Confidentiality and Availability (denoted as **C1** and **A1**, respectively) provide different rankings depending on τ. The model with both X_C and X_A is compared with the models without X_C (at $\tau = .5$) or X_A (at $\tau = .89$): an ANOVA test returned a p-value smaller than .001%, with F-values 25.906 at $\tau = .5$ and 10.494 at $\tau = .89$.

5 Conclusion and Future Work

This work is a preliminary study on statistical modelling for threat intelligence, with particular attention to the information resources regarding cyber-vulnerabilities and to the effects of risk acceptance/aversion.
The availability of resources to exploit cyber-vulnerabilities could be known to both attackers and defenders, and this may represent another factor affecting the actual evaluation of cyber-risk. We will extend the present model taking into account the knowledge of exploits from dedicated databases (e.g. ExploitDB, VulnDB).
On the other hand, a deeper investigation is needed to explore the relation between statistical (partial) ranking models, formal decision criteria, and sources of uncertainty that may give rise to multiple orders of priority in the cyber-security domain. A better comprehension of this topic could support its integration with information-theoretic methods for the analysis of secure disclosure properties.

TABLE 1. Summary of quantile regression at level $\tau = .5$.

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
C1	-4394.00	422.147	-10.409	0.000	-5222.807	-3565.193
A1	-1393.00	463.765	-3.004	0.003	-2303.517	-482.483

TABLE 2. Summary of quantile regression at level $\tau = .89$.

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
C1	-6321.00	3099.299	-2.039	0.042	-12400.00	-236.097
A1	-20320.00	3423.372	-5.935	0.000	-27000.00	-13600.00

References

Giudici, P. and Raffinetti, E. (2021) Cyber risk ordering with rank-based statistical models. *AStA - Advances in Statistical Analysis*, **105**(3), 469 – 484.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, **15**(4), 143 – 156.

Mell, P., Scarfone, K., and Romanosky, S. (2007) A complete guide to the common vulnerability scoring system version 2.0. Published by *FIRST-forum of incident response and security teams*, pp. 1 – 23.

Multidimensional latent mixed models for don't know responses in panel data concerning *Financial knowledge*

David Aristei[1], Silvia Bacci[2], Manuela Gallo[1], Maria Iannario[3]

[1] Department of Economics, University of Perugia, Italy
[2] Department of Statistics, Computer Science, Applications G. Parenti, University of Florence, Italy
[3] Department of Political Sciences, University of Naples Federico II, Italy

E-mail for correspondence: maria.iannario@unina.it

Abstract: In the setting of latent construct measurement, multidimensional item response theory (IRT) models are particularly useful when the proper scoring rule of observable items is unclear, as it happens when response categories include the 'Don't know' (DK) option. In this contribution we illustrate how a multidimensional IRT model may be suitably formulated to treat DK responses. We also provide an extension to encompass random effects due to a repeated measurement setting. The main contents of the approach are illustrated through an application on data concerning the measurement of the *Financial knowledge*.

Keywords: 'Don't know' responses; item response models; multidimensional.

1 Introduction

Don't know (DK) responses provided as possible options in multiple-item questionnaires may be due to several unknown reasons, such as poor self-confidence in own competencies, awareness of lack of knowledge or no feeling to express answers due to uncertainty or apathy. Naive approaches usually consider these responses as incorrect answers or missing values, leading to biased latent knowledge measures, as they fail to properly account for the unobserved differences between substantive and DK responses (Krosnick et al., 2002).
Some authors showed that changing the instructions to suppress the DK response improves reliability (Mondak, 2001). Others consider the DKs as missing values by analysing them as a source of uncertainty (Rubin et al.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

1995). Here we consider DKs as responses to all extents because they inform about a specific state of mind of the respondent, and therefore, it is not correct to treat them as missing values.
In the present study, we focus on the measurement of financial knowledge from survey items, taking into account respondents' nonrandom propensity to choose DK answers. To this aim, we analyze longitudinal data from the first and second wave of the survey 'Covid-19 Emergency: Italians between fragility and financial resilience' carried out in 2020 and 2021 by the Italy's Committee for the Planning and Coordination of Financial Education Activities and BVA-Doxa. In particular, we exploit the detailed information on Italian adults' knowledge obtained from *six* multiple-choice test questions on key financial concepts.
From a methodological point of view, we explicitly address the issue of estimating the latent knowledge construct taking into account the DK option, through a Bidimensional latent mixed model (described in section 2) suitably generalized. Model at issue relies on the assumption that the response process may be disentangled in two consecutive steps driven by two latent variables: propensity to answer and financial knowledge. At the first step, both latent traits affect the probability of selecting a substantive response versus DK option. At the second step, conditionally on the selection of a substantive response, financial knowledge affects the probability of a correct answer versus an incorrect one. Random-effects are encompassed in the main model structure for assessing the variability due to repeated responses provided by the same subject to same items.

2 Model description

Let denote the original item responses as $Y_{ip} = 0, 1, \ldots, m-1$, with $i = 1, \ldots, I$ for items, $p = 1, \ldots, P$ for persons, $j = 0, \ldots, m-1$ for response categories, and m number of response categories. In our setting $m = 3$ corresponding to DK option ($Y_{ip} = 0$), incorrect response ($Y_{ip} = 1$), and correct response ($Y_{ip} = 2$). Then, each item is disentangled in sub-items, denoted by $Y_{ip}^{(r)} = 0, 1$, with $r = 1, \ldots, R$ as an index for the decisional node; in our proposal $R = 2$. See Figure 1 for a graphical representation of the multi-node response process through a tree structure. Denoting by $\boldsymbol{\theta}_{pt} = (\theta_{1pt}, \theta_{2pt})'$ the vector of latent traits driving the response process, across $t = 1, 2, \ldots, T$ time points, both a substantive answer (i.e., $Y_{ipt}^{(1)} = 1$ vs. $Y_{ipt}^{(1)} = 0$ if DK is selected) to a generic item i and the value (i.e., $Y_{ipt}^{(2)} = 1$ for correct answer vs. $Y_{ipt}^{(2)} = 0$ for incorrect answer) observed on item i conditionally on having selected a substantive answer are explained through a Bidimensional latent regression two-parameter (2PL) model, as

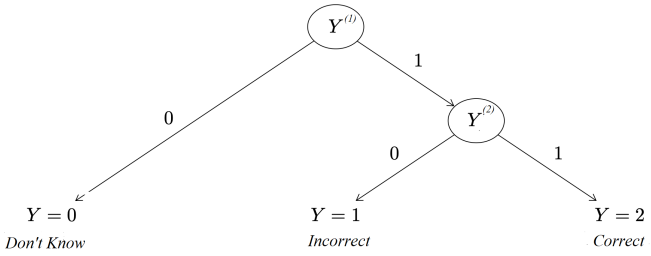


FIGURE 1. Linear response tree for the binary response categories and DK option

follows:

$$\begin{aligned} logit\left[P(Y_{ipt}^{(1)}=1|\boldsymbol{\theta}_{pt})\right] = \quad & \boldsymbol{\gamma}_i^{(1)}\boldsymbol{\theta}_{pt} - \beta_i^{(1)} = \textstyle\sum_d \delta_{di}^{(1)}\gamma_{di}^{(1)}\theta_{dpt} - \beta_i^{(1)} \quad (1) \\ logit\left[P(Y_{ipt}^{(2)}=1|\boldsymbol{\theta}_{pt}, Y_{ipt}^{(1)}=1)\right] = \quad & \boldsymbol{\gamma}_i^{(2)}\boldsymbol{\theta}_{pt} - \beta_i^{(2)} = \textstyle\sum_d \delta_{di}^{(2)}\gamma_{di}^{(2)}\theta_{dpt} - \beta_i^{(2)}. \end{aligned}$$

Here $\delta_{di}^{(1)}$ and $\delta_{di}^{(2)}$ are indicator functions assuming value 1 if item i measures latent trait d ($d = 1, 2$), and 0 otherwise; $\boldsymbol{\gamma}_i^{(1)} = (\gamma_{1i}^{(1)}, \gamma_{2i}^{(1)})'$ and $\boldsymbol{\gamma}_i^{(2)} = (\gamma_{1i}^{(2)}, \gamma_{2i}^{(2)})'$ are vectors of discrimination parameters for items answered at nodes 1 and 2, respectively; $\beta_i^{(1)}$ and $\beta_i^{(2)}$ represent difficulty parameters. According to the specification of pairs $(\delta_{di}^{(1)}, \delta_{di}^{(2)})$, different assumptions about the role of the latent traits are considered.
The presence of a unique latent trait is obtained as a special case, when $(\delta_{1i}^{(1)}, \delta_{2i}^{(1)}) = (1, 0)$ and $(\delta_{1i}^{(2)}, \delta_{2i}^{(2)}) = (1, 0)$, or vice-versa. This happens when unidimensionality assumption holds that is a same latent trait affects both the types of responses, whereas different combinations of $(\delta_{di}^{(1)}, \delta_{di}^{(2)})$ are referred to multiple latent traits. For the latter we may distinguish two situations:

- a between-item multidimensional where $(\delta_{1i}^{(1)}, \delta_{2i}^{(1)}) = (1, 0)$, and $(\delta_{1i}^{(2)}, \delta_{2i}^{(2)}) = (0, 1)$ for all items $i = 1, \ldots, I$.
- a within-item multidimensional structure of latent traits where θ_{1pt} only affects the DK answering (node 1 in Figure 1), whereas θ_{2pt} only affects the substantive answering (node 2 in Figure 1).

Model in eqs. 1 may be easily generalized to allow for individual characteristics that may affect the latent traits encompassing mixed effects too. Thus, a *latent mixed* multidimensional 2PL model is obtained substituting $\boldsymbol{\theta}_{dp}$ denoting the T-by-1 vector for person p across T time points with

$$\boldsymbol{\theta}_{dp} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\nu}_p + \boldsymbol{\epsilon}_p, \qquad d = 1, 2, \tag{2}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are fixed and random effects design matrices, respectively, and $\boldsymbol{\epsilon}_p$ is T-by-1 vector of residuals. The distribution of the random effects are a multivariate normal (Gaussian) distribution with mean $\mathbf{0}$ and $q \times q$ variance-covariance matrix $\boldsymbol{\Sigma}$, that is $\boldsymbol{\nu} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$.
For a simple linear growth model with a a single person-specific intercept and slope, we can rewrite eq. 2 as

$$\boldsymbol{\theta}_{pt} = \pi_{0p} + \pi_{1p} \times (t-1) + \epsilon_{pt},$$

where π_{0p} and π_{1p} are the individual intercept and slope parameters.
From inferential point of view, let $\boldsymbol{y}_{pt}^{(1)}$ denote the vector of responses of individual p ($p = 1, 2, \ldots, n$) to items $1, 2, \ldots, I$, across $t = 1, 2, \ldots, T$. Thus, the *manifest distribution* for individual p across t is given by

$$P(\boldsymbol{y}_{pt}^{(1)}, \boldsymbol{y}_{pt}^{(2)}) = P(\boldsymbol{y}_{pt}^{(1)}, \boldsymbol{y}_{pt}^{(2)} | \boldsymbol{\theta}_{pt}) P(\boldsymbol{\theta}_{pt}).$$

Here

$$P(\boldsymbol{y}_{pt}^{(1)}, \boldsymbol{y}_{pt}^{(2)} | \boldsymbol{\theta}_{pt}) = \prod_{i=1(y_{ipt}^{(1)}=0)}^{I} P(Y_{ipt}^{(2)} = 1 | \boldsymbol{\theta}_{pt})^{y_i^{(2)}} \cdot \Big[1 - P(Y_{ipt}^{(2)} = 1 | \boldsymbol{\theta}_{pt})\Big]^{1-y_i^{(2)}}$$
$$\cdot \prod_{i=1}^{I} P(Y_{ipt}^{(1)} = 1 | \boldsymbol{\theta}_{pt})^{y_i^{(1)}} \cdot \Big[1 - P(Y_{ipt}^{(1)} = 1 | \boldsymbol{\theta}_{pt})\Big]^{1-y_i^{(1)}},$$

with $P(Y_{ipt}^{(1)} = 1 | \boldsymbol{\theta}_{pt})$ and $P(Y_{ipt}^{(2)} = 1 | \boldsymbol{\theta}_{pt})$ obtained from eqs. (1).
From the manifest distribution $P(\boldsymbol{y}_{pt}^{(1)}, \boldsymbol{y}_{pt}^{(2)})$ we define the likelihood function as

$$\mathcal{L}(\boldsymbol{\psi}) = \prod_{t=1}^{T} \prod_{p=1}^{n} P(\boldsymbol{y}_{pt}^{(1)}, \boldsymbol{y}_{pt}^{(2)}).$$

Many computational methods for mixed models with multiple grouping factors are designed for hierarchical models in which the grouping factors form a strictly nested sequence. Preliminary results are obtained by means of a hybrid algorithm which benefits from both Bayesian and frequentist approaches to parameter estimation in order to avoid the computational complexity of evaluating multiple integrals in confirmatory item response models (see Cai (2010) for details). For further inferential issues see Chalmers (2015).

3 Preliminary data analysis

To calibrate the model preliminary results have been exploited on each wave separately where different models were computed. Fitting results favour a within-item multidimensional model, as remarked by AIC and BIC indexes

TABLE 1. Model selection: AIC and BIC indexes for multidim. 2PL models

	First wave		*Second* wave	
Assumption	AIC	BIC	AIC	BIC
Unidimensionality	35623.75	35774.96	34941.14	35092.36
Between-item bidim.	35553.31	35684.53	34624.63	34775.63
Within-item bidim.	**35177.27**	**35366.30**	**34325.69**	**34514.72**

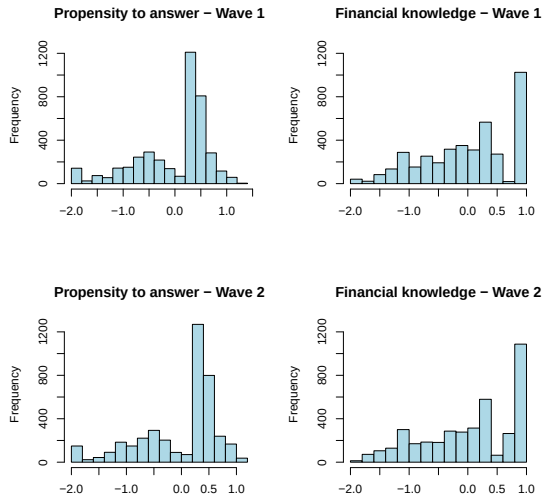


FIGURE 2. Estimated latent traits. Top panel: *first* wave; bottom panel: *second* wave

in Table 1. Thus, we may conceive the response process as driven by two latent traits: (i) *propensity to answer*, described by latent variable θ_{1p}, that affects the selection of DK option versus a substantive category; (ii) *financial knowledge*, described by latent variable θ_{2p}, that affects both outcome of node 1 and outcome of node 2. The empirical distributions of the two estimated latent traits for both waves are shown in Figure 2. Our results clearly point out that DK responses provide significant insights on respondents' financial competencies and they should be taken into account to properly measure the underlying levels of knowledge. In Table 2 are displayed the estimates of item parameters $\gamma_{1i}^{(r)}$, $\gamma_{2i}^{(r)}$, and $\beta_i^{(r)}$ of the within-item bidimensional 2PL model selected for each wave. Parameters $\beta_i^{(r)}$ denote the "easiness" of item i: higher the easiness, higher is the probability of observing answer '1' on the item (i.e., observing answer other than 'DK' for items at Node 1 and observing answer 'correct' for items at Node 2).

TABLE 2. Within-item bidimensional 2PL model: estimates of item parameters

		First wave			*Second* wave		
Node r	Item i	$\gamma_{1i}^{(r)}$	$\gamma_{2i}^{(r)}$	$\beta_i^{(r)}$	$\gamma_{1i}^{(r)}$	$\gamma_{2i}^{(r)}$	$\beta_i^{(r)}$
1	int ($i = 1$)	2.954	1.455	5.549	2.785	1.608	5.406
1	infl ($i = 2$)	2.437	1.050	3.673	2.075	1.025	3.341
1	div ($i = 3$)	2.148	1.740	2.595	2.604	1.886	3.015
1	mortg ($i = 4$)	2.114	1.829	3.049	2.217	1.485	2.979
1	compint ($i = 5$)	2.374	1.158	2.403	2.632	1.102	2.509
1	riskret ($i = 6$)	2.223	2.482	3.676	2.812	2.542	4.203
2	int ($i = 1$)	–	1.366	1.699	–	1.350	1.704
2	infl ($i = 2$)	–	1.676	1.870	–	1.899	1.959
2	div ($i = 3$)	–	0.517	1.520	–	0.382	1.615
2	mortg ($i = 4$)	–	0.939	0.935	–	0.999	1.056
2	compint ($i = 5$)	–	0.976	1.025	–	1.108	1.281
2	riskret ($i = 6$)	–	2.434	3.134	–	3.437	3.741

Parameters $\gamma_{1i}^{(r)}$ and $\gamma_{2i}^{(r)}$ $(r = 1, 2)$ denote how each item "discriminates" between individuals with different levels of the underlying latent trait (θ_1 or θ_2, respectively). They can be interpreted as factor loadings in a factor analysis, denoting the contribution of each item to the measurement of the related latent trait. Results are very similar in the two waves even if a slightly improving in the second wave appears for some item when financial knowledge is considered. It suggests the possible extension takeing into account mixed effects.

References

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, **75**, 33 – 57.

Chalmers, R. P. (2015). Extended Mixed-Effects Item Response Models with the MH-RM Algorithm. *Journal of Educational Measurement*, **52**, 200 – 222.

Krosnick, J.A., Holbrook, A.L., Berent, et al. (2002). The impact of 'no opinion' response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, **66**(3), 371 – 403.

Mondak, J. J. (2001). Developing valid knowledge scales. *American Journal of Political Science*, **45**, 224 – 238.

Rubin, D.B., Stern, H., Vehovar, V. (1995). Handling "don't know" survey responses: The case of the Slovenian plebiscite. *Journal of the American Statistical Association*, **90**, 822 – 828.

Evaluating the school effect using multilevel models: adjusting for pre-test or using gain scores?

Bruno Arpino[1], Silvia Bacci[1], Leonardo Grilli[1], Raffaele Guetto[1], Carla Rampichini[1]

[1] Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Italy

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

Abstract: We consider estimating the effect of a treatment on a given outcome measured on subjects tested both before and after treatment assignment. A vast literature compares the competing approaches of modeling the post-test score conditionally on the pre-test score versus modeling the difference, namely the gain score. Our contribution resides in analyzing the merits and drawbacks of the two approaches in a multilevel setting. This is relevant in many fields, for example education with students nested into schools. The multilevel structure raises peculiar issues related to the contextual effects and the distinction between individual-level and cluster-level treatments. We derive approximate analytical results and compare the two approaches by a simulation study.

Keywords: Achievement tests; Random effects model; Value added.

1 Introduction

We consider the problem of estimating the school effect on student achievement, when a pre-test is available. Our work is inspired by achievement tests implemented at the 5th grade (end of primary school) and 8th grade (end of lower secondary school) by the Italian agency for the evaluation of school system (Invalsi). We merge students with scores on these two grades to assess the school value added based on the progress from grade 5th to grade 8th. Specifically, we aim to evaluate if the school effect is different between public and private schools, controlling for possible confounders. While some confounders are observed, e.g. gender, others are not, such as

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

ability. Indeed the ability is not directly observed, but measured by some instrument like the pre-test.
Two main methodological approaches have been considered in the literature to deal with the estimation of the effect of a given covariate when pre-test measures of the outcome are available (Kim and Steiner, 2019): the conditioning approach and the gain score approach. The *conditioning* approach consists in estimating the effect of interest on the post-test score, conditionally on the pre-test score. On the other hand, in the *gain score* approach the considered outcome is the gain score, namely the difference between post-test and pre-test scores.
The conditioning approach is implemented via regression models or matching on the pre-test score, relying on the unconfoundedness assumption (Arpino and Aassve, 2013), namely the pre-test score is sufficient to remove confounding. On the other hand, the gain score approach is related to difference-in-difference methods, which are devised to remove the effect of unobservable confounders under the assumption that such confounders have a time-invariant effect, known as *common trend assumption*. In such a case, taking the first difference of the outcome removes confounding (e.g. Lechner, 2011).
Recently, Kim and Steiner (2019) reconsidered the choice between the conditioning and gain score approaches. In particular, they derive analytical results for a linear model without random effects. The treatment variable Z affects the post-test score Y, while an unobservable ability A affects both Z and Y. Thus, A is an unobserved confounder. In addition, the ability A affects the pre-test score P. If P is measure of A with high reliability (Cronbach alpha), conditioning on P removes most of the confounding effect of A. On the other hand, a low pre-test reliability suggests to prefer the gain score approach, which is not affected by the reliability. However, the gain score approach relies on the validity of the common trend assumption. Kim and Steiner (2019) derive formulas for the bias of the estimated effect of Z on Y under the two approaches, highlighting the assumptions required for unbiasedness.
In this contribution, we compare the conditioning and gain score approaches in a more complex setting with hierarchical data using a random intercept linear model. Specifically, we consider students (level 1 units) nested within schools (level 2 units), where ability, pre-test and post-test scores are level 1 variables, while the treatment is a level 2 binary variable (public vs private school). Moreover, we investigate through a simulation study the performances of the estimators in terms of bias.

2 Simulation study: main results

For a treatment at individual level, our results confirm the literature findings for a single-level setting. Specifically, the conditioning approach gives

a biased estimator of the treatment effect whenever the pre-test is affected by measurement error, though the bias disappears if the pre-test and post-test scores are affected by a common source of error of the same magnitude. As a consequence, designs with the same instrument at pre-test and post-test should be preferred as they help reducing the bias. The gain score approach provides an unbiased estimator if the common trend assumption holds at the individual level, regardless of the assumption holding at cluster level. For an individual-level treatment, including the cluster mean of the pre-test score as a regressor is not recommended, as it introduces further measurement error without reducing the bias. On the other hand, the findings for a treatment at cluster level are different because the cluster mean of the latent ability acts as a confounder. Thus, its observable counterpart, namely the cluster mean of the pre-test score, should be inserted as a regressor. However, this is not always sufficient to completely eliminate the bias, because the cluster mean of the pre-test is affected by measurement error. The issue may be relevant with small clusters (e.g., size 4 in our simulation study). Anyway, also in this context using the cluster mean as a regressor is generally convenient because it reduces the bias. Moreover, it is worth noting that, if the cluster mean of the pre-test score is used as a regressor, then the conditioning and gain score approaches provide the same estimates of the treatment effect, regardless of the cluster size.

3 Case study

We aim at evaluating the effect of the Italian lower secondary schools on student achievement measured by Invalsi tests, focusing on the differences between public and private schools. To this end, we alternatively apply the conditioning and the gain score approaches, outlined in Section 1.
The data set collects information on a cohort of students that participated in the Italian language and mathematics Invalsi tests at grades 5th and 8th (i.e., the last year of the primary school and the last year of the lower secondary school, respectively). The data set has been obtained by merging data on students who attended the 5th grade in school year 2013-2014 with data on students who attended the 8th grade in school year 2016-2017. We retain data on students present in both occasions. The resulting data set consists of $436,889$ students who took part on both occasions: $427,950$ participated in both occasions of the language test, $427,256$ participated in both occasions of the math test. A subset of $418,330$ students participated in both occasions of both tests.
The students are nested in $5,777$ Italian schools. The average number of tested students per school is 103.91 with a standard deviation of 54.97 (min $= 1$; max $= 334$).
Each of the two achievement tests is composed of a set of items measuring the unobservable ability in language and mathematics, respectively. Items

are dichotomously scored, with value 1 for a correct answer and value 0 for a wrong answer. The selection of the set of items relies on internationally validated methods based on the Rasch model (Rasch,1960). For this reason, the ability level of a student is measured by the raw score (i.e., the total number of correct answers to the test items). As the number of items is different across subject areas (language and mathematics) and grades, we divide the raw scores by their maximum so that they are normalised in the range 0-100.
Several background variables are available both at student and school levels. Student covariates include gender, citizenship, and marks in language and mathematics resulting from the school reports. Data also include information about the parents educational level and job condition, which are exploited by Invalsi to define an index of the socio-economic status. In addition, a wide set of indicators measured at the end of the 5th grade provides information on student material deprivation, motivation and interest in learning, and relations with the class mates. School characteristics include information on the geographical location (municipality, urban area, altimetric area, and population density), the average number of students per class and the type of school (public vs private). Other school-level variables are obtained by averaging the student level characteristics (e.g., proportion of immigrants per school).
We specify a multilevel model (Goldstein, 2010) with students at level 1 and schools at level 2. In order to compare the conditioning and the gain score approaches, we specify two versions of the model. In the first version, the response variable is the post-test score (8th grade test), while the pre-test score enters as a covariate. In the second version, the response variable is the gain score (difference between the 8th and 5th grade tests), while the pre-test score is omitted from the covariates. Both versions of the model include the treatment variable, that is the indicator of the type of school (public vs private), as well as student and school characteristics.
The estimated effect of private school in the final model is 0.69 ($s.e.0.246$) in the conditioning approach and 0.67 ($s.e.0.250$) in the gain score approach. This effect is significant and positive, but it is quite small, given that the test score has an average of 68.5 with a standard deviation of 16.3.

References

Allison, P.D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, **20**, 93 – 114.

Arpino, B. and Aaassve, A. (2013). Estimating the causal effect of fertility on economic wellbeing: data requirements, identifying assumptions and estimation methods. *Empirical Economics*, **44**, 355 – 385.

Goldstein, H. (2010). *Multilevel Statistical Models*. 4th Edition, Wiley.

Kim, Y. and Steiner, P. M. (2021). Gain scores revisited: a graphical models perspective. *Sociological Methods & Research*, **50(3)**, 1353 – 1375.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, **4**, 165 – 224.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

A novel semi-supervised learning approach for maritime lithium-ion battery monitoring

Clara Bertinelli Salucci[1], Azzeddine Bakdi[2], Ingrid Kristine Glad[1], Erik Vanem[13], Riccardo De Bin[1]

E-mail for correspondence: clarabe@math.uio.no

Abstract: We propose a novel semi-supervised learning method to monitor the State of Health of lithium-ion batteries, a prominent technology for the electrification of the transport sector. Our approach enables State of Health monitoring of batteries with no labeled data, starting from a minimal set of labeled data from another similar battery. This can be achieved by exploiting the relation between a pseudo-capacity measure and the total capacity of the labeled data. Our results with operational data from maritime batteries show that the approach is valid and can lead to significant progress in failure prevention, operational optimization, and for planning batteries at the design stage.

Keywords: Semi-supervised learning; Multivariable Fractional Polynomials; Li-ion battery State of Health.

1 Motivating Problem and Data Description

Monitoring of the State of Health (SoH) of lithium-ion (Li-ion) batteries is crucial for maritime applications. In fact, over time and over usage Li-ion batteries undergo ageing mechanisms that ultimately lead to battery failure; the consequences for a vessel at sea can be potentially catastrophic, therefore it is compelling to assess the battery conditions with good accuracy.
One way of quantifying the SoH is based on the degradation of the battery capacity (Vanem et al., 2021):

$$\mathrm{SoH}_i = \frac{C_{\mathrm{available}}}{C_{\mathrm{nominal}}} \times 100\,(\%), \tag{1}$$

where $C_{\mathrm{available}}$ is the actually available capacity, and C_{nominal} is the nominal capacity of the battery. However, estimating the capacity itself is often a

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

demanding exercise. It is common practice in the maritime field to carry out annual tests, that enable an estimate of the battery capacity. Such tests, however, are burdensome and time-consuming, and provide very sparse capacity assessments. Further, as they are performed under different conditions (temperature, durations, etc.), they can hardly be related to each other, resulting to some extent inaccurate. An efficient method for continuous battery diagnostic, thus, is strongly needed. Data-driven methods can be greatly advantageous as they are agnostic to the real and highly complex physical problem, they are ductile and can be used for different batteries (Vanem et al., 2021).
Operating data for this analysis are provided by a leading supplier of energy storage systems for maritime applications. The data pertain to the battery systems of three vessels, which we regard as three different datasets. For each of them, we have high-frequency sensor data: temperature, voltage and current intensity measurements, together with the battery State of Charge (SoC) which we regard as sensor data insofar as it is provided by the company with good accuracy. Such variables are continuously measured from the beginning of operation until 4.5 or 5.5 years later, though there are periods of missing data in all datasets.
Our minimal set of labeled data consists of three data-points for one of the vessels, obtained from three SoH tests conducted in years that are not necessarily consecutive. Our approach is based on relating the discharge phases of the batteries while they age over years: thus, we pre-process the data to go from continuous measurements series to single events, the discharge cycles, identified on the basis of changes of sign in the SoC derivative.

2 Semi-Supervised Learning Methodology

We will refer to the three datasets according to the following:

- Reference data: data from the dataset with the three labeled data-points (vessel A);
- Target data: data from the other two datasets (vessels B and C).

Consequently, the labeled cycles from vessel A will be called *reference cycles*, while all cycles from the other two vessels are *target cycles*; our aim is to predict the total capacity of the battery at the target cycles.
Our approach relies on a fundamental assumption: the SoH can be considered constant for a time window around the day where the measurement was taken. The assumption is valid as the SoH is known to degrade gently and almost linearly in the first years of operations, after an initial short stage in which the degradation is more pronunced, and before a final stage where the decay is faster and non-linear (Edge et al., 2021). This enables us to enlarge the set of reference cycles. The method develops in three steps:

1. Cycle classification: using a tree-like classification, cycles with similar characteristics are grouped together, and each class is treated independently. This is important to account for the large impact that different conditions (temperature, SoC range, C-rate, etc.) have on the estimated capacity. At the end of this step, the data are organised in tables containing cycles with similar characteristics. An example is provided in Table 1, where the first two cycles are from the reference dataset, and hence they have an estimated SoH, and three other similar cycles in datasets B and C have been matched.

2. Model training: using the reference cycles, we train a linear model in each class. The total capacity of the battery as from the SoH test is our dependent variable; the number of features entering the model depends on how many reference cycles we have in the considered class. In all cases we input the *pseudo-capacity*,

$$\tilde{C} = \int_{t_{\text{start}}}^{t_{\text{end}}} I(t)\,\mathrm{d}t; \tag{2}$$

   optionally, cycle characteristics such as duration, initial time, variance in the C-rate and temperature etc. are also included in the model. We discard all classes having models with $R^2 < 0.6$.

3. Total capacity estimation: in each class, we get capacity estimates for all target cycles from the model trained at step 2.

The capacity estimates from different classes are then gathered together and converted to the SoH scale. In real applications it is often convenient to have weekly or monthly SoH estimates, therefore we do a weighted average of the estimations where the weights are the reciprocal of the uncertainties estimated by the model. This is done in order to ensure that highly uncertain estimates contribute very little to the final estimate.

TABLE 1. Example of a few cycles from the same class: the first two rows are cycles from the reference dataset, and hence they have an estimated SoH. Other three similar cycles in datasets B and C have been matched and are a target for capacity estimation.

SoC_1	SoC_2	avg_cRate	max_temp	min_temp	SoH	dataset
89%	**73%**	**-0.372**	**28°**	**24°**	**92.4%**	**A**
89%	**73%**	**-0.379**	**27°**	**23°**	**92.4%**	**A**
90%	73%	-0.374	27°	24°	–	B
89%	72%	-0.379	27°	23°	–	B
89%	71%	-0.377	27°	24°	–	C

3 Multivariable Fractional Polynomials for SoH modelling

The semi-supervised approach provides SoH estimates which transform the large unlabelled datasets into training data for modelling the battery degradation: the Multivariable Fractional Polynomials (MFP) approach (Sauerbrei and Royston, 1999) has been chosen for the purpose, in view of the encouraging results achieved on lab data in a previous work (Bertinelli Salucci et al., 2022). The response variable of the model is the monthly change in the battery SoH with respect to the initial value $\mathrm{SoH}(t_0)$,

$$y = \Delta\mathrm{SoH}(t) = \mathrm{SoH}(t_0) - \mathrm{SoH}(t), \tag{3}$$

while the set of candidate covariates is derived from the battery sensor data for all charge and discharge cycles, including a few significative interaction terms. All features are cumulative over each month (e.g. sum of durations of charge or discharge phases, average C-rates, ...), except for the *equivalent full cycles* measure (efc) which is cumulative over the whole history of the battery system. The MFP algorithm selects the most suitable polynomial transformations of the covariates among a set of possible choices, and variable selection is also performed (significance level $P < 0.05$) to achieve potential variance reduction and ease the model interpretability. The regression model has been trained on data from vessel B and tested on vessel C.

4 Results and Conclusions

Monthly averaged results obtained with the semi-supervised learning approach are shown in Figure 1 and Figure 2 for the two target ships. The unavailability of frequent and reliable labels makes it difficult to provide a specific accuracy assessment for the method; however, our results are in line with the typical degradation patterns of Li-ion batteries depicted by Edge et al. (2021), as well as with battery experts' expectations.
The left panel of Figure 3 shows the results obtained in predicting the SoH degradation of vessel C with the MFP model trained on data from vessel B (Table refbertinellisalucci:tab2). The plot confirms the effectiveness of MFP regression for modelling SoH degradation of lithium-ion batteries: the predicted values are all very close to the estimates obtained with the semi-supervised approach, with a normalized Root Mean Squared error of 0.85%. The right panel of the figure presents an histogram of the normalised absolute error: most of the errors are below 1.5%, and all errors below 2%.

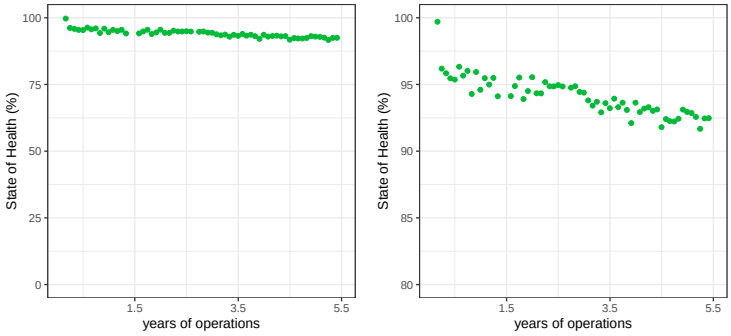


FIGURE 1. Montly averaged SoH estimates for vessel B on full scale (left) and reduced scale 80-100% (right).

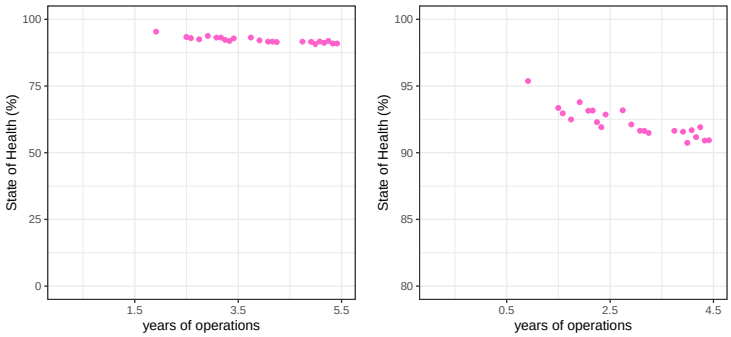


FIGURE 2. Montly averaged SoH estimates for vessel C on full scale (left) and reduced scale 80-100% (right).

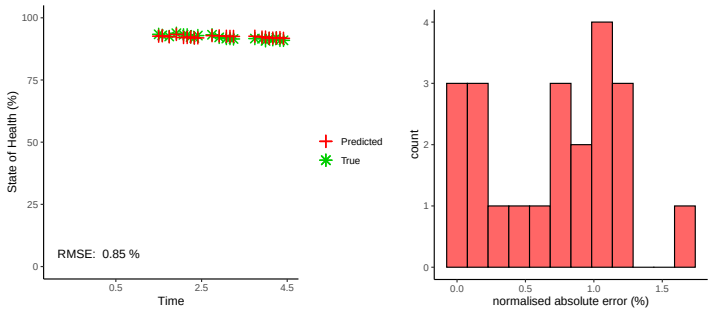


FIGURE 3. Left: State of Health degradation results for vessel C using the MFP regression model. Right: histogram of the normalised absolute error.

TABLE 2. MFP regression model trained on data from vessel B. The features entering the model after the variable selection mechanism are reported together with their estimated coefficients, standard errors and corresponding p-values: efc is a measure of the equivalent full cycles of the battery; $V_{\text{in,disch.}}$ is the average initial voltage of the discharges cycles in one month; $T_{\text{min,3}}$ and $T_{\text{min,1}}$ are the monthly averages of the minimum values of two temperature sensors.

	est. coefficient	**std. error**	***p*-value**
Intercept	-6.395	1.62	0.0002
$efc/10^5$	83.31	20.91	0.0002
$V_{\text{in,disch.}} : efc/10^5$	-16.53	4.42	0.0005
$T_{\text{min,3}}/10$	-26.19	8.25	0.0025
$T_{\text{min,1}}/10$	30.14	8.93	0.0014

References

Bertinelli Salucci, C., Bakdi, A., Glad, I.K., Vanem, E., and De Bin, R. (2022). Multivariable Fractional Polynomials for lithium-ion batteries degradation models under dynamic conditions. *Journal of Energy Storage*, **52**, 104903.

Edge, J., O'Kane, S., Prosser, R., Kirkaldy, et al. (2021). Lithium Ion Battery Degradation: What you need to know. *Physical Chemistry Chemical Physics*, **23**.

Sauerbrei W. and Royston P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **162**, 71-94.

Vanem, E., Bertinelli Salucci, C., Bakdi, A., and Alnes, Ø. Å. (2021). Data-driven state of health modelling—A review of state of the art and reflections on applications for maritime battery systems. *Journal of Energy Storage*, **43**, 103158.

Bernoulli-Gaussian model for dynamic sparsity in time varying parameter regression

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

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

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

Abstract: Time-varying parameter models are widely used in statistics for the analysis of dynamical systems. In such a models the risk of over-parametrization is high, thus perform model selection and achieve sparse estimates is a desired property. The latter is defined over two directions: vertical, where we look at the parameter vector at a fixed time, and horizontal, where we focus on a given variable and observe its behaviour across the timeline. In this paper, we tackle the estimation within a Bayesian framework and we extend the Bernoulli-Gaussian model for variable selection to deal with time varying sparsity. We assume a time dependence both in the dynamic of the regression coefficients and in the inclusion probabilities. We propose a variational Bayes approach for joint parameter estimation and signal extraction that relies on a global flexible representation of the latent states through a non-stationary Gaussian Markov random field.

Keywords: Bernoulli-Gaussian model; Dynamic sparsity; Variational Bayes.

1 Bayesian TVP regression with dynamic sparsity

The time-varying parameter regression model with variable selection can be expressed through the Bernoulli-Gaussian specification in the fashion of Ormerod et al. (2017):

$$y_t = \mathbf{x}_t^{\mathrm{T}} \boldsymbol{\Gamma}_t \boldsymbol{\beta}_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{N}(0, \sigma_t^2), \tag{1}$$

where y_t is the response and $\mathbf{x}_t$ is a set of covariates associated to the time-varying parameter $\boldsymbol{\beta}_t \in \mathbb{R}^p$. Moreover, each coefficient $\beta_{j,t}$ can be either included or not depending on the value of the diagonal elements id $\boldsymbol{\Gamma}_t$, namely $\gamma_{j,t} \in \{0, 1\}$. To account for a time dependence structure,

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

we assume a random walk dynamic for $\beta_{j,t}$ and for the logarithm of the variance $h_t = \log \sigma_t^2$:

$$\beta_{j,t} = \beta_{j,t-1} + v_{j,t}, \qquad v_{j,t} \sim \mathsf{N}(0, \eta_j^2), \qquad \beta_{j,0} \sim \mathsf{N}(0, k_0 \eta_j^2), \tag{2}$$

$$h_t = h_{t-1} + w_t, \qquad w_t \sim \mathsf{N}(0, \nu^2), \qquad h_0 \sim \mathsf{N}(0, k_0 \nu^2). \tag{3}$$

The latter formulation is equivalent to consider a Gaussian Markow random field (GMRF) for $\boldsymbol{\beta}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \eta_j^2 \mathbf{Q}^{-1})$ and $\mathbf{h} \sim \mathsf{N}_{n+1}(\mathbf{0}, \nu^2 \mathbf{Q}^{-1})$, where the matrix $\mathbf{Q}$ assumes a tridiagonal structure. The indicator variables $\gamma_{j,t}$ are assumed to be independent Bernoulli $\gamma_{j,t}|\omega_{j,t} \sim \mathsf{Bern}(p_{j,t})$ given the parameters $\omega_{j,t}$, where $\omega_{j,t} = \text{logit}(p_{j,t})$. Ročková and McAlinn (2021) assume a deterministic dynamic for each $p_{j,t}$, while Koop and Korobilis (2020) place an independent prior distribution on $\omega_{j,t}$. An important contribution of this work is that we assume a stochastic process for the inclusion probabilities. Similarly as for $\boldsymbol{\beta}_j$, we assume a GMRF specification for $\boldsymbol{\omega}_j \sim \mathsf{N}_{n+1}(\mathbf{0}, \xi^2 \mathbf{Q}^{-1})$. To complete the Bayesian model specification we place the following prior distributions for the variances parameters $\nu^2 \sim \mathsf{IG}(A_\nu, B_\nu)$, $\eta_j^2 \sim \mathsf{IG}(A_\eta, B_\eta)$, and $\xi_j^2 \sim \mathsf{IG}(A_\xi, B_\xi)$. The joint distribution of the data and the high-dimensional parameter vector $\boldsymbol{\vartheta} = (\mathbf{h}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\omega}^\top, \nu^2, \boldsymbol{\eta}^{2\top}, \boldsymbol{\xi}^{2\top})^\top$ can be written as the following product:

$$p(\mathbf{y}, \boldsymbol{\vartheta}) = p(\mathbf{y}|\boldsymbol{\vartheta}) p(\mathbf{h}) p(\nu^2) \prod_{j=1}^{p} p(\boldsymbol{\beta}_j|\eta_j^2) p(\boldsymbol{\gamma}_j|\boldsymbol{\omega}_j) p(\boldsymbol{\omega}_j|\xi_j^2) p(\eta_j^2) p(\xi_j^2), \tag{4}$$

where $p(\boldsymbol{\gamma}_j|\boldsymbol{\omega}_j) = \prod_{t=1}^{n} p(\gamma_{j,t}|\omega_{j,t})$ also factorizes over time. In order to simplify the computations, we exploit the Polya-Gamma representation of $p(\gamma_{j,t}|\omega_{j,t}) = \int_0^{+\infty} p(\gamma_{j,t}|z_{j,t}, \omega_{j,t}) p(z_{j,t}|\omega_{j,t})\, dz_{j,t}$, where $p(z_{j,t})$ is the density function of a Polya-Gamma $\mathsf{PG}(1, 0)$. Now we can consider an augmented version of (4) which has the advantage of recognizing all the full conditionals as known distribution functions. Instead of traditional MCMC estimation methods, we implement a variational Bayes (VB) algorithm exploiting the mean-field approximation paradigm which consists in providing a factorization of the joint variational density q. The specification of q we propose is the following:

$$q(\boldsymbol{\vartheta}) = q(\mathbf{h}) q(\nu^2) \prod_{j=1}^{p} q(\boldsymbol{\beta}_j) p(\boldsymbol{\omega}_j) p(\eta_j^2) p(\xi_j^2) \prod_{t=1}^{n} q(\gamma_{j,t}) q(z_{j,t}), \tag{5}$$

where the main feature is the fact that we keep a joint density for $\mathbf{h}$, $\boldsymbol{\beta}_j$, and $\boldsymbol{\omega}_j$ in order to preserve the time dynamic in the regression parameters and in the inclusion probabilities. Moreover, we consider a parametric approximation for the joint vector of log-volatility with a multivariate gaussian distribution $q(\mathbf{h}) \sim \mathsf{N}_{n+1}(\boldsymbol{\mu}_{q(h)}, \boldsymbol{\Sigma}_{q(h)})$. The estimation of remaining densities q in (5) is carried out exploiting the Coordinate Ascent Variational Inference (CAVI) algorithm presented in Ormerod and Wand (2010).

2 Simulation study

Here we consider 100 replicates from the following data generating process: $y_t = \mathbf{x}_t^\intercal \boldsymbol{\Gamma}_t \boldsymbol{\beta}_t + \varepsilon_t$, with $\varepsilon_t \sim \mathsf{N}(0, 0.16)$, for $t = 1, \ldots, 180$, where the entries of $\mathbf{x}_t$ are independently generated from a standard Gaussian. The dimension of the regression parameter $\boldsymbol{\beta}_t$ is equal to $p = p_1 + p_{01} + p_0 = 50$ where $p_1 = 1$ is the number of parameters always included (β_1), $p_{01} = 4$ is the number of coefficients which can be included or not at each time t ($\beta_{2:5}$), and $p_0 = 45$ is the number of parameters that are always zero ($\beta_{6:50}$). The p_1 parameters are generated from an AR(1) with unconditional mean far from zero, $\phi_1 = 0.98$ and conditional variance equal to 0.1. For the p_{01} parameters we proceed as follows: divide the interval in sub-periods $[1, 180] = [1, t_1] \cup [t_1 + 1, t_1 + t_2] \cup \ldots \cup [t_1 + \ldots + t_n + 1, 180]$, where $t_k \sim \mathsf{Pois}(60)$, and then alternate periods where $\gamma_{j,t} = 0$ and $\gamma_{j,t} = 1$ starting randomly. For the intervals where $\gamma_{j,t} = 1$ we generate a process as for p_1. We compare our method (BGTVP), the dynamic variable selection

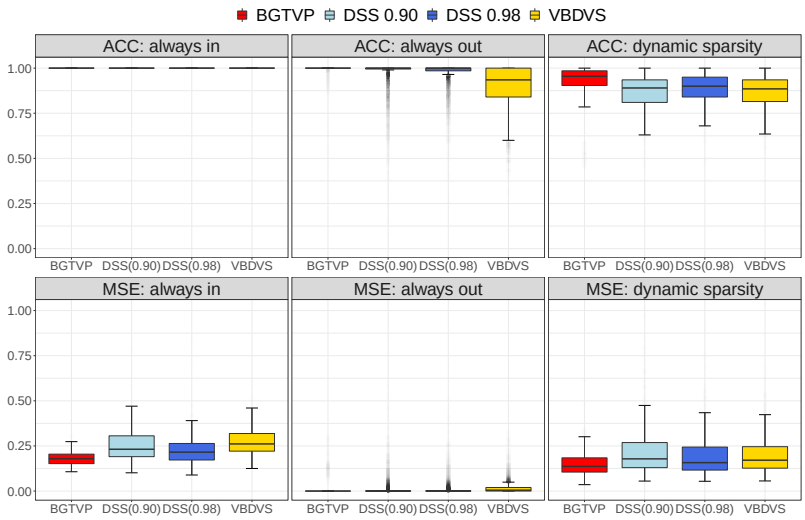


FIGURE 1. Simulation results. We compare the strategies in terms of ACC (top) and MSE (bottom). We provide different panels for different behavior of $\boldsymbol{\beta}_j$.

(VBDVS) of Koop and Korobilis (2020), and the dynamic spike-and-slab (DSS) of Ročková and McAlinn (2021) for two different values of marginal importance weight $\Theta = \{0.90, 0.98\}$. We look both at the mean squared error (MSE) and at the classification accuracy (ACC) to assess the capability of each approach to distinguish true signal from the noise. The results are depicted in Figure 1, both for all p and separately for p_1, p_{01} and p_0. The different methods have a perfect accuracy in recognize $\boldsymbol{\beta}_j$ that are always

included in the true model. Our approach outperforms the competitors in determine whether a coefficient is always equal to zero in its path. Looking at the results when dynamic sparsity is considered for $\boldsymbol{\beta}_j$, the performances are similar across methods. However it can e noticed that BGTVP tends to distinguish better the true signal from the noise even in this setting. Similar conclusions can be derived looking at the MSE measure.

3 Application to GDP deflator forecasting

Gross domestic product (GDP) deflator forecasting is a widely studied real-data example in the context of time-varying parameter models (see Kalli and Griffin, 2014). The sample period ranges from the second quarter of 1965 to first quarter of 2011. We consider as predictors 31 exogenous variables and first three lags of the response variable in order to account for an autoregressive behaviour. More details on the data are available in Appendix B of Kalli and Griffin (2014). Figure 2 shows the estimated stochastic variance and the in sample estimates of the GDP deflator together with its credibility intervals.

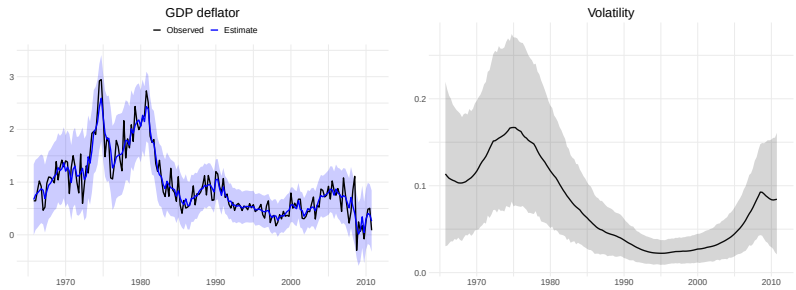


FIGURE 2. Fitted values with credibility intervals against observed values (left panel) and stochastic volatility estimate (right panel).

Figure 3 shows the estimate of the regression parameters that are selected in the model and their inclusion probabilities. The results are very sparse: in fact we select only 6 coefficients (intercept included) for the analysis. However, the results are reasonable compared to other studies, for example the variables selected by our method are a subset of those chosen in Kalli and Griffin (2014).
To conclude the real data analysis, we evaluate the out-of-sample prediction accuracy of our methodology. The latter strongly depends on the correct identification of the sparse structure. We test our method against some alternatives: a model with only a time-varying intercept (TVI), an autoregression of order 4 (AR) and its time-varying parameter counterpart (TVP-AR), the EMVS (Ročková and George, 2014) and the VBDVS (Koop

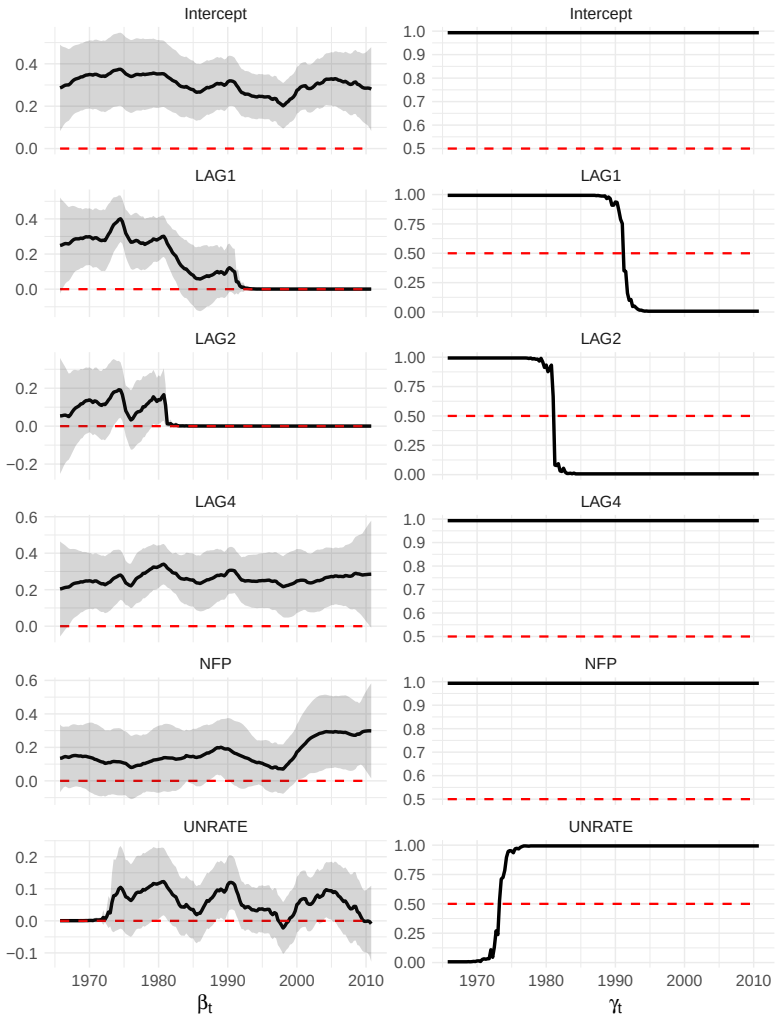


FIGURE 3. Dyanmic sparse estimate of regression coefficients $\beta_{j,t}$ with HPD 95% intervals (left) and inclusion probabilities $\gamma_{j,t}$ across time (right).

and Korobilis, 2020). We start with the first half of the sample as initial estimation period and we forecast one step ahead. Then, we expand it by adding the new observation and we repeat this procedure until the full sample is used for the estimation of the model. We measure forecast accuracy using the root mean squared error (RMSE), the mean absolute error (MAE) and the average log-predictive likelihood (ALPL). Table 3 shows the out-of-sample results. BGTVP outperforms other methods in terms of point estimates and it also shows good performances in terms of density

forecasting.

Measure	TVI	AR	TVP-AR	BGTVP	EMVS	VBDVS
RMSE	0.238	0.229	0.244	**0.223**	0.287	0.262
MAE	0.172	0.165	0.177	**0.158**	0.230	0.193
ALPL	-0.040	**0.124**	0.052	0.058	-0.395	-0.126

TABLE 1. Forecast accuracy. Best model according to each measure is highlighted.

4 Conclusion

In this paper we propose a Bayesian approach to dynamic sparsity in time-varying parameter regression with stochastic volatility. An important feature of the model specification is the time dependence structure among the logit of the inclusion probabilities, which is induced assuming a Gaussian Markow random field for the joint vector of the latent states. The inference is carried out within a Variational Bayes paradigm, which has the advantage of give rise to a fast algorithm, suitable for dealing with regressions having a large number of predictors. Our methodology is proved to outperform some established competitors both in a simulation study and in a real data forecasting application.

References

Kalli, M. and Griffin, J. (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, **178**, 779 – 793.

Koop, G. and Korobilis, D. (2020). Bayesian dynamic variable selection in high dimensions. *Working Papers*.

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

Ormerod, J.T., You, C. and Müller, S. (2017). A variational Bayes approach to variable selection. *Electronic Journal of Statistics*, **11**, 3549 – 3594.

Ročková, V. and George, E. I. (2014). EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, **109**, 828 – 846.

Ročková, V. and McAlinn, K. (2021). Dynamic Variable Selection with Spike-and-Slab Process Priors. *Bayesian Analysis*, **16**, 233 – 269.

Bivariate mixed binary-survival additive regression modelling

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

[1] TU Dortmund University, Germany

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

Abstract: We propose a bivariate additive regression model where the response is given by a binary outcome and a right-censored survival time modelled using the piecewise-exponential approach. A joint bivariate density is constructed using parametric copulae, allowing for separate specification of the dependence structure and marginal distributions. Model fitting is carried using trust-regions implemented as a custom extension of the `R` package `GJRM`.

Keywords: Copula; GAMLSS, Survival analysis; Distributional regression; Piecewise-exponential model.

1 Motivation

Consider the i-th observation of a bivariate response vector (y_{1i}, y_{2i}), with $i = 1, \ldots, n$. This response is comprised of a binary outcome $y_{1i} \in \{0, 1\}$ and a continuous, right-censored survival time $y_{2i} \in \mathbb{R}_+$ with censoring indicator $\delta_i = \mathbf{1}(y_{2i} \leq C_i)$, where C_i denotes the (random, non-informative) censoring time and $\mathbf{1}$ denotes the indicator function. We propose a bivariate regression model that accommodates a piecewise-exponential model (PEM) of the survival margin, as well as a non-survival response. The proposed approach allows for simultaneous modelling of very flexible hazard rates as well as an additive model of a binary response. Such a bivariate model is not possible without the modifications developed here due to the following issue: Consider the right-censored survival marginal response. In the PEM approach, the log-likelihood function of the survival model coincides with the Poisson log-likelihood (with given offset, see, e.g., Bender et al., 2018). The follow-up time is partitioned into J intervals $(\kappa_{j-1}, \kappa_j]$, where $\kappa_0 < \cdots < \kappa_j$ are cut points. Thus, the baseline hazard rate is assumed to be constant within each interval. In order to fit a PEM, the original dataset

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

needs to be suitably augmented. Consider two hypothetical observations with survival times $t_1 = 0.5$ (censored) and $t_2 = 2.7$ (not censored) as well as cut points at $\kappa_0 = 0$, $\kappa_1 = 1.5$ and $\kappa_2 = 3$ (i.e. $J = 2$ intervals).

TABLE 1. Comparison of non-augmented and augmented data for a PEM.

original data				$\Longrightarrow$	augmented data					
i	t_i	δ_i	$\mathbf{x}_i$		i	j	δ_{ij}	t_{ij}	$\log(t_{ij})$	$\mathbf{x}_{ij}$
1	0.5	0	$\mathbf{x}_1$		1	1	0	0.5	$\log(0.5)$	$\mathbf{x}_{11}$
2	2.7	1	$\mathbf{x}_2$		2	1	0	1.5	$\log(1.5)$	$\mathbf{x}_{21}$
					2	2	1	1.2	$\log(1.2)$	$\mathbf{x}_{22}$

The data augmentation for a PEM is exemplified in Table 1 for the two hypothetical observations $i = 1, 2$, where the index j denotes the j-th discrete time interval. The new binary indicator δ_{ij} is equal to 1 if the observation i is not censored in time interval j and 0 if it is censored. As shown in Table 1, the augmented data for a PEM consists now of n' rows instead of the original n rows, since the data entry for observation i now consists of $j = 1, \ldots, j(i)$ entries, depending on the discrete time intervals and the i-th subject's survival time, with $j(i)$ denoting the index of the event or censoring time interval, i.e. for which $t_i \in (\kappa_{j(i)-1}, \kappa_{j(i)}]$. If there are p_2 covariates in the model of the survival margin (PEM), the hazard rate ϑ_{2ij} for observation i in interval j is written as

$$\log(\vartheta_{2ij}) = \eta_{2ij} = \underbrace{\beta_0 + s_0(t)}_{\text{log-baseline hazard}} + \sum_{r=1}^{p_2} s_r(t, \mathbf{x}_i) + o_{ij} \;\; \forall t \in (\kappa_{j-1}, \kappa_j],$$

where the term $o_{ij} = \log(t_{ij})$ is a given offset. The functions $s_r(\cdot)$ are some smooth functions of the covariates that can be compactly written as a combination of a design matrix and a vector of regression coefficients, e.g. $\eta_{2ij} = \mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2$. The index "2" in the hazard rate and linear predictor denotes that the survival response is the second marginal response of the bivariate model. The data augmentation for the PEM results in a design matrix $\mathbf{X}_2$ of dimensions $(n' \times p_2)$ with $n' > n$. In case of the model for the binary marginal response y_{1i}, the dataset is left un-augmented and its linear predictor is given by $\eta_{1i} = \mathbf{x}_{1i}^\top \boldsymbol{\beta}_1$. Assuming there are p_1 covariates, the design matrix $\mathbf{X}_1$ is of dimensions $(n \times p_1)$. A bivariate regression model using both augmented and un-augmented data cannot be directly specified due to the mismatch in the dimensions of the binary and piecewise-exponential margins, i.e. $n' > n$. In the following, we derive a cumulative distribution function (CDF) based on the PEM such that both margins are conform and a bivariate model with the aforementioned specifications can be fitted.

2 Methodology

Assume $\delta_{ij} \in \{0,1\} \sim Poisson(\vartheta_{ij})$ and denote the sequence of $j(i)$ independent censoring indicators of the i-th observation by $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{ij(i)})^\top$. Given these assumptions, the joint density of $\boldsymbol{\delta}_i$ turns out to be:

$$f_2(\boldsymbol{\delta}_i) = \prod_{j=1}^{j(i)} \underbrace{\frac{\vartheta_{ij}^{\delta_{ij}} \exp(-\vartheta_{ij})}{\delta_{ij}!}}_{\delta_{ij} \sim Poisson(\vartheta_{ij})} = \begin{cases} \underbrace{\exp\left(-\sum_{j=1}^{j(i)} \vartheta_{ij}\right)}_{=: f_2(\boldsymbol{\delta}_i = 0)}, & \text{if } \delta_{ij(i)} = 0, \\ \underbrace{\exp\left(-\sum_{j=1}^{j(i)} \vartheta_{ij}\right) \vartheta_{ij(i)}}_{=: f_2(\boldsymbol{\delta}_i = 1)}, & \text{if } \delta_{ij(i)} = 1. \end{cases} \tag{1}$$

The CDF is $F_2(\boldsymbol{\delta}_i = 0) = f_2(\boldsymbol{\delta}_i = 0)$ and $F_2(\boldsymbol{\delta}_i = 1) = f_2(\boldsymbol{\delta}_i = 0) + f_2(\boldsymbol{\delta}_i = 1)$. The index "2" once again denotes that the PEM is the second marginal response. Equation (1) summarises the information of the augmented data, turning $j(i)$ entries per subject i into a single data entry. Hence, a likelihood function, which is based on the bivariate density presented in Marra et al. (2020),

$$L_i(\boldsymbol{\beta} | y_{1i}, \boldsymbol{\delta}_i) = \left[C\Big(F_1(0), F_2(\boldsymbol{\delta}_i)\Big) - C\Big(F_1(0), F_2(\boldsymbol{\delta}_i) - f_2(\boldsymbol{\delta}_i)\Big) \right]^{1-y_{1i}} \cdot \left[f_2(\boldsymbol{\delta}_i) - C\Big(F_1(0), F_2(\boldsymbol{\delta}_i)\Big) + C\Big(F_1(0), F_2(\boldsymbol{\delta}_i) - f_2(\boldsymbol{\delta}_i)\Big) \right]^{y_{1i}},$$

can be computed, where $C(F_1(\cdot), F_2(\cdot))$ is a parametric copula evaluated at the respective marginal CDFs of y_{1i} and $\boldsymbol{\delta}_i$, as well as a dependence parameter ϑ_{3i}. The bivariate density in the likelihood function $L_i(\cdot)$ depends on the vector of unknown regression coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)^\top$ and thus on the distribution parameters $\boldsymbol{\vartheta}_i = (\vartheta_{1i}, \vartheta_{2ij}, \ \vartheta_{3i})^\top$, which are specified using the Generalized Additive Model for Location, Scale and Shape (GAMLSS; Stasinopoulos et al., 2018) framework. Generally, the bivariate response is assumed to be a draw from a parametric distribution made up of $k = 1, \ldots, K$ parameters ϑ_{ki}, i.e. $(y_{1i}, y_{2i}) \sim f_{12}(y_{1i}, y_{2i} | \boldsymbol{\vartheta}_i)$. Given a vector $\mathbf{x}_i$ of $r = 1, \ldots, p$ covariates, each parameter is modelled by combining a structured additive predictor η_{ki} and a link function $g_k(\cdot)$:

$$\vartheta_{ki} = g_k^{-1}(\eta_{ki}) \quad \Leftrightarrow \quad g_k(\vartheta_{ki}) = \eta_{ki} = \beta_{k0} + \sum_{r \in L_k} s_{kr}(x_{ir}),$$

here $L_r \subseteq \{1, \ldots, p\}$ indicating that each parameter can be modelled using different subsets of covariates. The smooth functions $s_{kr}(\cdot)$ feature different specifications to accommodate e.g. linear, non-linear or spatial functional forms of the regressors. Similarly to the predictor of the hazard rate, these smooth functions can be written as a combination of a design matrix and a vector of coefficients, i.e. $\eta_{1i} = \mathbf{x}_{1i}^\top \boldsymbol{\beta}_1$ and $\eta_{3i} = \mathbf{x}_{3i}^\top \boldsymbol{\beta}_3$. In the present case $K = 3$, and the distribution parameters are obtained using the respective inverse link functions, namely: $\vartheta_{1i} = g_1^{-1}(\eta_{1i})$, $\vartheta_{2ij} = \exp(\eta_{2ij})$, and $\vartheta_{3i} = g_3^{-1}(\eta_{3i})$. While typically logit and probit functions are used for $g_1(\cdot)$, the link $g_3(\cdot)$ depends on the chosen parametric copula. This means that it is also possible to specify a flexible additive model for the copula dependence parameter. The link employed for the PEM is the natural logarithm. The data augmentation required for the PEM in the survival margin also produces a dimension mismatch in the remaining components required for optimisation of the log-likelihood function $\ell_i(\cdot)$, i.e. score vector and information matrix. These issues are also addressed within this work in order to carry out estimation via trust-regions or other optimisation methods based on first and second order derivatives of the objective function. For instance, let the score vector that contains the first order partial derivatives of the log-likelihood w.r.t. the coefficient vectors be defined as:

$$\mathbf{s}_i(\boldsymbol{\beta}) = \left(\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}, \quad \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2}, \quad \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3} \right).$$

The second entry in the score requires the first order partial derivatives of the previously defined PDF and CDF of the survival margin w.r.t. the coefficient vector $\boldsymbol{\beta}_2$. Recall that $\vartheta_{2ij} = g_2^{-1}(\eta_{2ij}) = \exp\left(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2\right)$, then:

$$\frac{\partial f_2(\boldsymbol{\delta_i} = 0)}{\partial \boldsymbol{\beta_2}} = -\exp\left(-\sum_{j=1}^{j(i)} \exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2) \right) \cdot \sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2) \mathbf{x}_{2ij} \right],$$

$$\frac{\partial f_2(\boldsymbol{\delta_i} = 1)}{\partial \boldsymbol{\beta_2}} = \exp\left(-\sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2) \right] + \mathbf{x}_{2ij(i)}^\top \boldsymbol{\beta}_2 \right) \cdot \left[-\sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2) \mathbf{x}_{2ij} \right] + \mathbf{x}_{2ij(i)} \right].$$

The partial derivatives of the second marginal CDF are equal to:

$$\frac{\partial F_2(\boldsymbol{\delta_i} = 0)}{\partial \boldsymbol{\beta}_2} = \frac{\partial f_2(\boldsymbol{\delta_i} = 0)}{\partial \boldsymbol{\beta}_2},$$

$$\frac{\partial F_2(\boldsymbol{\delta_i} = 1)}{\partial \boldsymbol{\beta}_2} = \frac{\partial f_2(\boldsymbol{\delta_i} = 0)}{\partial \boldsymbol{\beta}_2} + \frac{\partial f_2(\boldsymbol{\delta_i} = 1)}{\partial \boldsymbol{\beta}_2}.$$

The matrix of second partial derivatives w.r.t. the unknown coefficient vectors is defined as:

$$\mathbf{H}_i(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^\top} & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_2^\top} & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_3^\top} \\ \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_1^\top} & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^\top} & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_3^\top} \\ \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3 \partial \boldsymbol{\beta}_1^\top} & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3 \partial \boldsymbol{\beta}_2^\top} & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_3 \partial \boldsymbol{\beta}_3^\top} \end{pmatrix}$$

For the blocks of the Hessian that correspond to the vector $\boldsymbol{\beta}_2$, the previous first order partial derivatives, as well as the following second order partial derivatives are used:

$$\begin{aligned} \frac{\partial^2 f_2(\boldsymbol{\delta}_i = 0)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^\top} = & \sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\mathbf{x}_{2ij}\right] \cdot \exp\left(-\sum_{j=1}^{j(i)} \exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\right) \cdot \\ & \sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\mathbf{x}_{2ij}^\top\right] \quad - \\ & \exp\left(-\sum_{j=1}^{j(i)} \exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\right) \cdot \sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\mathbf{x}_{2ij}\mathbf{x}_{2ij}^\top\right], \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 f_2(\boldsymbol{\delta}_i = 1)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^\top} = & \left[-\sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\mathbf{x}_{2ij}\right] + \mathbf{x}_{2ij(i)}\right] \cdot \\ & \exp\left(-\sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\right] + \mathbf{x}_{2ji(i)}^\top \boldsymbol{\beta}_2\right) \cdot \\ & \left[-\sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\mathbf{x}_{2ij}^\top\right] + \mathbf{x}_{2ji(i)}^\top\right] + \\ & \exp\left(-\sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\right] + \mathbf{x}_{2ji(i)}^\top \boldsymbol{\beta}_2\right) \cdot \\ & \left[-\sum_{j=1}^{j(i)} \left[\exp(\mathbf{x}_{2ij}^\top \boldsymbol{\beta}_2)\mathbf{x}_{2ij}\mathbf{x}_{2ij}^\top\right]\right]. \end{aligned}$$

The second order partial derivatives of $F_2(\cdot)$ w.r.t. $\boldsymbol{\beta}_2$ are obtained in the

same fashion as the first order partial derivatives, i.e.

$$\frac{\partial^2 F_2(\boldsymbol{\delta}_i = 0)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^\top} = \frac{\partial^2 f_2(\boldsymbol{\delta}_i = 0)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^\top},$$

$$\frac{\partial^2 f_2(\boldsymbol{\delta}_i = 1)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^\top} = \frac{\partial^2 f_2(\boldsymbol{\delta}_i = 0)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^\top} + \frac{\partial^2 f_2(\boldsymbol{\delta}_i = 1)}{\partial \boldsymbol{\beta}_2 \partial \boldsymbol{\beta}_2^\top}.$$

The corresponding information matrix for observation i is then given by $-\mathbf{H}_i(\boldsymbol{\beta})$. It should be noted that the mismatches in dimensions between the second PEM margin and the first margin as well as the model of the copula dependence parameter (i.e. $\vartheta_{3i} = g_3^{-1}(\mathbf{x}_{3i}^\top \boldsymbol{\beta}_3)$) are addressed by summing up the individual $j = 1, \ldots, j(i)$ "sub-contributions" of the PEM margin (i.e. sub-contributions of both the score vector and the information matrix).

3 Discussion

To the best of our knowledge, so far there exist no alternatives to the proposed modelling approach. This means that it is not possible to combine the augmented dataset for the very flexible piecewise-exponential model with the un-augmented dataset for the second margin (in this case, binary) as well as an additive model of the (potential) dependence between both marginal responses. The developments presented here should allow for modelling of a survival-survival outcome with both margins as a piecewise-exponential model with (possibly) different number of discrete time intervals in each margin. Implementing other combinations for the response vector such as continuous-survival, discrete-survival, or categorical-survival should be straightforward as well. The aforementioned cases are left as an open research area for future developments.

References

Bender, A., Groll, A., and Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, **18 (3-4)**, 299 – 321.

Marra, G., Radice, R., and Zimmer, D. (2020). Estimating the binary endogenous effect of insurance on doctor visits by copula-based regression additive models. *Journal of the Royal Statistical Society, Series C*, **69 (4)**, 953 – 971.

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

Locating γ-Ray Sources on the Celestial Sphere via Mixture Models

Silvia Brosolo[1], Alessandra R. Brazzale[1], Giovanna Menardi[1]

[1] Department of Statistical Sciences, University of Padova, Italy

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

Abstract: Searching for as yet undetected γ-ray sources is a major target of the Fermi LAT Collaboration. In this paper, we explore the capability of a filtering method based on robust mixtures of von Mises–Fisher distributions with the additional inclusion of concomitant variables. The proposed procedure is illustrated on data drawn from a Fermi LAT catalogue of detected sources.

Keywords: Astrostatistics; Finite mixture models; von Mises-Fisher distribution

1 Background and Motivation

In γ-ray astronomy, the data typically consist of an event list which gives the direction in the sky of each detected photon together with additional information. If the distance to the emitting source is not relevant, the data points are placed on the celestial sphere with Earth at its center and unit radius, as shown in the left panel of Figure 1. Directions are often expressed in *galactic coordinates*, which place the origin of the Cartesian system in the center of our galaxy — the Milky Way — and align the x-axis with the galactic plane (right panel of Figure 1). To overcome mismatches due to projecting data onto the 2-dimensional sky map, we rather express directions through *polar coordinates*, that is, co-latitude (θ) and longitude (ϕ) in geographical terms, which can easily be back-transformed to Cartesian coordinates $\mathbf{y} = (\cos\theta, \sin\theta\cos\phi, \sin\theta\sin\phi)^\top$ on the unit sphere.
Discovering and locating high-energy emitting sources in the whole sky map is a declared target of the Fermi Gamma-ray Space Telescope collaboration. An astronomical source is an object in outer space which, in our case, emits γ-ray photons, that is, quanta of light in the highest energy range. Traditionally, analyses are based on so-called *single-source models* (Hobson et al, 2009, § 7.4) , which require the whole sky map to be split into small

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

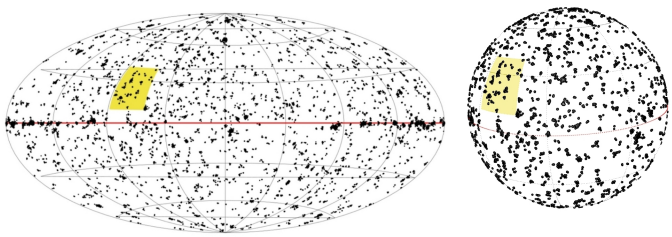


FIGURE 1. Fermi-LAT γ-ray source maps for a 5-year observation period in Galactic (left) and polar (right) coordinates. In yellow, the analysed region.

regions. The presence of a possible new source is assessed on a pixel-by-pixel basis using Poisson regression and likelihood ratio testing. Conversely, *variable-source-number models* address the problem from a more global perspective, as they simultaneously model and locate all sources in a sky map (Hobson et al, 2009, § 7.3).
In this paper, we address the problem of identifying γ-ray sources from the global perspective of variable-source-number models while working on the sphere. Sources will be represented by highly concentrated clusters, each modelled as a mixture of von Mises-Fisher distributions, with the further flexibility of introducing robustness as well as additional information on the photons, as provided by some concomitant factors. The methodological background is reviewed in Section 2 and illustrated through a case-study of Fermi LAT data in Section 3.

2 Modelling photon directions

In our setting, photon directions in $\mathbb{R}^3$ are represented as unit vectors $\mathbf{y}$, that is, as points on the sphere $\Omega^2 = \{\mathbf{y} \in \mathbb{R}^3 : ||\mathbf{y}||_2 = 1\}$ with unit radius and centre at the origin. Since sources are known to present themselves as spatially concentrated photon emissions, a natural model to address their identification, by accounting for their directional nature, mixes a number of von Mises–Fisher (vMF) densities:

$$p(\mathbf{y}) = \sum_{j=1}^{J} \alpha_j p_j(\mathbf{y}) = \sum_{j=1}^{J} \alpha_j c_3(\kappa_j) e^{\kappa_j \mu_j^\top \mathbf{y}} \tag{1}$$

with $\alpha_j > 0$ and $\sum_j \alpha_j = 1$. Here, the parameters μ_j are directly linked to the mean directions of the emitting sources and have unit norm; $\kappa_j > 0$ are concentration parameters which (inversely) characterise how widely the photon emission spreads around the mean direction of the sources, and $c_3(\cdot)$ are normalising constants which depend on the modified Bessel function of the first type.

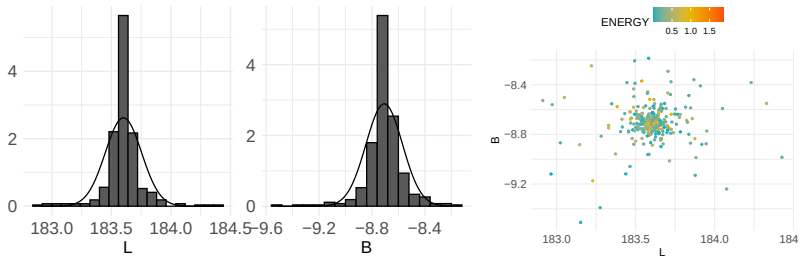


FIGURE 2. Leptokurtik marginal distributions of Galactic coordinates of photons emitted by a single source (left) and associated scatterplot highlighting that higher energy photons are concentrated around the center of the source

While specification (1) has proved a rather satisfactory fitting (Costantin et al. 2020), here, we consider to extend its flexibility to better account for data specificities. On one side, to reflect a mostly leptokurtic shape (Figure 2, left), the density of each source emission j is itself modelled by a mixture of two vMF distributions with the same location but different concentration, i.e.

$$p_j(\mathbf{y}) = \lambda_j c_3(\kappa_j) e^{\kappa_j \mu_j^\top \mathbf{y}} + (1-\lambda_j) c_3(\eta_j \kappa_j) e^{\eta_j \kappa_j \mu_j^\top \mathbf{y}}, \tag{2}$$

$0 < \lambda_j < 1$ and $\eta_j > 0$. As highlighted by Farcomeni and Punzo (2020) for the Euclidean normal setting, the specification of a contaminated distribution to model each cluster provides robustness with respect to mild outliers thanks to the induced inflated spread of one component. Identifiability is guaranteed by the specification of the same location parameter for both mixture components. The choice is also consistent with the so-called *point spread function*, which describes the response of the LAT to a point source as a function of its energy and the geometry of the detector and is modelled by a mixture of t distributions (Ackermann et al, 2013).
In addition, the availability of supplementary information on photon emission, such as the associated energy and the quality of the event reconstruction, allows us to consider its inclusion to strengthen the discrimination of the putative sources (Fig. 2, right). Denoted by $\mathbf{x} = (1, x_1, \ldots, x_p)$ the vector of concomitant factors, the mixing proportions are hence modelled as

$$\alpha_j = \alpha(\mathbf{x}; \omega_j) = \frac{\exp \mathbf{x}^\top \omega_j}{\sum_{j=1}^{J} (\exp \mathbf{x}^\top \omega_j)}, \quad \lambda_j = \lambda(\mathbf{x}; \gamma_j) = \frac{\exp \mathbf{x}^\top \gamma_j}{(1 + \exp \mathbf{x}^\top \gamma_j)}. \tag{3}$$

A suitable adjustment of the Expectation Maximisation algorithm allows us to recover the maximum likelihood estimates of the parameters, while setting the number J of mixture components can be suitably addressed by model selection criteria such as the BIC.

TABLE 1. Model validation with respect to known true sources: ARI, TPR, FPR and weighted average distance between the true and the closest detected source.

	Model (1)	Model (2)	Model (3)
ARI	0.555	0.821	0.838
TPR	0.774	0.774	0.806
FPR	0.226	0.226	0.194
$\overline{d}(\mu, \hat{\mu})$	0.105	0.068	0.209

3 Application to Fermi LAT data

The yellow region in Figure 1 shows a portion of the Northern sky of size $(l, b) \in [90°, 120°] \times [10°, 40°]$. The observations are drawn from the available 3FHL LAT catalogue and provide information on 31 earlier detected sources in the area. Table 1 shows the results obtained by fitting a simple mixture model (1), a mixture of mixtures model (2), and a mixture of mixtures models with additional information provided by the energy emitted by each photon and the quality of event reconstruction as specified in (3). With respect to the baseline specification, the robust alternative allows for a remarkable improvement of accuracy, both with respect to the detected source location (measured by the weighted angular distance from the true sources), and with respect to the ability of the fitted models to associate the events to the pertaining detected source (measured by the Adjusted Rand Index, ARI). The proportion of true sources correctly detected (TPR) and the proportion of estimated components which are not associated with a source (FPR) remain unchanged. Including the concomitant variables aids a marginal refinement of the event classification and source detection, yet at the expense of a worsened accuracy of the source location estimate. Overall, the promising results suggest further space for improvement, to be addressed by a more flexible inclusion of the additional information, possibly related to the source location rather than the mixing proportions.

References

Ackermann, M. et al. (2013). Determination of the point-spread function for the fermi large area telescope from on-orbit data and limits on pair halos of active galactic nuclei. *The Astrophysical Journal* **765**(1)

Costantin D., Menardi G., Brazzale A.R., Bastieri D., Fan J.H. (2020). A novel approach for pre-filtering event sources using the von Mises-Fisher distribution. *Astrophysics and Space Science* **365**

Farcomeni, A. and Punzo, A. (2020). Robust model-based clustering with mild and gross outliers, *TEST*, **29**, 989-1007

Hobson M.P., Jaffe A.H., Liddle A.R., Mukherjee P., Parkinson D. (2009). *Bayesian Methods in Cosmology*. Cambridge University Press

Spatial anisotropic modelling of the repeat/near repeat victimisation phenomenon

Paul T. Brown[1], Chaitanya Joshi[1], Deane Searle[2], Stephen Joe[1]

[1] Department of Mathematics, University of Waikato, Hamilton, New Zealand.
[2] Evidence Based Policing Centre, Wellington, New Zealand.

E-mail for correspondence: paul.brown@waikato.ac.nz

Abstract: We model the repeat/near repeat victimisation phenomenon, a theory often used in crime modelling, as a spatiotemporal point process. Furthermore, we assume spatial anisotropy to provide a more realistic treatment of the spatial domain. We apply this to a model of residential burglaries in the city of Hamilton, New Zealand. Model fitting and inference is provided using integrated nested Laplace approximations with the stochastic partial differential equations approach (INLA-SPDE). This not only provides a computationally efficient approach as opposed to other Bayesian methods, but also gives a more spatially continuous and anisotropic treatment of the spatial domain.

Keywords: Spatiotemporal Modelling, Repeat/Near Repeat Victimisation, Spatial Anisotropy, Integrated Nested Laplace Approximations.

1 Introduction

Understanding space and time patterns of crime is key for accurate prediction and forecasting of crime, providing police with good information to implement effective crime prevention strategies. Certain types of crime exhibit strong space and time dependencies such as residential burglaries and car theft. The theory of repeat/near repeat victimisation (Townsley *et al.*, (2003)) which states that the likelihood of a location experiencing a crime is dependent on recent occurrences of crime at that location or within a given neighbourhood, encompasses the idea that both space and time dependencies exist. However, applying this theory in a real-world setting as is may not be a good assumption to make. Spatial dependencies (or correlations)

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

over a domain are generally anisotropic, meaning that spatial correlations do not extend in all directions in the same way. Spatial features such as a river or a major transport corridor, could block spatial correlations from extending to the other side of the feature. Though spatial anisotropy is a more realistic assumption to make, it is not generally considered in spatial modelling of crimes. In this short manuscript we introduce a model of residential burglaries that incorporates an anisotropic version of repeat/near repeat victimisation.

2 Methodology

2.1 Log-Gaussian Cox Processes

Consider the instances of residential burglaries in a spatiotemporal domain D as the realisation of a spatiotemporal point process, a random collection of observations having a spatial coordinate $\mathbf{s} = (x, y)$ and a time point t where

$$Y(\mathbf{s}, t) \equiv \left\{ y(\mathbf{s}, t), (\mathbf{s}, t) \in D \subset \mathbb{R}^2 \times \mathbb{R} \right\}. \tag{1}$$

We model this point process using an intensity function $\lambda(\mathbf{s}, t)$, which is a measure of the average number of observations per unit of space within a specific time interval. Given a bounded region D defined in (1), a point process model is an inhomogenous Poisson process where the number of points in a region and time interval is Poisson distributed with mean $\Lambda(\mathbf{s}, t) = \int_D \lambda(\mathbf{s}, t) d\mathbf{s}dt$. The likelihood of an inhomogenous point process Y given an intensity function λ can be expressed as

$$\pi(Y|\lambda) = \exp\left\{ |D| - \int_D \lambda(\mathbf{s}, t) d\mathbf{s}dt \right\} \prod_{\mathbf{s}_i, t \in Y} \lambda(\mathbf{s}_i, t). \tag{2}$$

Note that the likelihood is usually intractable as the integrand inside (2) cannot typically be calculated explicitly. In practical terms, we model the intensity as a log-Gaussian Cox process (LGCP), which models the log intensity as a realisation of a Gaussian random field $Z(\mathbf{s}, t)$. This can be framed as a Bayesian hierarchical model, and is a latent Gaussian model if we assume a multivariate Gaussian prior for the random field. Assuming the Gaussian random field can be approximated with a Gaussian Markov random field, the model fits inside the INLA modelling framework for fast Bayesian inference (Rue *et al.,* (2009)).

2.2 Non-stationary SPDE Approach

The stochastic partial differential equation (SPDE) approach to spatial and spatiotemporal modelling (Lindgren *et al.,* (2011)) was developed as a way of representing a continuous spatial process with a discretely indexed

spatial stochastic process. It also turns out to be a more computationally efficient method as opposed to a fine grid discretisation of the spatial domain (Simpson *et al.,* 2016). We use the non-stationary SPDE approach which allows us to define geographic features in the spatiotemporal domain that acts as barriers to spatial correlation (called the "barrier model"). The barrier model uses a spatially varying linear fractional SPDE

$$(\kappa(\mathbf{s})^2 - \Delta)^{\alpha/2}(\tau(\mathbf{s})Z(\mathbf{s})) = \mathcal{W}(\mathbf{s}), \tag{3}$$

where κ and τ are scale and precision parameters respectively and vary at different locations, α is the smoothness parameter, Δ is the Laplacian and $\mathcal{W}(\mathbf{s})$ is Gaussian spatial white noise. The solution to (3) is a non-stationary Gaussian field with mean 0 and covariance given by the Matèrn covariance function

$$Cov(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = \frac{\sigma^2}{\Gamma(\xi)2^{\xi-1}}(\kappa h)^{\xi} B_{\xi}(\kappa h). \tag{4}$$

In the above equation, σ^2 is the marginal variance, $h = ||\mathbf{s}_i - \mathbf{s}_j||$ represents the Euclidean distance between two spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$, and B_{ξ} is the modified Bessel function of the second kind with order $\xi > 0$. Rather than treating the above as a correlation function of the shortest distance between two points, the barrier model views it as a collection of paths through a simultaneous autoregressive process, so that local dependencies are manipulated in such a way as to cut off paths that cross geographic barriers. For more information see Bakka *et al.,* (2019).
In practice, the Gaussian field is approximated using a finite element mesh using a basis function representation defined by a triangularisation of the spatiotemporal domain. Details of the method can be found in Krainski *et al.,* (2019).

2.3 Dataset and Model

The dataset comprises of spatial locations of residential burglaries in Hamilton, New Zealand, for two months (March and April) in 2017. The city landscape is characterised by the Waikato River which runs through the city and divides it in easern and western halves. Burglaries are geo-coded as New Zealand transverse mercator coordinates (eastings and northings), and times are split over eight weeks. The spatial domain considered is the urban and suburban area of Hamilton City and we consider Waikato River and Lake Rotoroa as geographic barriers. We also consider two covariates related to residential burglaries. The measure of the socioeconomic conditions of a given location is measured using the NZ socioeconomic deprivation index (SDI, Atkinson *et al.,* (2020)). The count of graffiti instances within a 500 metre radius ($Graf$) of a location is used as a measure of anti-social behaviour. The model is given by:

$$\hat{\eta}_{i,t} = \beta_0 + \beta_1 SDI_{i,t} + \beta_2 Graf_{i,t} + f(\mathbf{s}_i, t; r, r_b = 0, \sigma, \rho),$$

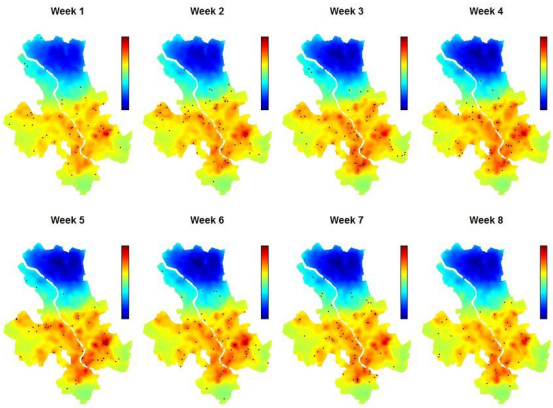


FIGURE 1. Spatial model maps of the eight weeks of data. Blue indicates a low intensity whereas red indicates high intensity. Black points are the actual observed locations of burglaries.

where β's are the coefficients for the fixed effects (including the intercept), and $\eta_{i,t}$ is the Gaussian linear predictor, indexed at location $i = 1, \ldots, N$ at time $t = 1, \ldots 8$. Note that N is the number of nodes in the spatial mesh and number of observed residential burglaries. The function f is governed by hyperparameters, and is a spatiotemporal function which is the Kronecker product of the Matèrn covariance function with an autoregressive process of order 1 (AR(1)). The use of an AR(1) process reflects that residential burglaries this week are dependent mainly on the number of residential burglaries that occurred last week, which is consistent with the temporal aspect of the repeat/near repeat victimisation theory. The hyperparameters include a spatial range parameter r, range parameter at the barriers $r_b = 0$, marginal standard deviation σ and time correlation ρ, all of which are given penalised complexity priors with the following specifications

$$\begin{aligned} \pi(r) &\sim PC(1, 0.95), \\ \pi(\rho) &\sim PC(0.7, 0.5), \\ \pi(\sigma) &\sim PC(1, 0.01). \end{aligned}$$

Note that the prior for the spatial correlation range r reads that $P(r < 1km) = 0.95$. The priors for ρ and σ read in reverse, so the prior for ρ reads that $P(\rho > 0.7) = 0.5$, and similarly for σ. The priors for r and ρ are based on results from the near-repeat calculator (Ratcliffe, (2020)). Once model fitting is done, a 9th week is then predicted based on the spatial model of the previous eight weeks.

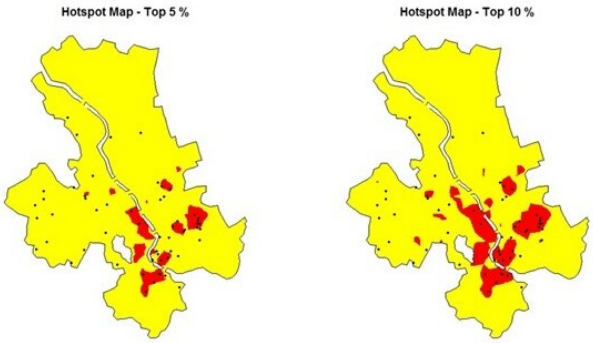


FIGURE 2. Hotspot maps with the predicted top 5% (left) and 10% (right) of intensities.

3 Preliminary Results

The posterior means of the estimated intensities are projected onto the spatial domain of the city and can be seen in Figure 1. Higher intensities occurred in the south central and south east of the city. The hotspot maps in Figure 2 show the prediction of the highest 5% and 10% of intensities. We can see that higher intensities in the south central part of Hamilton do not extend over the boundaries of the Waikato River. Both covariates were significant in the Bayesian sense (credible intervals do not include zero) and deviance information criterion outputs showed that this model fit better than the model with no covariates.

4 Further Work

We have so far placed a lot of emphasis on modelling the spatial aspects of burglaries. The model we have provides predictions that are continuous and spatially anisotropic which can provide greater insights as to where residential burglaries may take place. Further development and detail of geographic barriers in the spatial field may yield further prediction gains. However, the temporal aspects will also need further work for this model to be usable from a preventative policing point of view. Predictions over a finer time resolution would make the model far more usable, but would greatly increase the model complexity. For example, if we believe that the intensity of residential burglaries are dependent on what has happened in the past week, we would need to fit an autoregressive model of order 7, increasing the number of model hyperparameters and extinguishing any computational advantages that INLA may bring. Also, data sparsity is an issue as we make time resolutions finer. Clusters of burglaries become

apparent over time, but are much harder to detect when only modelling daily occurrence of burglaries.
Another temporal aspect that requires further work is that burglary data is right-censored with respect to time. Burglaries typically occur when the resident of the house is away, so the observed time is the time reported. Aoristic analysis (Ratcliffe, (2000)) has been used in the past to help with this problem, and is an idea we are looking at implementing in this model.

Acknowledgments: The author's wish to thank the Waikato District Police for the residential burglaries dataset.

References

Atkinson, J., Salmond, C., and Crampton, P. (2020). *NZDEP2018 Index of Deprivation.*. Wellington: University of Otago.

Bakka, H., Vanhatalo, J., Illian, J.B., Simpson, D., and Rue, H. (2019). Non-stationary Gaussian models with physical barriers. *Spatial Statistics*, **29**, DOI: 10.1016/j.spasta.2019.01.002.

Krainski, E.T. *et al.,* (2019). *Advanced Spatial Modelling with Stochastic Partial Differential Equations Using R and INLA*, Boca Raton: CRC Press.

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

Ratcliffe, J.H. (2000). Aoristic analysis: the spatial interpretation of unspecific temporal events. *International Journal of Geographic Information Science*, **14(7)**, 669 – 679.

Ratcliffe, J.H. (2020). *Near Repeat Calculator (version 2.0).*, Temple University, Philidelphia, PA.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B*, **71(2)**, 319 – 392.

Simpson, D., Illian, J.B., Lindgren, F., Sørbye, S.H., and Rue, H. (2016). Going off grid: Computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, **103(1)**, 49 – 70.

Townsley, M., and Homel, R., and Chaseling, J. (2003). Infectious burglaries: A test of the near repeat hypothesis. *British Journal of Criminology*, **43(3)**, 615 – 633.

Analysis of count data by quantile regression coefficient modelling: student's gained credits after online teaching

Viviana Carcaiso[1], Leonardo Grilli[2]

[1] Department of Statistical Sciences, University of Padua, Italy
[2] Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Italy

E-mail for correspondence: leonardo.grilli@unifi.it

Abstract: The impact of online teaching on the productivity of university students can be measured by the number of gained credits, which is a count variable with an irregular distribution. Quantile regression is a powerful and flexible method, though the application to count data entails some difficulties. We compare the standard approach based on jittering with a recently proposed approach based on parametric modelling of the quantile regression coefficients. We exploit the two approaches to analyse data from the University of Florence, discussing the pros and cons.

Keywords: COVID-19; jittering; model selection; parametric modelling; university credits.

1 Two approaches to quantile regression for counts

Quantile regression is a distribution-free method to analyse the relationships between the quantiles of a response variable and a set of covariates. Most theoretical developments refer to the case of a continuous response variable. The extension of quantile regression to count data raises several issues since the conditional quantile function of a discrete random variable cannot be a continuous function of the regression parameters. The traditional approach, proposed by Machado and Santos Silva (2005), is based on *jittering*: an artificial continuous variable Z is generated by adding a uniform random variable U to the original count variable Y. Then, the conditional quantile regression function is specified as $Q_Z(p|\boldsymbol{x}) = p+exp(\boldsymbol{x}'\boldsymbol{\beta}(p))$

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

and $\boldsymbol{\beta}(p)$ is estimated by linear quantile regression. To reduce the dependence of the estimates on the sampled values of the uniform distribution, the jittering procedure is repeated m times to get "average-jittering" estimators which are proved to be consistent and asymptotically normal. The jittering approach has been successfully applied in several settings, for example the analysis of credits gained by university students (Grilli et al., 2016).

A different route for applying quantile regression to count data is based on the *quantile regression coefficients modelling* paradigm of Frumento and Bottai (2016). The idea is to impose a parametric structure to the quantile regression coefficient functions and to fit the parameters all in once, instead of one quantile at a time. Specifically, the q regression parameters are defined as $\boldsymbol{\beta}(p|\boldsymbol{\theta}) = \boldsymbol{\theta}\boldsymbol{b}(p)$, where $\boldsymbol{b}$ is a vector of k known functions of p that are assumed to be continuous and differentiable, and $\boldsymbol{\theta}$ is a $q \times k$ matrix. The parameters in $\boldsymbol{\theta}$ are estimated by minimising an integrated loss function. Differently from standard quantile regression, the minimisation problem can be tacked with conventional algorithms such as Newton–Raphson or gradient search.

Frumento and Salvati (2021) propose to exploit quantile regression coefficients modelling for the analysis of count data. Given that the uniform random variable used for jittering has expectation 0.5, they apply the model to a transformation $T(\cdot)$ of $Y^\circ = Y + 0.5$, that is

$$Q_{T(Y^\circ)}(p|\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\beta}(p|\boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\theta}\boldsymbol{b}(p).$$

Two common choices for $T(\cdot)$ are the identity and the logarithm.

The regression coefficient modelling approach performs a smoothing that facilitates the interpretation. In addition, it might increase the estimation efficiency, as shown by the simulation study of Frumento and Salvati (2021). However, selecting an appropriate parametric form for each of the q coefficients is a difficult task because of the wide range of options and the unavailability of the standard model selection procedures. Frumento and Bottai (2016) propose a goodness-of-fit test which compares by the Kolmogorov–Smirnov statistic the distribution of $F(y_i|\,\boldsymbol{x}_i, \boldsymbol{\theta})$, $i = 1, \ldots, n$, with a Uniform distribution. This test is approximate in case of count data.

2 Comparing the two approaches: analysis of gained credits by university students after online teaching

At the beginning of March 2020, to limit the spread of the COVID-19 epidemic, the Italian government forced the universities to adopt online teaching. We aim at evaluating the impact of this decision on the productivity of students measured by the number of gained credits. Specifically, we analyse data collected in the administrative records about first-year students who enrolled in the Bachelor's degree programs of Psychological

Sciences and Industrial Design in the academic years 2018/19 and 2019/19 at the University of Florence. The two degree programs are different in terms of disciplines and admission rules (Psychological Sciences admits a fixed number of freshmen on the basis of a multiple choice test). The idea is to compare the productivity of freshmen in the second semester of 2019/20, which is the one affected by online teaching, to that of first-year students in the second semester of 2018/19, who received the usual frontal lectures. Both cohorts attended regular lectures in the first semester, but were exposed to different forms of teaching in the second one. In both academic years of interest the study plan remained the same. After excluding students who got no credits in the first semester, the data set consist of 946 first-year students (649 for Psychological Sciences and 297 for Industrial Design). Students' productivity is measured by the number of gained credits (ECTS). According to the study plan, students should obtain approximately 30 credits per semester. Since the number of credits is always a multiple of 3, the response variable used in the models, denoted by Y, is defined as the number of credits gained in the second semester divided by 3. The distribution of Y is quite irregular, thus quantile regression is an appealing methodology. The following covariates are included in the model: number of credits obtained during the first semester, centred around 15; a dummy variable for cohort 2019 (online teaching) vs 2018 (in-presence teaching); a dummy variable for male vs female; high school grade, centred around 80 (it ranges from 60 to 100); six dummy variables for the type of high school (baseline category: "Scientific").

The analysis is performed separately for each degree course. The effect of interest is the coefficient of the binary variable distinguishing the students who experimented online teaching (cohort 2019) from the others (cohort 2018). We apply the two approaches to quantile regression for counts outlined in Section 1. In particular, for the *jittering* approach we specify the model

$$Q_Z(p|\boldsymbol{x}) = p + \boldsymbol{x}'\boldsymbol{\beta}(p) = p + \beta_0(p) + \sum_{i=1}^{10} \beta_i(p)\, x_i,$$

whereas for *quantile regression coefficients modelling* we specify the model

$$Q_{Y^\circ}(p|\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}'\boldsymbol{\beta}(p|\boldsymbol{\theta}) = \beta_0(p|\boldsymbol{\theta}) + \sum_{i=1}^{10} \beta_i(p|\boldsymbol{\theta})\, x_i.$$

To facilitate the comparison, in both approaches the transformation of the working variable is the identity function. The first approach is implemented through the function `rq` of the R package `quantreg`, while the second approach is carried out with the function `iqr` of the `qrcm` package.

For the coefficient modelling approach, we tried several parametric forms for the coefficients $\beta_i(p|\boldsymbol{\theta})$, compared on the basis of the number of free model parameters, the integrated loss function and the p-values of the

Kolmogorov–Smirnov (KS) test for goodness-of-fit. To limit the range of alternative specifications, we decided to be very accurate only with two coefficients: (*i*) the coefficient of the intercept, which serves as a baseline; (*ii*) the coefficient of the binary variable distinguishing the students who experimented online teaching from the others, which is the parameter of main interest. Those coefficients have been modelled in a flexible way with logarithmic transformations and polynomials. The other coefficients, which play the role of controls, are approximated by simple functions like lines and constants. Furthermore, the choice of the parametric functions is guided by the comparison with the non-parametric patterns provided by the jittering approach.

The results from *quantile regression coefficients modelling* show that, for Psychological Sciences, the effect of online teaching is negative at most quantiles, notably on the tails, namely for students with low or high productivity; however the estimated effect is always small and not statistically significantly at 5%. For freshmen in Industrial Design the effect is instead positive, though it is small and almost never statistically significant.

As for the relative merits of the two approaches, we found that *quantile regression coefficients modelling* yields smooth functions of the coefficients with less variation than the point-wise estimates of the *jittering* approach, especially in the tails. Indeed, a comparison of the standard error reveals a greater efficiency of the first approach except for central quantiles. The drawback lies in the complexity of model selection due to the wide range of options and the lack of well-grounded selection procedures. In this regard, the non-parametric *jittering* approach remains necessary in practice, as it provides a reference for assessing the adequacy of the parametric functions.

References

Frumento, P. and Bottai, M. (2016). Parametric modeling of quantile regression coefficient functions. *Biometrics*, **72**, 74 – 84.

Frumento, P. and Salvati, N. (2021). Parametric modeling of quantile regression coefficient functions with count data. *Stat. Methods Appl.*, **30**, 1237 – 1258.

Grilli, L., Rampichini, C., Varriale, R. (2016). Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: an approach based on quantile regression for counts. *Stat. Modelling*, **16**, 47 – 66.

Machado, J. A. F. and Santos Silva, J. M. C. (2005). Quantiles for counts. *J. Am. Stat. Assoc.*, **100**, 1226 – 1237.

Bayesian Discrete Conditional Transformation Models

Manuel Carlan[1], Thomas Kneib[1]

[1] Chair of Statistics, Georg-August-Universität Göttingen, Germany

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

Abstract: We propose a Bayesian modeling framework for discrete ordinal and count data based on conditional transformations of the responses. The conditional transformation function is estimated from the data in conjunction with an a priori chosen reference distribution. For count responses, the resulting transformation model is a Bayesian fully parametric yet distribution-free approach that can additionally account for excess zeros with additive transformation function specifications. For ordinal categoric responses, our cumulative link transformation model allows for the inclusion of linear and nonlinear covariate effects that can additionally be made category-specific, resulting in (non-)proportional odds or hazards models. Inference is conducted by a generic modular Markov chain Monte Carlo algorithm where multivariate Gaussian priors enforce specific properties such as smoothness on the functional effects. To illustrate the versatility of Bayesian discrete conditional transformation models, applications to counts of patent citations in the presence of excess zeros and on treating forest health categories in a discrete partial proportional odds model are presented.

Keywords: Bayesian transformation models; penalised splines; overdispersion; zero-inflation; partial proportional odds.

1 Introduction

Distributional regression models that overcome the traditional focus of regression analyses on the conditional mean of the response distribution and rather focus on the complete response distribution have seen considerable interest in the past decade (see, e.g., Kneib et al. 2022 for an overview). One particularly interesting special case of distributional regression are conditional transformation models (CTMs, Hothorn et al. 2018), where the cumulative distribution function (CDF) of the response Y given covariates

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

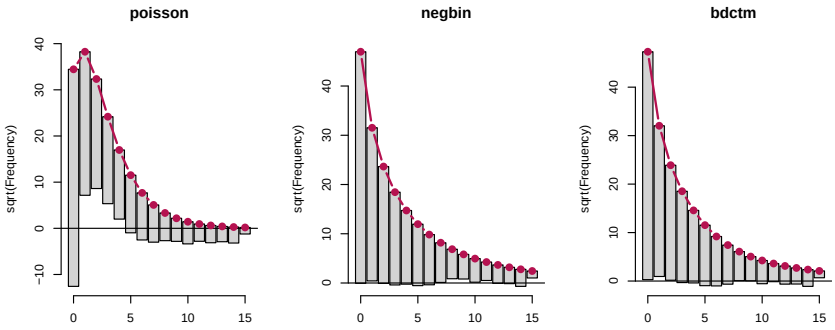


FIGURE 1. Rootograms of the linear Poisson, the linear negative binomial and the simple linear BDCTM model.

$\boldsymbol{X} = \boldsymbol{x}$ is modelled as

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|\boldsymbol{x}) = P(Y \leq y|\boldsymbol{x}) = F_Z(h(y|\boldsymbol{x}))$$

with a pre-specified reference CDF F_Z (e.g. the standard normal) and a covariate-dependent bijective transformation function $h(y|\boldsymbol{x})$ that has to be estimated from the data. The resulting model is semiparametric in the sense that a parametric specification for the transformation function is considered while the overall setup allows to generate very flexible response distributions. To make the model specification feasible, one typically assumes an additive composition of the transformation function as $h(y|\boldsymbol{x}) = \sum_{j=1}^{J} h_j(y|\boldsymbol{x})$, where $h_j(y|\boldsymbol{x})$ are response-covariate interactions that are monotone in direction of y.
For continuous responses, the density implied by a CTM can be derived via the transformation theorem for densities as

$$f_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|\boldsymbol{x}) = f_Z(h(y|\boldsymbol{x})) \left| \frac{\partial h(y|\boldsymbol{x})}{\partial y} \right|$$

such that likelihood-based and Bayesian inference become available (see Hothorn et al. 2018 and Carlan at al. 2020). In this paper, we introduce extensions of CTMs for discrete responses within the Bayesian framework and present two applications for count data and ordinal responses.

2 Count Transformation Models

For count responses, the basic idea of discrete CTMs is to truncate the transformation via the floor operator $\lfloor y \rfloor$ such that

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y|\boldsymbol{x}) = F_Z(h(\lfloor y \rfloor|\boldsymbol{x})).$$

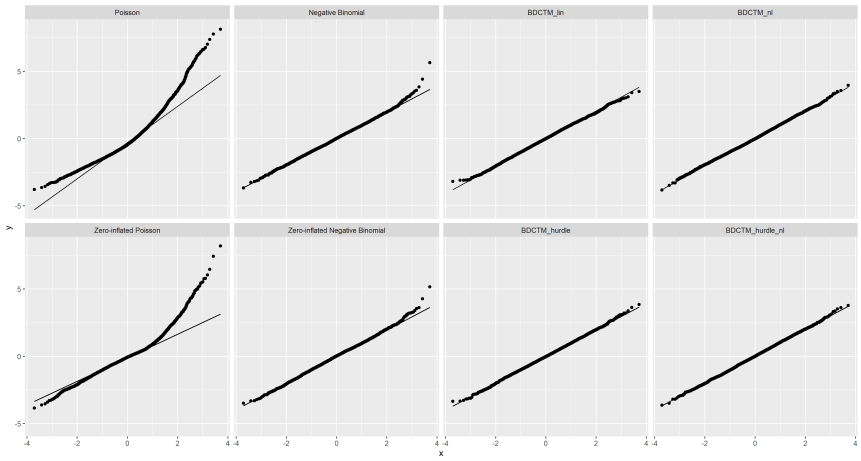


FIGURE 2. Comparison of quantile residuals obtained by BDCTM models with and without addititional zero component with various generalized linear and zero-inflated models.

While the transformation theorem for densities can no longer be applied in this case, we can still derive the corresponding probability mass function as

$$\log(f_Z(y|\boldsymbol{x})) = \begin{cases} \log[F_Z(h(y|\boldsymbol{x}))] & y = 0 \\ \log[F_Z(h(y|\boldsymbol{x})) - F_Z(h(y-1|\boldsymbol{x}))] & y = 1, 2, \dots \end{cases}$$

by evaluating the step heights of the CDF. To further increase the flexibility of the model, we can also consider two component mixtures including an explicit separate probability for zeros that is useful in the presence of excess zeros.

As an illustration, we analyze the number of citations (*ncit*) for $n = 4,805$ patents granted by the European Patent Office. The data set includes five dummies and three continuous variables (grant year, the number of the designated states, number of patent claims) as explanatory variables. A high rate of zeros ($\approx 46\%$) and a big spread $ncit \in \{0, \dots, 40\}$ hint on the presence of zero-inflation and overdispersion. The rootograms presented in Figure 1 as well as additional inspections of quantile residuals in Figures 2 indicate that, even after adjusting for covariates, standard specifications for count data are not necessarily flexible enough to capture these features while the Bayesian discrete CTM offers a decent fit without requiring the strong assumption of pre-specified response distribution.

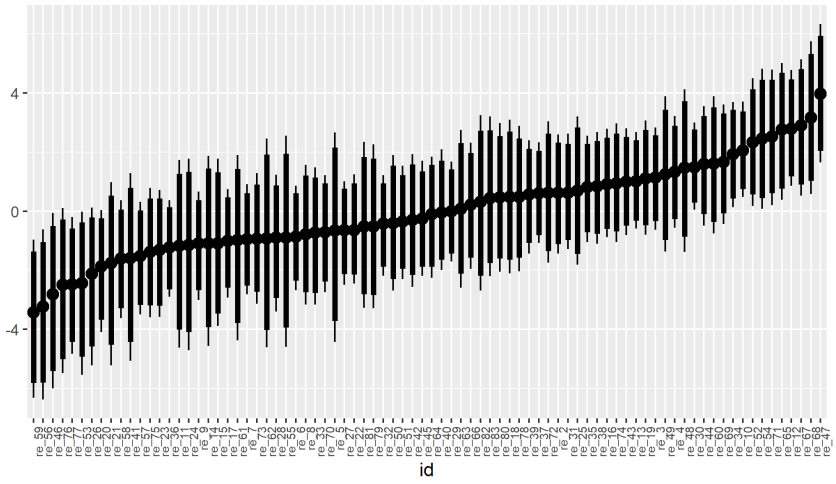


FIGURE 3. Median-sorted estimated random intercepts for tree location ID.

3 Cumulative Link Transformation Models

For ordinal responses, we consider models in the spirit of cumulative regression models, yet with non-proportional specification such that

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_k|\boldsymbol{x}) = F_Z(h_Y(y_k) + h_k(\boldsymbol{x})),$$

where y_k, $k = 1, 2, \ldots, K$ are the labels for the ordered response categories, $h_Y(y_k)$ is a component of the transformation function independent of covariates and $h_k(\boldsymbol{x})$ induces the non-proportionality by interacting the response category with covariates. Partial proportional models result, when some of the components $h_k(\boldsymbol{x})$ are restricted to be independent of the response category, i.e. $h_k(\boldsymbol{x}) \equiv h(\boldsymbol{x})$.
For cumulative link transformation models, we consider an analysis of forest health, where the health status of trees is evaluated in three ordered defoliation grades (1 = no $(0\%), 2$ = weak $(12.5\% - 37.5\%)$ and 3 = severe $(\geq 50\%)$). Among others, the dataset comes with the covariate canopy density in percent, longitude and latitude of the tree location, and tree location identification number. The goal of our analysis is to determine the effect of the covariates on the degree of defoliation. For this, we set up a partial proportional odds model where we assume nonlinear category-specific shifts of canopy density, a transformation random effect for the tree location groups and a spatial nonlinear effect on basis of a tensor spline for the coordinates. Figures 3 and 4 shows the estimated random intercepts and spatial effect, respectively.

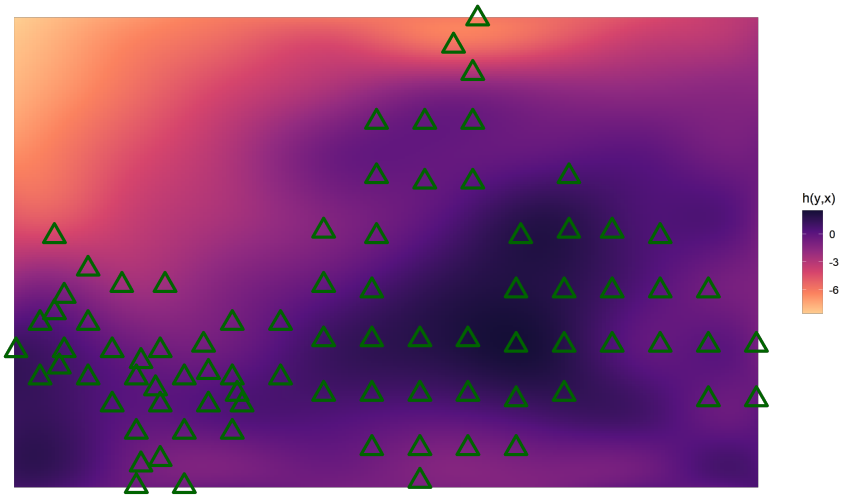


FIGURE 4. Estimated two-dimensional spatial effect with triangles indicating observed tree locations (right).

4 Parameterisation and Inference

For all our models, we utilize (Bayesian) penalized splines and their tensor product interactions with various other effect types (such as random effects or spatial effects in the forest health example) to obtain flexible transformation functions while enabling regularization by appropriate smoothness priors. Monotonicity along the response dimension is ensured by reparameterizing the basis coefficients in a way that leads to a monotonically increasing sequence (Pya and Wood, 2015). For Bayesian inference, we rely on Markov chain Monte Carlo simulations implemented with the No-U-turn sampler with dual averaging for efficient exploration of the posterior distribution. Details are available in Carlan and Kneib (2022).

References

Carlan, M., and Kneib, T. (2022). Bayesian Discrete conditional transformation models. arXiv:2205.08594. Accepted for publication in *Statistical Modelling*

Carlan, M., Kneib, T., and Klein, N. (2020). Bayesian conditional transformation models. arXiv:2012.11016.

Hothorn, T., Möst, L., and Bühlmann, P. (2018). Most likely transformations. *Scandinavian Journal of Statistics*, **45**, 110 -– 134.

Kneib, T., Silbersdorff, A., and Säfken, B. (2022). Rage Against the Mean – A Review of Distributional Regression Approaches. *Econometrics and Statistics*, to appear.

Pya, N. and Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing*, **25**, 543 – 559.

Competing risks models with two time scales

Angela Carollo[12], Hein Putter[2], Paul Eilers[3], Jutta Gampe[1]

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

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

Abstract: Competing risks models can involve more than one time scale. We propose a model for competing events in which the cause-specific hazards vary smoothly over two times scales. We estimate these two-dimensional hazard functions by P-splines, exploiting the equivalence between hazard smoothing and Poisson regression. As the data are arranged on a grid we can make use of generalized linear array models (GLAM) for efficient computations. We present an application to the study of transitions out of non-marital cohabitation in Germany.

Keywords: Cause-specific hazards; Two-dimensional smoothing; P-splines; GLAM

1 Introduction

Competing risks describe the situation where individuals are at risk of experiencing one of several types of events (Putter et al., 2007). The prototype of a competing risks model is the study of cause-specific mortality. For example, in clinical studies of cancer it is common to analyse mortality from cancer and mortality due to other causes. But competing events are also present in demographic studies: A marriage can end either in divorce or in widowhood, a non-marital cohabitation ends by marriage or separation.
Time is a key quantity in any event history analysis, and it can be recorded over several time scales. For example, after a cancer diagnosis, the risk of death might be studied over time since diagnosis, over age, which is time since birth, or over time since treatment. In demography, age-specific rates are the most common choice, but other time scales, like time since marriage, are often relevant too. All time scales progress at the same speed and differ only in their origin.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

The key quantities in a competing risks analysis are the cause-specific hazards. They are defined as the instantaneous risk of experiencing an event of a specific type at time t, given no event (of any type) has happened yet. From them the overall survival function, i.e., the probability of no event up to time t, and the cumulative incidence functions, the probability of an event of given type before t, can be derived.
Usually, cause-specific hazards are defined for the same single time scale, and little research has been done on how to handle multiple time scales in competing risks models. Lee et al. (2017) propose a model where two competing causes are modelled on two different time scales and then combined under one of the two scales to estimate the cumulative incidence functions. Carollo et al. (2022) propose a model for a single event where the hazard varies smoothly over two time scales and is estimated by tensor products P-splines. Here we develop this model further for a competing risks setting. Each cause-specific hazard varies over two time scales, and estimation again is achieved by bivariate P-splines smoothing. Therefrom, we calculate the cumulative incidence functions for each cause.
As an application we study transitions out of cohabitation, either into marriage or into separation, for women living in West Germany.

2 A competing risks model with two time scales

Consider individuals in non-marital cohabiting unions. At each point in time a cohabiting individual is at risk of marrying (event 1) or of separating from the current partner (event 2). These transitions can be studied along two time scales, namely $t =$ age and $s =$ duration of cohabitation. A graphical depiction of this process in presented in Figure 1.
The cause-specific hazards for event type $\ell \in \{1, 2\}$ over the two time scales t and s are defined as

$$\lambda_\ell(t, s) = \lim_{\Delta \downarrow 0} \frac{P\left(\text{event}_\ell \in \{t + k\Delta, s + k\Delta : 0 \leq k \leq 1\} | \text{no event before } (t, s)\right)}{\Delta}.$$

Here $t > s$, so the two-dimensional hazards $\lambda_\ell(t, s)$ are only defined in the lower half-open triangle of $\mathbb{R}^2_+$.
The two time scales differ in their origin, which in this example is the age t_0 when the individual enters the cohabitation. This age differs between individuals and in a Lexis diagram individuals move along 45°-lines from $(t_0, 0)$ to $(t_0 + v, v)$ until they leave the risk set (due to event or censoring). This allows to view cause-specific hazards equivalently as two-dimensional functions $\tilde{\lambda}_\ell(u, s)$, where $\tilde{\lambda}_\ell(u = t - s, s) = \lambda_\ell(t, s)$. The $\tilde{\lambda}_\ell(u, s)$ are defined over the full positive quadrant $\mathbb{R}^2_+$.
From the cause-specific hazards $\lambda_\ell(t, s)$ or $\tilde{\lambda}_\ell(u, s)$, respectively, we obtain the cumulated cause-specific hazards

$$\Lambda_\ell(t, s) = \int_0^s \lambda_\ell(t_0 + v, v)\, dv \qquad \text{or} \qquad \tilde{\Lambda}_\ell(u, s) = \int_0^s \tilde{\lambda}_\ell(u, v)\, dv,$$

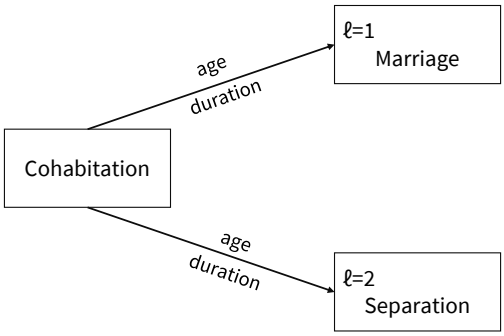


FIGURE 1. Competing risks process for the transitions out of cohabitation by age and duration of the cohabitation.

and the overall survival function

$$S(t,s) = \exp\left\{-\sum_{\ell=1}^{2}\Lambda_\ell(t,s)\right\} \qquad \text{or} \qquad \tilde{S}(u,s) = \exp\left\{-\sum_{\ell=1}^{2}\tilde{\Lambda}_\ell(u,s)\right\},$$

where $u = t - s$. The cumulative incidence functions (CIF) are

$$I_\ell(t,s) = \int_0^s \lambda_\ell(t_0+v,v)\, S(t_0+v,v)\, dv \;;\; \tilde{I}_\ell(u,s) = \int_0^s \tilde{\lambda}_\ell(u,v)\, \tilde{S}(u,v)\, dv.$$

In the context of our application, $\tilde{I}_\ell(u,s)$ is the cumulative probability of marriage ($\ell = 1$) or separation ($\ell = 2$) within s years after start of cohabitation for a subject who entered cohabitation at age u.
To estimate the cause-specific hazard surfaces $\tilde{\lambda}_\ell(u,s)$ we divide the (u,s)-plane into $J \times K$ small bins (squares) and within each bin we count the number of events of type ℓ, denoted by $y_{jk}^{(\ell)}$, and the exposure times r_{jk}.
The $y_{jk}^{(\ell)}$ are assumed to be realizations of Poisson variates with means

$$\tilde{\mu}_{jk}^{(\ell)} = r_{jk} \cdot \tilde{\lambda}_{jk}^{(\ell)} = r_{jk} \cdot \exp\{\tilde{\eta}_{jk}^{(\ell)}\}. \tag{1}$$

The $\tilde{\lambda}_{jk}^{(\ell)}$ represent the cause-specific hazard $\tilde{\lambda}_\ell(u,s)$ evaluated at the center of bin (j,k).
The log-hazards $\tilde{\eta}_\ell(u,s)$ are assumed to be smooth functions and they are modelled as sums of tensor products of B-splines. Difference penalties on the coefficients, one in the row- and one in the column-direction, ensure smoothness of the estimates (Currie et al., 2004).

Due to the binning the data, event counts and at-risk times, are on a regular grid and are naturally arranged as $J \times K$ matrices

$$Y_\ell = [y_{jk}^{(\ell)}] \qquad \text{and} \qquad R = [r_{jk}].$$

Correspondingly, we denote $M_\ell = [\tilde{\mu}_{jk}^{(\ell)}]$ and $E_\ell = [\tilde{\eta}_{jk}^{(\ell)}]$. The tensor products are formed from two marginal B-splines bases of size m and $\breve{m}$, respectively, along the u- and s-axis. The two basis matrices are denoted by B, which is $J \times m$, and by $\breve{B}$, which is $K \times \breve{m}$. The coefficients are arranged in the matrix $A_\ell = [\alpha_{fg}^{(\ell)}]$, where $f = 1, \ldots, m$ and $g = 1, \ldots, \breve{m}$.
The log-hazard can then be expressed in a compact way as $E_\ell = B \cdot A_\ell \cdot \breve{B}^{\mathrm{T}}$.
The penalty on the coefficients in A_ℓ is constructed from two matrices D and $\breve{D}$ that form differences (of order d) of the columns of a matrix and it is controlled by two smoothing parameters ρ and $\breve{\rho}$ to allow anisotropic smoothing:

$$\mathrm{pen}(\rho, \breve{\rho}) = \rho \, ||DA_\ell||_{\mathrm{F}}^2 + \breve{\rho} \, ||A_\ell \breve{D}^{\mathrm{T}}||_{\mathrm{F}}^2$$

($||.||_{\mathrm{F}}^2$ represents the sum of all squared elements of a matrix.)
The objective function to minimize resulting from (1) is

$$\begin{aligned} \mathrm{Q}_\ell &= dev(M_\ell; Y_\ell) + \mathrm{pen}(\rho, \breve{\rho}) \\ &= 2 \sum_{j=1}^{J} \sum_{k=1}^{K} \left(y_{jk}^{(\ell)} \ln(y_{jk}^{(\ell)} / \tilde{\mu}_{jk}^{(\ell)}) - (y_{jk}^{(\ell)} - \tilde{\mu}_{jk}^{(\ell)}) \right) + \mathrm{pen}(\rho, \breve{\rho}) \end{aligned}$$

which leads to normal equations that can be solved, for given ρ and $\breve{\rho}$, in a penalized Poisson IWLS scheme (in compact notation):

$$\left[(\breve{B} \otimes B)^{\mathrm{T}} W_\ell' (\breve{B} \otimes B) + P \right] \alpha_\ell' = (\breve{B} \otimes B)^{\mathrm{T}} W_\ell' z_\ell'.$$

Here, $P = \rho(\breve{I} \otimes D^{\mathrm{T}} D) + \breve{\rho}(\breve{D}^{\mathrm{T}} \breve{D} \otimes I)$, with I and $\breve{I}$ identity matrices of appropriate dimension, W_ℓ is a diagonal matrix of weights, α_ℓ is the vector of coefficients, z_ℓ is the working variable and the prime symbol indicates the current value in the iteration.
The special format of Y_ℓ, R and E_ℓ allows to employ generalized linear array methods (Currie et al., 2006) for efficient computations. We determine the optimal values for the smoothing parameters by numerically optimizing the AIC of the model as a function of $(\rho, \breve{\rho})$.
Once the coefficients $\hat{A}_\ell = [\hat{\alpha}_{fg}^{(\ell)}]$ are obtained, we can evaluate the estimated $\tilde{\eta}_\ell(u, s)$ and hence also the cause-specific hazards $\tilde{\lambda}_\ell(u, s)$ on a detailed grid. Consequently, the cumulative hazards $\tilde{\Lambda}_\ell(u, s)$, the overall survival function $\tilde{S}(u, s)$ and the CIF $\tilde{I}_\ell(u, s)$ can be obtained by simple numerical integration (rectangle or trapezoid rule) with sufficient accuracy.

3 Application: transitions out of cohabitation

We applied this approach to study transitions out of cohabitation, either to marriage or to separation, for West German women over age and length of the cohabitation. The data come from the 11$^{\text{th}}$ wave of the German Family Panel (pairfam). The pairfam is a longitudinal panel study for researching issues of family and relationships dynamics in Germany (Huinink et al., 2011).

For this application we selected all women living in West Germany who at the time of their first interview were already in a non-marital cohabiting union, or have entered one at some time during the study period (2008-2019). We follow the trajectories of these cohabiting unions from either the start of the cohabitation, or the time of the first interview, until marriage, separation or end of the follow-up period. For each individual in the sample we know the age at beginning of the cohabitation, the age at the event or censoring time, and the duration of the cohabitation (which is left truncated if the cohabiting union was already formed by the time the individual entered the study).

We first estimate the cause-specific hazards along age and duration of the cohabitation for West German women by dividing the transformed positive quadrant into a grid of 62 by 54 bins. Then, we compute the cubic marginal B-splines by placing a knot every 3 bins circa, resulting in 20 and 17 B-splines on the u and s axis respectively.

The cause-specific hazards estimated from the data, given the optimal smoothing parameters, are represented in the transformed positive quadrant in Figure 2.

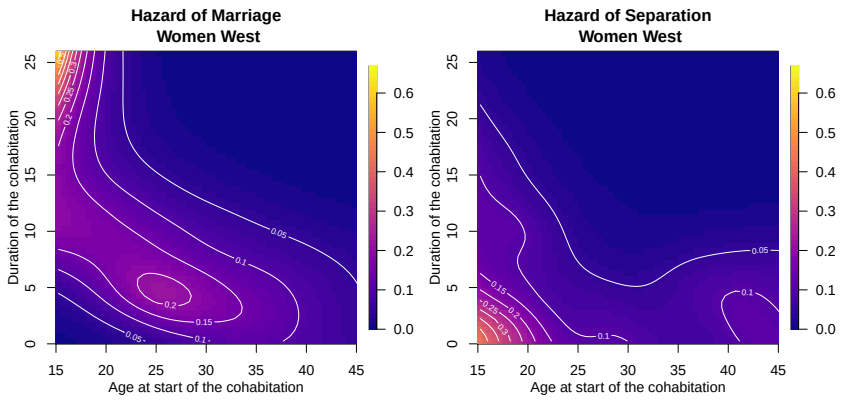


FIGURE 2. Cause-specific hazards of marriage (left panel) and separation (right panel) by age at entry into cohabitation and duration of cohabitation.

From these, we calculated the estimated cumulative incidence functions,

shown in Figure 3.

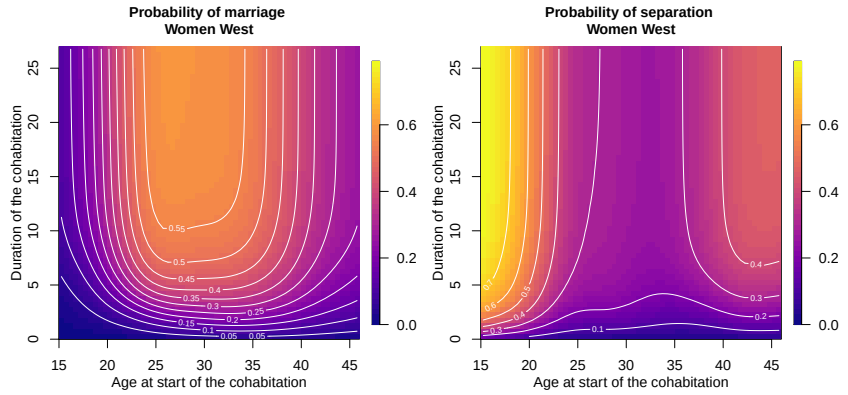


FIGURE 3. Cumulative incidence functions of marriage (left panel) and separation (right panel) by age at entry into cohabitation and duration of cohabitation.

References

Carollo, A., Putter, H., Eilers, P.H.C. and Gampe, J. (2022). Smooth hazards with multiple time scales. (preprint)

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

Currie, I.D., Durban, M. and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *JRSS B*, **68**, 259-280.

Huinink, J. et al. (2011). Panel Analysis of Intimate Relationships and Family Dynamics (pairfam): Conceptual framework and design. *Zeitschrift f"ur Familienforschung*, **1/2011**, 77-101.

Lee, M., Gouskova, N.A., Feuer, E.J and Fine, J.P. (2017). On the choice of time scales in competing risks predictions. *Biostatistics*, **18**, 15 – 31.

Putter, H., Fiocco, M. and Geskus, R.B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389 – 2430.

A Responding Attitude Component in Hidden Markov Models

Roberto Colombi[1], Sabrina Giordano[2], Maria Kateri[3]

[1] Department of Management, Information and Production Engineering, University of Bergamo, Italy
[2] Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria, Italy
[3] Institute of Statistics, RWTH Aachen University, Germany

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

Abstract: For longitudinal categorical data, we propose a hidden Markov model with a bivariate latent Markov chain that jointly models an unobservable trait of interest and a binary indicator representing a interviewee's response style over time. The novelty consists of modelling the substantive latent trait, taking simultaneously under account the underlying responding behavior and its evolution over time. Alternative existing approaches either ignore the respondents' responding behavior or consider it as a continuous time-invariant latent trait.

Keywords: Response style; Stereotype model; Model selection

1 Motivation and Models

Opinions, behaviors and perceptions are generally subjective indicators of non-directly measurable variables and are usually collected through Likert-type items.
The categories selected by responders may not represent their true preferences but their tendency to use only a certain rating scale options, governed by an underlying behavioral attitude, known as Response Style (RS) (e.g., Van Vaerenbergh and Thomas, 2013). If ignored, RS distorts the latent trait measurement and induces bias in the estimates of parameters.
The interest here is on the longitudinal perspective where responses, categorical indicators of a latent trait of interest at several time occasions, can be driven or not by RS and the RS attitude can vary dynamically, allowing also dependence on individual characteristics.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

The methodological proposal is an extension of a hidden Markov model (HMM) by Bartolucci et al, 2012, with a bivariate latent Markov chain that jointly models an unobservable trait of interest and an unobservable binary indicator of the respondent's form of answering (RS driven or not, i.e., RS or $\overline{RS}$) over time.

Consider r ordinal responses observed on n units at T time occasions. In particular, let Y_{jit}, $Y_{jit} \in \mathcal{C}_j = \{1, \ldots, c_j\}$, denote the j-th ordinal response variable, $j \in \mathcal{R} = \{1, \ldots, r\}$, of the i-th unit, $i \in \mathcal{I} = \{1, \ldots, n\}$, at the t-th occasion, $t \in \mathcal{T} = \{1, \ldots, T\}$. The responses are assumed to reflect the levels of unobservable latent constructs L_{it}, $i \in \mathcal{I}$, $t \in \mathcal{T}$, with finite discrete state space $\mathcal{S}_L = \{1, \ldots, k\}$. Furthermore, they can be observed under two latent regimes: RS or $\overline{RS}$ that are captured by binary latent variables U_{it}, $i \in \mathcal{I}$, $t \in \mathcal{T}$, with state space $\mathcal{S}_U = \{1, 2\}$, where 1 and 2 denote the RS and $\overline{RS}$ states, respectively.

The proposal is a HMM defined by two components that describe the Markov chain of the latent variables and the conditional distributions of the responses given the latent variables. Covariates $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$, $t \in \{2, \ldots, T\}$ may influence the transition probabilities of the latent variables U_{it} and L_{it} for the i-th unit, respectively. They are assumed to affect only the distribution of the latent variables. In our view, in fact, the covariate effect is captured by the latent constructs L_{it} which are indirectly observed trough the responses Y_{jit}.

The assumptions allowing the transition probabilities of the bivariate Markov chain $\{L_{it}, U_{it}\}_{t \in \mathcal{T}}$ to be $\pi_{it}(u, l|\bar{u}, \bar{l}) = \pi_{it}^{U|L}(u|l, \bar{u})\pi_{it}^{L}(l|\bar{l}), t = 2, \ldots, T$, are illustrated in Figure 1.

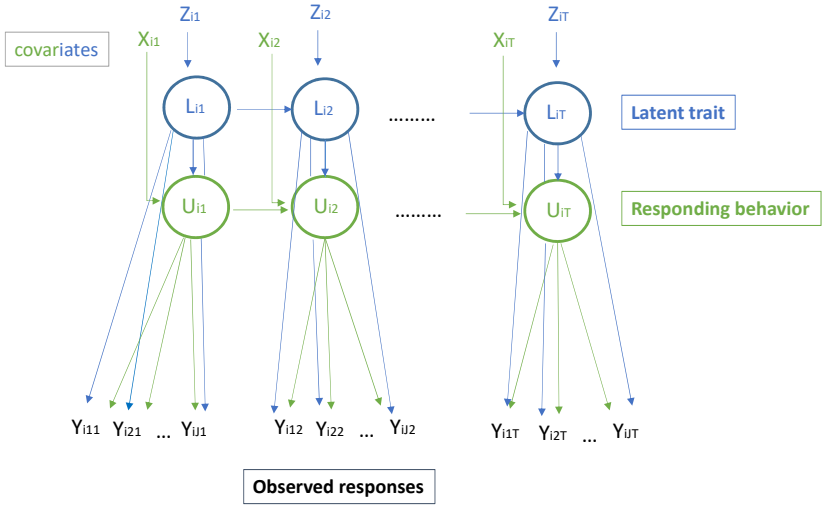


FIGURE 1. HMM with an RS component

Let us describe the models for $\pi_{it}^{U|L}(u|l,\bar{u})$ and $\pi_{it}^{L}(l|\bar{l})$, and for the observation probability functions, under the two regimes RS and $\overline{RS}$.

Latent model.

A stereotype logit model is adopted for the latent trait transition probabilities:

$$\log \frac{\pi_{it}^{L}(l|\bar{l})}{\pi_{it}^{L}(\bar{l}|\bar{l})} = \beta_{0l\bar{l}} + \nu_{l\bar{l}}\boldsymbol{\beta}'_{1\bar{l}}\boldsymbol{z}_{it}, \quad l \neq \bar{l}, \quad l,\bar{l} \in \mathcal{S}_L, \quad t = 2,\ldots,T.$$

For $\bar{l} \neq 1$, $\nu_{1\bar{l}} = 1$, for $\bar{l} = 1$, $\nu_{2\bar{l}} = 1$, while the rest of the ν-scores are to be estimated. A logit model is considered for the conditional RS transition probabilities for each possible RS state $\bar{u}$ of the previous occasion and for each current state l of the latent construct:

$$\log \frac{\pi_{it}^{U|L}(2|l,\bar{u})}{\pi_{it}^{U|L}(1|l,\bar{u})} = \bar{\beta}_{0l\bar{u}} + \bar{\boldsymbol{\beta}}'_{1l\bar{u}}\boldsymbol{x}_{it}, \quad l \in \mathcal{S}_L, \quad \bar{u} \in \mathcal{S}_U, \quad t = 2,\ldots,T.$$

Observation model.

The observation probability functions are parameterized (without covariates) as follows.

Given the RS regime, every probability function $f_{j|1}(y_j|l)$, $j \in \mathcal{R}$, $l \in \mathcal{S}_L$, is specified by the linear local logit model:

$$\log \frac{f_{j|1}(y_j+1|l)}{f_{j|1}(y_j|l)} = \phi_{0lj} + \phi_{1lj}s_j(y_j), \quad y_j = 1,2,\ldots,c_j-1.$$

The ϕ_{0lj}, ϕ_{1lj} are parameters to estimate and the scores $s_j(y_j)$ are known constant defined as: $s_j(y_j) = 1$ for $y_j < c_j/2$, $s_j(y_j) = 0$ for $y_j = c_j/2$, $s_j(y_j) = -1$ for $y_j > c_j/2$, $y_j = 1,2,\ldots,c_j-1$. Parameter ϕ_{0lj} governs the skewness of the probability function $f_{j|1}(y_j|l)$, so that it is symmetric with $\phi_{0lj} = 0$, left and right skewed with $\phi_{0lj} > 0$ and $\phi_{0lj} < 0$, respectively. This allows a parsimonious representation of various RS-types.

Given the $\overline{RS}$ regime, every probability function $f_{j|2}(y_j|l)$, $j \in \mathcal{R}$, $l \in \mathcal{S}_L$, is parameterized by $c_j - 1$ adjacent categories logits:

$$\log \frac{f_{j|2}(y_j+1|l)}{f_{j|2}(y_j|l)} = \varphi_{y_j l}, \quad y_j = 1,2,\ldots,c_j-1.$$

2 Inference and Model Selection

Maximum likelihood estimates of the parameters, listed in $\boldsymbol{\theta}$, are calculated via an EM algorithm. In the E step, the following expected values are computed:

$$\delta_{it}^{(1)}(u,l;\bar{\boldsymbol{\theta}}) = E_{obs}(d_{it}^{(1)}(u,l)), \tag{1}$$

where the latent binary variable $d_{it}^{(1)}(u,l)$ is equal to 1 when the i-th unit is at time t in state (u,l) and $E_{obs}()$ is the conditional expected value given the observed values of the responses Y_{jit}, the covariates and given the estimate $\bar{\boldsymbol{\theta}}$ of the parameter vector $\boldsymbol{\theta}$.
Model selection can be based on indices for measuring the quality of classification and the distinguishability of the latent classes. Such an index, based on the posterior probabilities of the latent classes (Bartolucci et al, 2012), takes in our set-up the form:

$$S_k = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(\delta_{it}^{*} - 1/2k)}{(1-1/2k)nT}, \tag{2}$$

where k is the number of states of the latent construct and δ_{it}^{*} is, for unit i at time t, the maximum with respect to (u,l) of the posterior latent class probabilities (1). Measure S_k lies between 0 and 1, where 1 represents certainty in classification and a perfect separation among latent classes, while values close to 0 indicate that most of δ_{it}^{*} are close to $1/2k$, that is like choosing the states in $\mathcal{S}_L$ for both regimes RS and $\overline{RS}$ at random. This index is very suitable for our context where the observed responses are manifest realizations of the latent variables, therefore a good quality in terms of separation of the $2k$ latent states is crucial.
In line with the literature which ignores the answering behavior, we can measure the quality of the separation of the latent construct states marginally with respect to U, so that (2) reduces to:

$$S_k^{L} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(\delta_{it}^{L} - 1/k)}{(1-1/k)nT}, \quad with \quad \delta_{it}^{L} = \max_{l\in\mathcal{S}_L}\sum_{u\in\mathcal{S}_U}\delta_{it}^{(1)}(u,l;\bar{\boldsymbol{\theta}}).$$

Moreover, in our context, the distinguishability among the k states of the latent construct can be interestingly measured at the RS and $\overline{RS}$ regimes separately. The S_k index is specified for this aim as follows:

$$S_k^{L|RS} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(\delta_{it}^{L|RS} - 1/k)}{(1-1/k)nT},\ with\ \delta_{it}^{L|RS} = \max_{l\in\mathcal{S}_L}\frac{\delta_{it}^{(1)}(1,l;\bar{\boldsymbol{\theta}})}{\sum_{l^*\in\mathcal{S}_L}\delta_{it}^{(1)}(1,l^*;\bar{\boldsymbol{\theta}})},$$

$$S_k^{L|\overline{RS}} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}(\delta_{it}^{L|\overline{RS}} - 1/k)}{(1-1/k)nT},\ with\ \delta_{it}^{L|\overline{RS}} = \max_{l\in\mathcal{S}_L}\frac{\delta_{it}^{(1)}(2,l;\bar{\boldsymbol{\theta}})}{\sum_{l^*\in\mathcal{S}_L}\delta_{it}^{(1)}(2,l^*;\bar{\boldsymbol{\theta}})}.$$

3 Data analysis

The practical usefulness of the proposed model is illustrated on data from the Bank of Italy and extensively discussed in Colombi et al, 2021.

The household financial capability is the latent trait of interest that influences the household's decision-making to face financial issues, measured through two observed indicators: the self-perceived ability to make ends meet and the self-report of perceived risk related to financial investments. The way households perceive and disclose their perceptions (affected by *RS* or not) can change over time, allowing also dependence on some demographic and economic household characteristics.

References

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2012). *Latent Markov Models for Longitudinal Data*. CRC Press.

Colombi, R., Giordano, S., and Kateri, M. (2021). Hidden Markov models for longitudinal rating data with dynamic response styles. *AirXiv:2111. 13370v1.*

Van Vaerenbergh, Y. and Thomas, T.D. (2013). Response styles in survey research: a literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, **25**, 195-217.

Alternative Approaches to Dynamic Predictions: An Application on a Cohort of Patients with Chagas disease

Enrico A. Colosimo[1], Emilly M. Lima[1], Maria C. P. Nunes[2]

[1] Statistics Dept., Universidade Federal de Minas Gerais, Belo Horizonte, Brazil,:
[2] Medical School,Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

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

Abstract: Prediction models have been used in several areas, especially in health science. In many situations, they can serve as a tool for clinical support, helping define risk groups by severity and assisting decision-making concerning the most appropriate treatment. This study was motivated by a clinical demand on building a death risk score for patients with Chagas disease, from the SaMi-Trop prospective cohort project. Patients living in state of Minas Gerais, Brazil, were followed for two years, and based on baseline information, a risk score was built to predict 2-year mortality. The follow-up study allowed the nature of the death risk to be dynamic. In this work, we propose four approaches to building dynamic scores: two naive ones, another based on a new landmark and the last one considering a joint modeling approach.

Keywords: Cox model, joint modeling, landmark, risk prediction..

1 Introduction/Motivation

Statistical modeling for clinical data is often used to build prediction models in order to assess the risk of an individual experiencing an event within a given time window. In practice, the most commonly used model for building risk scores is the Cox proportional hazards model, which uses the information from the markers recorded at baseline and a specific time window of interest. We call prediction based on just baseline information as static.
Clinical studies involving a temporal response are prospective leading to new measurements for markers that change over time. An example, which is actually the motivation of this study, is the prospective cohort of the SaMi-Trop Project (*Sao Paulo-Minas Gerais Tropical Medicine Research Center*). This project is a large study involving chagasic heart disease patients living in 21 cities in the north of the state of Minas Gerais, Brazil. Information from these patients, such as age, gender, heart rate and electrocardiogram findings were recorded at the first visit (baseline) between 2013 and 2014.
A risk score was developed using the baseline information and death records after two years of the first visit. Results were recently published in Oliveira et al. (2020). This is exactly the scenario of a static prediction. A second

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

visit took place between 2015 and 2016. At this second visit, all time-dependent markers were updated and, consequently, a dynamic prediction is necessary for those patients who survived for two years.
The aim of this paper is to compare the prediction methodologies in longitudinal survival data and use them to obtain dynamic predictions of death risk for SaMi-Trop cohort patients who survived two years of follow-up.

2 Notation and Cox Regression Model

The Cox regression model the hazard function by

$$\lambda(t \mid \boldsymbol{x}, y(t)) = \lambda_0(t) e^{\boldsymbol{x}^\top \boldsymbol{\beta} + y(t)\gamma}, \tag{1}$$

in which $\lambda_0(t)$ is a baseline hazard function, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters associated with the fixed covariates $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^{\mathrm{T}}$ and γ measures the effect of the time-dependent covariate $y(t)$. Without loss of generality, model (1) considers just one time-dependent covariate.
The estimation of coefficients $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma)$ in Eq. (1) is based on the partial likelihood function $\mathrm{L}(\boldsymbol{\theta})$ (MPLE).
Let's consider $\boldsymbol{x}_0 = \{\boldsymbol{x}, y(t = 0)\}$ a baseline vector of covariates from a new individual. The main interest lies in estimating the risk, $r(t|\boldsymbol{x}_0) = P(T < t|\boldsymbol{x}_0)$, which is the probability of occurrence of an event until time t. Its estimator is obtained by:

$$\hat{r}(t \mid \boldsymbol{x}_0) = 1 - \hat{S}(t \mid \boldsymbol{x}_0) = 1 - \left[\hat{S}_0(t)\right]^{\exp\{\boldsymbol{x}_0^\top \hat{\boldsymbol{\theta}}\}}. \tag{2}$$

where $\hat{\boldsymbol{\theta}}$ is the MPLE and $\hat{S}_0(t) = \exp\{-\hat{\Lambda}_0(t)\}$ is the Breslow estimator of the baseline survival function evaluated at t.
Data splitting and bootstrap methods are used for internal validation to avoid overestimating discrimination and calibration measures (overfitting). We used training and test method and bootstrap to assess the predictive accurary of the baseline model and bootstrap method for the dynamic model. The Area under the ROC Curve (AUC), also known as C statistic, is the most used measure to evaluate the discrimination of a prediction model. We used an estimator for predictive error (PE), proposed by Henderson et al. (2002), for calibration purpose.

3 Dynamic Prediction

There is an interest in obtaining a new estimate for the probability of an individual experiencing the event until a new time u, $u > t$ who survived up to t. Four approaches are presented in this section for the purpose of dynamic predictions.
One way to update the risk is to use the model built at baseline and plug-in the estimator in (2) the new information obtained at time t, $\boldsymbol{x}_t = \{\boldsymbol{x}, y(t)\}$.

***Naive* 1:** the prediction for the individual who survived up to t is made by just inserting the new information from the longitudinal marker in the

formula (2). That is, the estimate is obtained by:

$$\hat{r}_{N_1}(u|t,\boldsymbol{x}_t) = 1 - \left[\hat{S}_0(u-t)\right]^{\exp\{\boldsymbol{x}^\top\hat{\boldsymbol{\beta}}+y(t)\hat{\gamma}\}}. \quad (3)$$

***Naive* 2:** the second method for the dynamic predicted risk takes into account the conditional probability given the individual survival up to t and the covariates updated information:

$$\hat{r}_{N_2}(u|t,\boldsymbol{x}_t) = 1 - \frac{\hat{S}(u \mid \boldsymbol{x}_t)}{\hat{S}(t \mid \boldsymbol{x}_t)}.$$

Landmark: the landmark approach considers that a "new" sample is formed including only individuals who have survived until time t. That is

$$\hat{r}(u|t,\boldsymbol{x}_t) = 1 - \hat{S}_0^*(u-t)^{\exp\{\boldsymbol{x}^\top\hat{\boldsymbol{\beta}}^*+y(t)\widehat{\gamma}^*\}} \quad (4)$$

where, $\hat{\boldsymbol{\beta}}^*$ and $\hat{\gamma}^*$ are the maximum partial likelihood estimators of the Cox model coefficients and $\hat{S}_0^*(u-t)$ is the estimated baseline survival under the new baseline t.

Joint Modeling: in the joint modeling approach, the longitudinal marker process $y(t)$ is assumed to be measured with error such that, for the i-th individual measured at the j-th time point, $j = 1, \ldots, J_i$, its formulation has the following structure:

$$\begin{aligned} y_i(t_{ij}) &= y_i^*(t_{ij}) + \varepsilon_i(t_{ij}) = \boldsymbol{w}_i^T\boldsymbol{\alpha} + \boldsymbol{z}_i^T(t_{ij})\boldsymbol{b}_i + \varepsilon_i(t_{ij}), \\ \varepsilon_i(t_{ij}) &\sim N(0,\sigma^2), \end{aligned} \quad (5)$$

where $y_i^*(t_{ij})$ is the true and unobserved value for the longitudinal marker process at time t_{ij}, $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ are the fixed and time-dependent vectors of covariates associated with the regression coefficients $\boldsymbol{\alpha}$ and random effect $\boldsymbol{b}_i$, respectively. In order to accommodate random effects, it is assumed that $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_b)$ is independent of $\varepsilon_i(t)$. Cox proportional hazards model is now written in terms of $y_i^*(t)$, using the same notation as in (1).
After building the model, we are actually interested in estimating the conditional probability

$$r_{JM}(u|t,\boldsymbol{x}_t) = 1 - S_{JM}(u \mid t,\boldsymbol{x}_t) = 1 - P(T > u \mid T > t, \boldsymbol{x}_t), \quad (6)$$

which can be rewritten as

$$P(T > u \mid T > t, \boldsymbol{x}_t) \quad = \quad \int \frac{S\left[u \mid y^*(u), \boldsymbol{x}_t\right]}{S\left[t \mid y^*(t), \boldsymbol{x}_t\right]} p(\boldsymbol{b} \mid T > t, \boldsymbol{x}_t) d\boldsymbol{b}.$$

A first-order estimator (Rizopoulos, 2011) for the expected survival (6) replaces $\boldsymbol{\theta}^*$ by its estimate $\hat{\boldsymbol{\theta}}^*$ as

$$\widehat{S}_{JM}(u \mid t,\boldsymbol{x}_t) = \frac{\widehat{S}\left[u \mid \hat{y}^*(u), \boldsymbol{x}_t\right]}{\widehat{S}\left[t \mid \hat{y}^*(t), \boldsymbol{x}_t\right]}. \quad (7)$$

To obtain standard errors and confidence intervals for $\widehat{S}_{JM}(u \mid t,\boldsymbol{x}_t)$, Proust-Lima and Taylor (2009) proposed Monte Carlo realizations.

4 Real Data Analysis

Heart rate is the only longitudinal marker that was updated in the second visit. At the current time, survival information after 4 years for the 2-year survivors is only available for 283 patients. The median follow-up time was 63 months and it was observed that 110 (among 1544) patients experienced the event (death) before 2 years and 80 (among 206) after the second visit. The main results concerning discrimination and calibration are presented in Table 1. Validation results are based on 200 bootstrap resamples. AUC estimates are similar for the four approaches, with LM and JM being less precise. The Naive 1 approach is not calibrated and JM has the smallest PE value. In general terms, it seems that JM and Naive 2 have the best results.

TABLE 1. Bootstrap validation results for predicting 4-year death risk for those 2-year survivors.

Approach	AUC (CI 95%)	PE
Naive 1	0.762 (0.736-0.783)	0.074
Naive 2	0.762 (0.736-0.783)	0.068
LM	0.759 (0.673-0.792)	0.070
JM	0.781 (0.667-0.885)	0.060

5 Final Remarks

We propose four different approaches to building dynamic risk scores. Discrimination results are very similar for the four approaches. In terms of calibration, the prediction error estimate is smaller for the Naive 2 and JM approaches and Naive 1 has poor calibration. All approaches have some limitations under the SaMi-Trop cohort data set. Probably the most restricted one is the LM since there is a substantial in the sample size in order to take into account the new baseline at two years of follow-up.

References

Oliveira, et al. (2020). Risk score for predicting 2 year mortality in patients with Chagas Disease cardiomyopathy from endemic area . *Jour, Am. Heart Assoc.*, **9(6)**.

Proust-Lima, C. and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach *Biostatistics*, **10(3)**, 535 – 549.

Henderson, R., Diggle, P. and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatiistics*, **3(1)**. 33 – 50.

Rizopulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal an time-to-event data. *Biometrics*, **67(3)**, 819 – 829.

Multi-state models as an alternative to joint models; estimating treatment effects using longitudinal data

Ilse Cuevas Andrade[1], Ardo van den Hout[1], Nora Pashayan[2]

[1] Department of Statistical Science, University College London, UK
[2] Department of Applied Health Research, University College London, UK

E-mail for correspondence: ilse.andrade.21@ucl.ac.uk

Abstract: An area of clinical interest is understanding the association between a time-dependent biomarker and a time-to-event outcome. To model this association joint models define a linear mixed model for the longitudinal outcome and a Cox model for the time-to-event outcome. As an alternative, following a multi-state framework opens up new modelling perspectives. By discretising the longitudinal outcome, and using a continuous Markov model, interval-censored observations can be considered as well as flexible modelling approaches that take into consideration the biomarker thresholds and the time-dependency of the process. Data from a randomised clinical trial for liver cirrhosis patients are used to compare both methodologies with respect to estimation of treatment effects.

Keywords: Survival analysis; Continuous Markov model; Time-dependent covariates.

1 Introduction

Clinical trials commonly include longitudinal data, including data on biomarkers. It is of biomedical interest to understand the association between a biomarker and the survival outcome. The Cox regression model for a time-to-event can investigate the effects of longitudinal biomarkers. However, the model cannot be used for prediction as it does not model the longitudinal process of the biomarker; alternative modelling frameworks are joint models and multi-state survival models.
Using data from a randomised clinical trial with 488 patients histologically verified with liver cirrhosis, we analyse and compare the treatment effect outcome using multi-state models and joint models.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

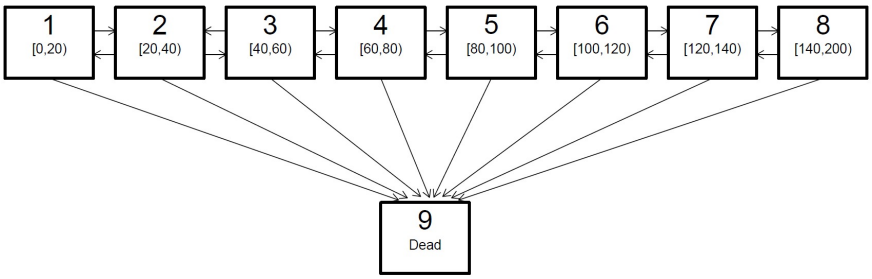


FIGURE 1. Nine-state model for the Liver Cirrhosis data. Living states are defined by discretising the Prothrombin index.

For the joint model methodology, we follow the model proposed by Rizopoulos (2012). A linear mixed model with cubic spline effect of time, with different average profiles per treatment group is fitted.
For the multi-state methodology; we define eight living states by discretising the Prothrombin index (PI) . The Prothrombin index is a blood test of coagulation factors produced by the liver, and it is a marker of severity of liver disease. The death state is the ninth state, see Figure 1. Time-dependency of the process is modelled using Gompertz baseline hazard functions and using B-splines.

2 Multi-state model

Let $\{Y_t \mid t \in (0, \infty)\}$ be a continuous Markov chain on the state space S, and let $P(t, u)$ be the $D \times D$ transition probability matrix with entries $p_{rs}(t, u) = P(Y_t = s \mid Y_u = r)$, for $0 \leq t \leq u$, and $r, s = 1 \dots D$.
The transition probabilities can be derived from the transition hazards, $P(t, u) = \exp((u - t)Q(t))$, where Q is the $D \times D$ generator matrix with off-diagonal (r, s) entries q_{rs} and diagonal entries $q_{rr} = -\sum_{s \neq r} q_{rs}$.
The transition-specific hazards can be defined combining a baseline hazard with log-linear regression

$$q_{rs}(t \mid x) = q_{rs.0}(t) \exp\left(\beta_{rs}^{\top} x\right) , \tag{1}$$

where $\beta_{rs} = (\beta_{rs.1}, \dots, \beta_{rs.p})^{\top}$ is a parameter vector, and $x = (x_1 \dots x_p)^{\top}$ is the covariate vector. The baseline hazard $q_{rs.0}(t)$ describes the hazard's time-dependency.

2.1 Estimation

Considering interval-censored observations, the maximum likelihood is constructed with the transition probabilities (Kalbfleisch and Lawless, 1985).

Assuming the transition into the death state D is known, and letting the living states be indexed by $1, 2 \dots D-1$. For an individual i with observation times $t_1, \dots, t_n$, and an observed trajectory of states $y_1, \dots, y_n$, the likelihood contributions for interval the (t_{j-1}, t_j) is given by

$$\begin{aligned} L_i(\theta \mid y, x) &= P\left(Y_J = y_J, \dots, Y_2 = y_2 \mid Y_1 = y_1, \theta, x\right) \\ &= \left(\prod_{j=2}^{J-1} P\left(Y_j = y_j \mid Y_{j-1} = y_{j-1}, \theta, x\right) \right) C\left(y_J \mid y_{J-1}, \theta, x\right) , \end{aligned}$$

where θ is the vector of parameters, and x is the covariate vector. If t_j is a living state then $C\left(y_j \mid y_{j-1}, x\right) = P\left(Y_j = y_j \mid Y_{j-1} = y_{j-1}, \theta, x\right)$, and if death is observed at t_j then

$$C\left(y_j \mid y_{j-1}, x\right) = \sum_{s=1}^{D-1} P\left(Y_j = s \mid Y_{j-1} = y_{j-1}, \theta, x\right) q_{sD}\left(t_{j-1} \mid \theta, x\right) .$$

The probability transition matrix is computed using the eigenvalue-decomposition method, and the log-likelihood is maximised using the Nelder-Mead optimisation method from the general purpose optimiser `optim`, R-package.

3 Application

The data used are from a randomised clinical trial for liver cirrhosis patients, the outcome was survival and the aim was to analyse if the hormonal treatment prednisone improved survival for patients with cirrhosis. Patients were monitored by measuring blood Protrombin level. This dataset is available in the `JM`, R-package.
The number of patients considered in the study was 488; 251 received prednisone and 237 received placebo treatment. The patients were followed-up for 12 years. The follow-up visits were every 3 months for the first year and annually afterwards. However, the visits are rather irregular.
The Prothrombin index is a measurement based on a blood test of coagulation factors produced by the liver. The cut-offs reported within the normal range are from 70% to 100%, having a low Prothrombin index is associated with severe liver fibrosis. Having an abnormal Prothrombin index is reversible, patients can return to a normal value.
For the joint model approach, the longitudinal outcome can be modelled using a linear mixed model with a natural cubic spline effect for time.
The linear treatment effect estimate associated with the risk for death is $\gamma = 0.209$ with a standard error of 0.1401. For further details of this model please see Rizopoulos (2012, Chapter 5).

The multi-state model includes treatment as a covariate for the forward, backward, and dead transitions. Now, taking into account the normal range cut-offs of the Prothrombin index, we propose an innovative modelling perspective by including a new variable d. This additional variable allows modelling the biomarker threshold and its association with the risk for death. Given that the normal ranges of this biomarker $[70, 100]$ are defined in the middle states of the process, we define d as the absolute difference between the current state and 4.5. Recall, that state 4 is defined for a Prothrombin index between $[60, 80]$ and state 5 between $[80, 100]$.
Furthermore, the effect of time is taken into consideration using Gompertz baseline for the backward transitions and death. The hazards for this model are defined as

$$\begin{aligned} q_{r,r+1}(t) &= \exp\left(\beta_1 + \gamma_1 \texttt{trt}\right) && \text{for } r = \{1,2,3,4,5,6,7\} \\ q_{r,r-1}(t) &= \exp\left(\beta_2 + \gamma_2 \texttt{trt} + \xi_1 t\right) && \text{for } r = \{2,3,4,5,6,7,8\} \\ q_{r,9}(t) &= \exp\left(\beta_3 + \gamma_3 \texttt{trt} + \xi_2 t + \eta_1 \mathrm{d}\right) && \text{for } r = \{1,2,3,4,5,6,7,8\}, \end{aligned} \tag{2}$$

where `trt` is equal to 1 if the treatment given was prednisone, and zero otherwise, and $d = \mid r - 4.5 \mid$. Model selection was made using the AIC value criteria. The AIC value for this model is 9511.22. The estimated value of the covariate associated with the treatment effect for the risk for death is $\gamma_3 = 0.295$ with a standard error of 0.115.
As shown above, both frameworks yield a similar estimate value of the treatment effect associated with the risk for death. However, Figure 2 shows the effect of including the distance to the middle states of the process (d) in the hazard for death. Having a low $[0, 20]$ or high $[140, 200]$ PI implies a higher hazard for death than having a PI index between 60 and 100.
As an extension of the above, B-splines can be fitted to model the time-dependency of the process allowing for a more flexible baseline hazard. The knots are equidistant, and the number of knots and the degree of the polynomial segments can be chosen ad-hoc based on features of the data and computational aspects (Eilers and Marx 1996).
For instance, the hazard for death can be modelled as follows

$$q_{r,9}(t) = \exp\left(\sum_{k=1}^{K} \alpha_{r9.k} B_k(t) + \gamma_3 \texttt{trt} + \eta_1 \mathrm{d}\right),$$

where K is the number of knots, and $B_k(t)$ are the spline basis matrices, for $r = \{1,2,3,4,5,6,7,8\}$.
The left plot in Figure 2 shows the estimated hazards for death from state 1 and state 4 using B-splines with a third-degree polynomial and $K = 7$. The AIC value of the model is 9566.55.
As expected, in Figure 2 we can see that the same effect concerning the value of d in the hazard for death is consistent in both estimated hazards;

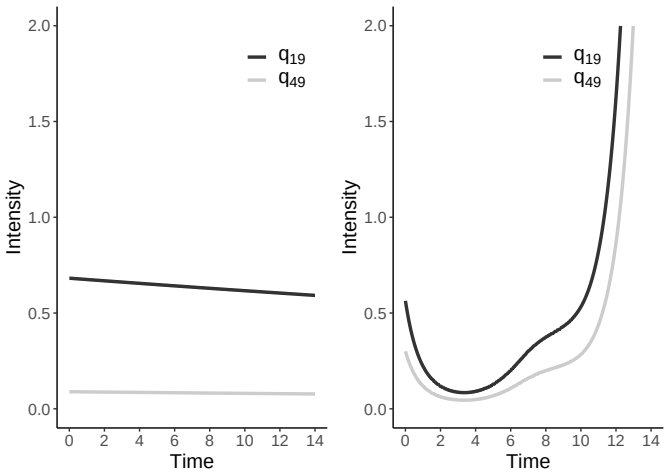


FIGURE 2. Parametric Gompertz hazard for death from state 1 (black curve) and state 4 (grey curve). Estimated hazard for death from state 1 (black line) and from state 4 (grey line) using B-splines with a third-degree polynomial and $K = 7$ for patients that received prednisone treatment from the Liver Cirrhosis data.

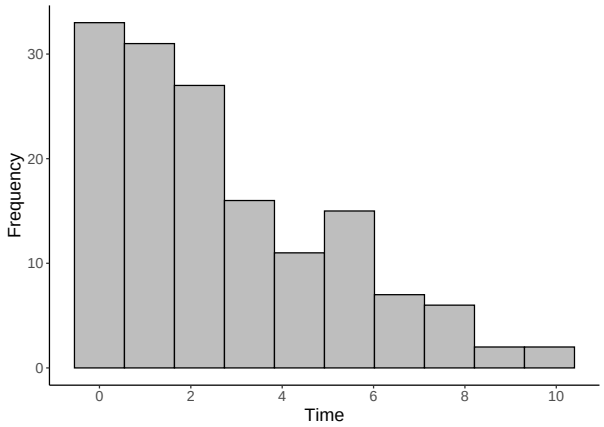


FIGURE 3. Histogram of deaths from patients that received prednisone treatment from the Liver Cirrhosis data.

being at the boundaries of the process implies a higher hazard of death than being at state 4 or 5.

B-splines can perform poorly if there is sparse or missing data which is typically at the lower and upper bound of the data. In this case, from Figure 3 we can see that the highest number of deaths happened in the first

two years of the study and the latest observed death is in the year tenth. However, since the latest observation (censored) is in the year thirteen, the splines are fitted for a timeline from zero to fourteen. Therefore, the lack of observations after year tenth explains the behaviour at the boundaries of the estimated hazards in Figure 2.
Despite B-splines not being optimal at the boundaries, the comparison illustrates the modelling flexibility that B-splines can add to the estimated hazards, and the AIC value provides some evidence that the Gompertz hazard is a good choice for estimating the hazard for death.
As mentioned above, the number of B-splines and the degree of the polynomial segments are chosen based on the data features and computational aspects. One way to address the question of how many splines to use is by controlling the smoothness of the likelihood by using P-splines (Eilers and Marx 1996), this alternative will be further explored.
In conclusion, following a multi-state approach allows us to model the time-dependent biomarker from a different perspective by discretising it, enhancing clinical research considerations such as modelling the biomarker threshold. Also, the time-dependency of the process can be considered using flexible modelling approaches such as B-splines.
Furthermore, under a multi-state framework, there is no need to compute marginal estimates to derive population-based inference as it is when using joint models.

Acknowledgments: The research was funded by the National Cancer Institute Grant $\#U01CA253915$. The contents are solely the responsibility of the authors and do not necessarily represent the official views of PTE or the Awarding Agency.

References

Eilers, P. H. C., & Marx, B. D. (2021). *Practical Smoothing: The Joys of P-splines.* Cambridge: Cambridge University Press.

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.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: with applications in R.* New York: Chapman and Hall/CRC.

Variable Selection with Quasi-Unbiased Estimation: the CDF Penalty

Daniele Cuntrera [1], Vito M.R. Muggeo [1], Luigi Augugliaro [1]

[1] Università degli studi di Palermo, Dip.to Sc Econom, Az e Statistiche, Palermo

E-mail for correspondence: daniele.cuntrera@unipa.it

Abstract: We propose a new non-convex penalty in linear regression models. The new penalty function can be considered a competitor of the LASSO, SCAD or MCP penalties, as it guarantees sparse variable selection while reducing bias for the non-null estimates. We introduce the methodology and present some comparisons among different approaches.

Keywords: Variable selection, non-convex penalty function, LASSO, SCAD, MCP

1 Introduction

Nowadays, variable selection is an essential issue in regression modelling as high dimensional data. Modern applications of statistical theory and methods can involve massive data sets, often with a massive number of variables measured on a relatively small number of experimental units (Johnstone and Titterington, 2009). Methods for analyzing high-dimensional data are being developed in various fields of science, such as Bioinformatics and Genomics. Regardless of the field of application, the goal of all these methods is to identify a subspace of the data that contains all the valuable information. Therefore, spotting noise and significant covariates is a major concern when studying the effect on the response variable of interest. The LASSO penalized regression model is one of the most widely used techniques (Tibshirani, 1996), but unfortunately, the selected coefficients suffer from substantial bias (Fu and Knight, 2000). In this paper, we present a new non-convex penalty function based on the standard normal cumulative distribution: for this reason, we will refer to it as the CDF penalty function. We will show that it inherits part of the advantage of both LASSO and SCAD.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 The proposal: the cumulative distribution function penalty

Let be data (y_i, x_i), $i = 1, ..., n$, where $x_i = (x_{i1}, ..., x_{ip})^T$ are the covariates and y_i is the response. Operating in linear regression framework, we assume independent observations, and that the set of covariates x_i is standardised. In a classical linear regression context, the dependence of a response variable on one or more covariates is modeled as $y_i = x_i^T \beta + \epsilon_i$, where β is the parameter vector and ϵ_i are independent and identically distributed random Gaussian noise. The ordinary least squares (OLS) estimates are obtained by minimizing the sum of the model residuals. Penalized regression models are based on adjusting the trade-off between bias in the estimates and their variance, so they mediate the need to have a model close to the data and reduce the estimated coefficients' variance. Following Fan and Li (2001), in linear regression models, the penalized objective loss function can be defined as follows

$$\frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \sum_{j=1}^{p} p(\beta_j, \lambda) \tag{1}$$

where $p(\beta, \lambda)$ is the penalty function that allows for selecting variables, thus reducing the estimates' variance. The tuning parameter λ determines the weight of the penalty in optimizing the function: high tuning parameter values cause some estimates to be bound to zero; reducing its value decreases the number of coefficients bounded to zero. In general, the problem is to find the optimal intermediate value.
We propose to specify the penalty term in (1) by using a transformation of the cumulative distribution function of the standard Normal random variable. We define the penalty as

$$p(\beta_j, \lambda) = \lambda \; p(\beta_j) = \lambda \sqrt{2\pi} \kappa \Phi \left(\frac{|\beta_j|}{\kappa} \right). \tag{2}$$

The CDF penalty (2) ensures variable selection (due to singularity at 0, as LASSO, SCAD and MCP). κ influences the rate of convergence of coefficient estimates to maximum likelihood estimates of the selected model and in practice it is chosen to ensure continuity of the coefficients path. The proposed penalty is intended to mimic some of the good features of SCAD (Fan and Li, 2001) and MCP (Zhang, 2010) penalties by reducing the bias of the LASSO. Figure 1 shows the shape of SCAD, MCP, LASSO and CDF penalty. The CDF penalty has been represented for three different values of κ. It can be seen that our proposal is a trade-off between the LASSO (which is obtained for higher values of κ) and the SCAD and MCP penalties (which are obtained for smaller values of κ).
There are several advantages of (2) with respect to SCAD or MCP, the most notable being that (2) is *multiplicative*, namely it is possible to write

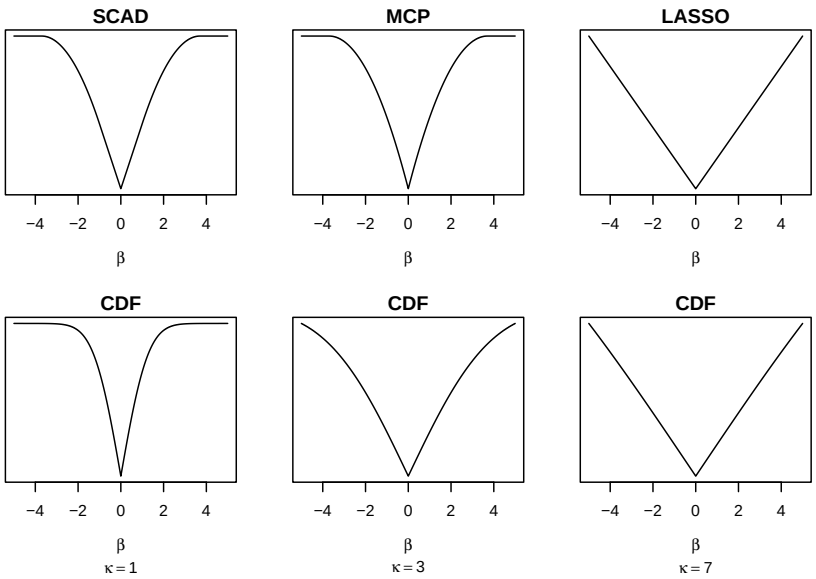


FIGURE 1. The CDF penalty at different values of the parameter κ.

$p(\beta, \lambda) = \lambda\, p(\beta)$; therefore the shape of penalty is independent of λ and it does not change along the coefficient path.
Recalling Fan and Li (2001), a penalty function has to satisfy the following conditions to have the properties of sparsity, continuity and nearly unbiasedness:

1. $\lim_{\beta \to +\infty} p'_\lambda(|\beta|) = 0$;
2. $\min[|\beta| + p'_\lambda(|\beta|)] > 0$;
3. the minimum of $|\beta| + p'_\lambda(|\beta|)$ is attained at 0.

The first-order derivative of CDF penalty is equal to $\exp -\left\{\frac{|\beta|^2}{2k}\right\}$, so it is easy to see that the first condition is guaranteed (independently from the value of κ). In figure 2 is reported the plot of $|\beta| + p'_\lambda(|\beta|)$ for three different values of κ. Independently on κ, the minimum value is equal to 1, and it is attained always on 0. So, the second and the third conditions are guaranteed. We can conclude that the CDF penalty enjoys sparsity, continuity, and nearly unbiasedness properties.
Therefore the proposed sparse and quasi-unbiased estimator of the regression coefficient vector is defined

$$\hat{\beta} = \arg\min \left\{ \frac{1}{2n} \sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_j \sqrt{2\pi}\kappa \Phi(|\beta_j|/\kappa) \right\}. \qquad (3)$$

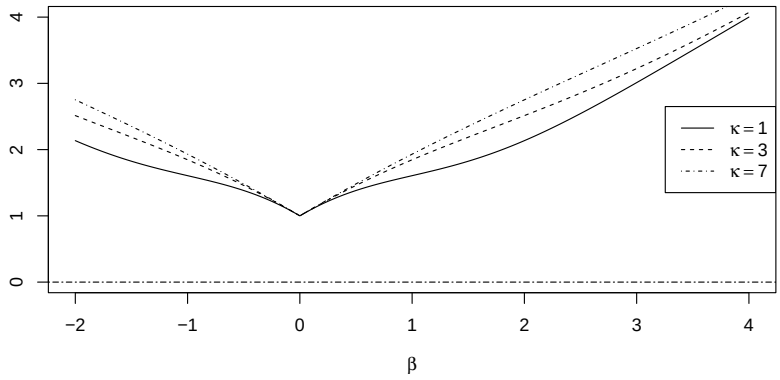


FIGURE 2. Plot of $|\beta| + p'_\lambda(|\beta|)$ for different values of κ.

To solve (3), we propose to use the ADMM (Alternating Direction Method of Multipliers) algorithm. The problem can be rewritten as follow

$$min\, f(\beta) + g(\tilde{\beta}) \qquad \text{s.t. } \beta - \tilde{\beta} = 0$$

where

$$f(\beta) = \frac{1}{2n}\|y - X\beta\|^2 \qquad g(\tilde{\beta}) = \lambda \sum_{j=1}^{p} \sqrt{2\pi}\kappa\Phi\left(\frac{|\tilde{\beta}_j|}{\kappa}\right). \tag{4}$$

The augmented scaled Lagrangian function to be optimized is

$$\mathcal{L}_\rho(\beta, \tilde{\beta}, \gamma) = f(\beta) + g(\tilde{\beta}) + \frac{\rho}{2}\|\beta - \tilde{\beta} + \gamma\|_2^2. \tag{5}$$

where $\rho > 0$ is a fixed constant only affecting the convergence speed and not the final solution. Once the problem has been formalized, the ADMM algorithm alternates, until convergence, the followings

$$\beta^{k+1} = \arg\min_{\beta} f(\beta) + \frac{\rho}{2}\|\beta - \tilde{\beta} + \gamma\|_2^2 \tag{6}$$

$$\tilde{\beta}^{k+1} = \arg\min_{\tilde{\beta}} g(\tilde{\beta}) + \frac{\rho}{2}\|\beta - \tilde{\beta} + \gamma\|_2^2 \tag{7}$$

$$\gamma^{k+1} = \beta - \tilde{\beta} + \gamma. \tag{8}$$

3 Simulation study

A simulation study is conducted to compare the performance of the proposed penalty with respect to alternatives already established in the literature (LASSO, SCAD and MCP). Performance is measured by mean square error (MSE) and false discovery rate (FDR). For the true parameter $\beta_0 = (1, 1, 1, 1, 1, 0, ..., 0)^T$, and $p = 100$ covariates $x_i \sim \mathcal{N}(0, \Sigma)$ with the Toeplitz correlation matrix $\Sigma_{jk} = 0.5^{|j-k|}$, the response variable is generated as $y_i = x_i^T \beta_0 + \sigma_k \epsilon_i$ where $n = 50$, $\epsilon_i \sim \mathcal{N}(0, 1)$, and the σ_k are twelve different equally-spaced values ranging in $[0.25, 3.0]$ of the random noise variance regulating the signal-to-noise ratio. For each scenario reported, 500 simulations were ran and at each replicate the optimal λ for each penalty has been selected by minimization of BIC using the number of non-null estimates as degrees of freedom. Moreover the additional parameters for SCAD and MCP are fixed at the usual values of 3.7 and 3 respectively, while the κ parameter of CDF is computed as a known function of the selected λ.

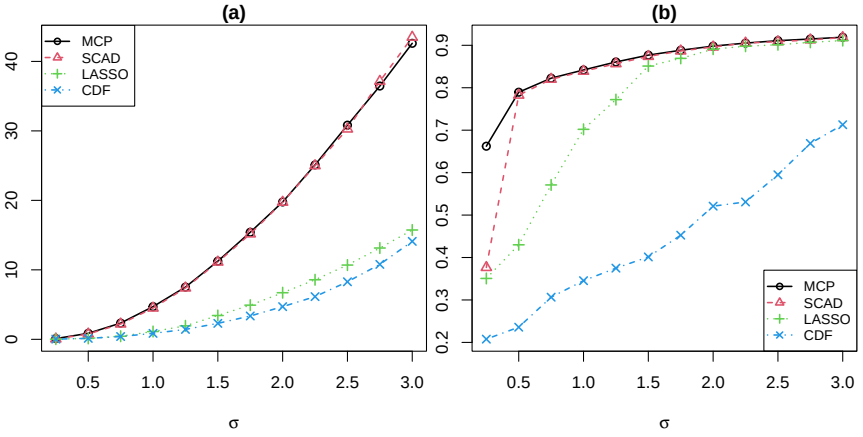


FIGURE 3. Simulation results at different values of signal-to-noise ratio.

Results of the simulations, in terms of MSE and FDR, are shown in Figure 3. Looking at the MSE values of SCAD and MCP (panel (a)), they have similar values that are much higher than those of LASSO and CDF penalties. Our proposal has the lowest value of MSE across all the σ values. Also in terms of FDR (panel (b)), the CDF penalty outperforms the considered competitors: the values observed are much lower than those obtained with LASSO, SCAD and MCP.

4 Conclusion

In this paper we have presented a new SCAD (or MCP) type penalty: it is based on the cumulative distribution function of the standard normal with an additional shape parameter (κ). Our proposal borrows the stability of LASSO, keeping some important features of SCAD and MCP in terms of unbiasedness of the non-null estimates. Based on some simulation experiments, the results show that our proposal outperforms the competitors in terms of both MSE and FDR.

References

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *JASA*, **96**, 1348-1360.

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

Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philos. Trans. R. Soc. A*, **367**, 4237-4253.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc Series B*, **58**, 267-288.

Zhang, C-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, **38**, 894-942.

Exploring relationships using non-linear correlation coefficients

Fernanda De Bastiani[1], Mikis D. Stasinopoulos[2], Robert A. Rigby [2], Thomas Kneib[3]

[1] Federal University of Pernambuco, Recife - Brazil
[2] London Metropolitan University, London - United Kingdom
[3] Georg-August University, Göttingen - Germany

E-mail for correspondence: fernanda.bastiani@ufpe.br

Abstract: The use of correlations coefficients is limited to linear relationships but with the explosion of smoothing techniques and machine learning methodology the need for pre- and post-fit identification of non-linear relationships between the explanatory terms (features) is urgently needed. In this paper, we consider alternative measures of association which could be used to replace/supplement the correlation coefficients for non-linear relationships. We aim to explore and compare the techniques available in the literature and provide practical advice for the user. It is ongoing research.

Keywords: concurvity; GAMLSS; identifiability; Pearson correlation coefficient; P-splines.

1 Introduction

During our statistical training, we learn that correlation coefficients are only good for detecting linear relationships. Unfortunately, within the data science community often people use the correlation coefficients indiscriminately for detecting any relationships between the features in the data. To add to the confusion, in his more recent popular book "Helgoland", the physicist Carlo Rovelli uses the term "correlation" to describe what we would be called in statistics a "relationship" between variables. It seems that there is scope for the statistical community to go back to the basics and rethink the use of correlations and its contribution to statistical modelling. This paper is such an attempt.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

Linear correlation coefficients are well established in the statistical literature as a measure of the strength of linear relationship between two variables. Our motivation for studying non-linear correlations stems from the fact that if there is a high correlation between the explanatory variables of a statistical model, this often results in instability and difficulty of interpreting a fitted model for a response variable. For linear models the problem is well established under the term *collinearity* or the more general term *multicollinearity* when more than two variables are involved.
For additive smoothing terms, Hastie and Tibshirani (1990), used the term *concurvity* to describe the problem of non-linear relationships between two explanatory variables involved in the model, and the term *multiconcurvity* can be used when more that two variables are involved. It is generally recognized that there are two major problems associated with concurvity: the first has to do with fitting the additive model especially if one uses backfitting; the second has to do with interpreting the model since concurvity could make the fitted model unstable and sensitive to small changes in the explanatory variables. The first problem is reduced by 'modified' backfitting, see Hastie and Tibshirani (1990) and Stasinopoulos *et. al.* (2017, Ch. 3). The second is the main reason for the exploration for non-linear correlations in the features in this paper. A "non-linear correlation" between the covariates would help to expose (multi)concurvity and possible pitfalls in the interpretation of a model.

2 A review of non-linear correlation coefficients methods

As with linear correlation coefficients, here we are looking for summary measures (and possibly graphical tools) which could flash up when a non-linear relationship exists between the covariates.
The relationship between two or more variables can be i) *pairwise* ii) *full*, that is, between one of the term and the rest, and iii) pairwise *partial* of two terms correcting for the contribution of other terms.
This paper concentrates on pairwise non-linear correlations, but at a later stage generalizations will be considered. Non-linear relationships can be in parts negative and in other parts positive, therefore, unlike the well known linear correlation coefficients that returns a value between, -1 to 1 to reflect the direction of the association, the non-linear correlation should take values from 0 to 1 indicating whether there is a strong non-linear relationship (close to one) or a weak non-linear relationship (close to zero). Besides the well known Pearson, Kendall and Spearman correlation coefficients, we considered smoothing techniques, concurvity measures, canonical correlation, mutual information, maximal and distance correlations to measure the association between two variables.

The first idea is to fit variable x_2, as explanatory term, to x_1, using a smooth function of x_2, in order to model their pairwise non-linear relationship and then calculate the correlation coefficient of the fitted values of $\hat{x}_1$ with x_1, i.e. $r_{1,2}$, as the non-linear correlation between them. The asymmetry of the relationship creates problems since interchanging x_1 and x_2 can produce a completely different non-linear correlation. A simple solution, is to take the maximum of the two correlation coefficients as the final measure of non-linear association, i.e. $\texttt{max}\{r_{1,2}, r_{2,1}\}$. The following are the three different smoothing techniques used in this paper. i) Regression trees using the predictive power score, (**pps**), of the `R` package **ppsr** (van der Laken, 2021); ii) dynamic partition, (**dp**), using the `R` package **nlcor** which regresses x_1 against x_2 using linear segmental regression (Ranjan and Najari, 2022); and iii) P-splines, Eilers and Marx (1996), (**pb**), where the smoothing parameter is selected using restricted maximum likelihood.

A concurvity measure can be defined as a measure of statistical dependency among covariates. According to Wood (2017), concurvity occurs when some smooth term in a model could be approximated by one or more of the other smooth terms in the model. The first approach here is based on the `concurvity()` function from the **mgcv** `R` package. It analyses the linear manifolds of the additive term bases $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_{J_k}$. It tries to identify whether a manifold of one of the smoothers, $\mathcal{M}(\mathbf{B}_i)$, is close to the manifold of another $\mathcal{M}(\mathbf{B}_j)$, for $i \neq j$, by fitting a multivariate least squares model and then calculating the R-square statistic using the fitted sum of squares divided by the total sum of squares. The sums of squares are defined by the Frobenius norm, (**F**), given by $||\mathbf{A}||_F = [trace(\mathbf{A}^\top \mathbf{A})]^{1/2}$. The second method uses canonical correlation analysis, **cc**, on the manifolds $\mathcal{M}(\mathbf{B}_i)$ and $\mathcal{M}(\mathbf{B}_j)$ and reports the first (highest) correlation coefficient between them. A similar approach is used also by Huang *et. al.* (2009). Note the bases $\mathbf{B}_i$ for $i = 1, 2$ were created using equal space B-splines with 15 number of knots.

In information theory the concept of *entropy*, $H(x) = -E_p \log p(x)$ measures the randomness of (one or more) random variables x which have probability function $p()$ (Cover and Thomas, 1999). Note that entropy is defined by probabilities not by the values of x. Low entropy is associated with more order, while high with more randomness. Here we used the maximal information coefficient approach, **MIC**, of Reshef *et al.* (2011) which uses the concept of *mutual* information. Note that this approach require the continuous covariates to be discretized. The maximal correlation, **mC**, is defined as $mC = \max_{f,g} Cor(f(x), g(y))$ for smooth functions $f()$ and $g()$ which can be obtain using the R package **acepack**. The distance correlation **dC** which uses distance matrices for both variables is obtained using the `dcor()` function from package **energy**.

3 Examples

To check the suitability of the different measures of non-linear correlation we used 12 different examples. Figure 1 shows the twelve different cases of linear and nonlinear relationship between two explanatory variables. Cases 1, 4, 5, 6, 8, 10 and 11 seem to be the most likely cases to create problems in the fitting and the interpretation of a model for a response variable. Case 9 (the double moustache) and case 12 (the doughnut) are very unlikely to occur in practice but they are of interest for checking non-linear association.

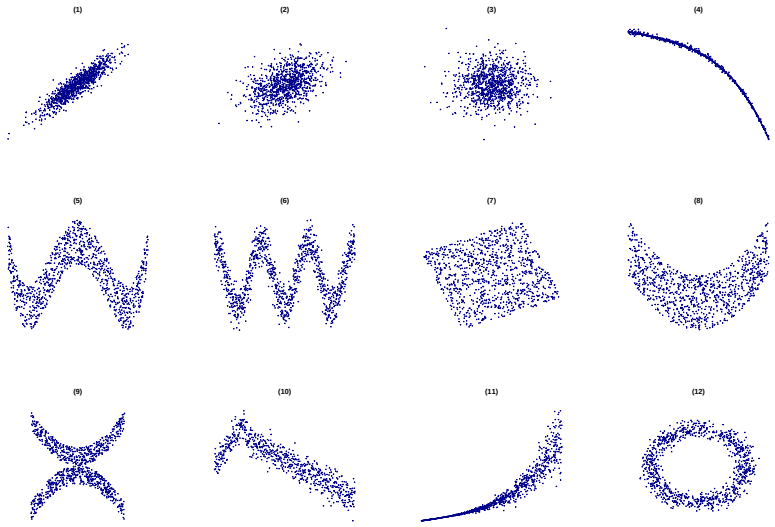


FIGURE 1. Investigated cases of linear and non linear correlation between two variables. The cases numbers are the same as in Table 1.

The estimated coefficients using the methodology described in Section 2 are given in Table 1. As expected, the Pearson, Spearman and Kendall correlation coefficients perform well in the linear cases 1, 2, 3, but also in the non-linear cases 4, 10 and 11. However they fail to detect the relationship in the non-linear cases of 5, 6, 8, 9 and 12.
From the smoothing techniques, the regression trees method **pps** performed rather poorly compared to the dynamic partition **dp** and P-splines **pb** methods. The **dp** and **pb** performed similarly and well in most cases, apart from cases 9 and 12 where they failed to capture the relationship.
Both concurvity measures **F** and **cc** performed well, with realistic coefficients, apart from the case 3 where they should be close to zero but they instead at values 0.119 and 0.343, respectively. In particular they captured the non-linear relationships in cases 9 and 12, although the coefficient 1 for

TABLE 1. Comparison between correlation coefficients for the artificial data. The notation used is r: Pearson's, τ: Kendall's and ρ: Spearman's correlations; **pps**: predictive power score; **dp**: dynamic partition; **pb**: P-splines; **F**: concurvity using the Frobenius norm; **cc**: concurvity using canonical correlation; **MIC**: maximal information coefficient; **mC**: maximal correlation; **dC**: distance correlation.

Case	r	τ	ρ	**pps**	**dp**	**pb**	**F**	**cc**	**MIC**	**mC**	**dC**
1	0.88	0.69	0.87	0.50	0.88	0.88	0.79	0.89	0.62	0.88	0.85
2	0.33	0.21	0.31	0.04	0.33	0.38	0.18	0.42	0.19	0.33	0.29
3	-0.08	-0.05	-0.08	0.00	0.08	0.08	0.12	0.34	0.14	0.09	0.08
4	-0.94	-0.97	-0.99	0.85	0.94	1.00	1.00	0.99	1.00	0.99	0.96
5	0.04	0.00	0.02	0.46	0.84	0.87	0.80	0.89	0.77	0.89	0.37
6	0.00	0.00	0.00	0.51	0.88	0.90	0.85	0.92	0.77	0.92	0.15
7	-0.03	-0.02	-0.04	0.05	0.35	0.31	0.17	0.41	0.19	0.31	0.16
8	-0.01	0.00	-0.00	0.24	0.68	0.71	0.54	0.74	0.39	0.72	0.39
9	-0.03	-0.02	-0.03	0.01	0.03	0.06	0.85	0.92	0.43	0.91	0.85
10	-0.83	-0.65	-0.84	0.59	0.89	0.93	0.88	0.94	0.82	0.94	0.86
11	0.91	0.93	0.99	0.76	0.91	0.99	1.00	0.99	0.99	0.98	0.94
12	0.04	0.01	0.03	0.00	0.04	0.06	0.71	0.84	0.36	0.83	0.16

F in case 11 seems too extreme. Unfortunately they were very sensitive on the number of knots used. The **MIC** and **dC** performed reasonably well, but they failed to capture the doughnut case 12. The **mC** performed very well in all cases.

4 Conclusions

The maximum correlation, **mC**, performed best, providing a realistic non-linear measure of association (between two features or explanatory variables) in all the examples considered. The Frobenius norm, **F**, from the concurvity techniques also performed well, followed by the maximal information coefficient, **MIC**.

The canonical correlation, **cc**, from concurvity, performed well except for the random scatter (case 3), where its correlation value was too high. The distance correlation, **dC**, performed well, apart from the doughnut (case 12), while the dynamic partition, **dp**, and the P-splines, **pb**, from the smoothing techniques, performed well, apart from the double moustache (case 9) and the doughnut (case 12).

All of measures are relatively fast to calculate. One possibility is to combine two or more measures of association together, in order capture different features of the association. There is more investigation and verification to be done in the near future.

We conclude with the following notes. The construction of any correlation coefficient between two variables x and y seems to follow four steps: i) transformations of the variables involved: i.e. $(x, y) \to (\tilde{x}, \tilde{y})$, ii) a definition of a metric $< \tilde{x}, \tilde{y} >$, iii) use of the metric for the definition of the correlation measure $\Gamma = \frac{<\tilde{x},\tilde{y}>}{||\tilde{x}||||\tilde{y}||}$ where $||\ ||$ is a norm, and iv) test whether the metric is zero or not. For a proper comparison of non-linear correlations all steps should be investigated and evaluated properly. Note also that the in the

current paper the *maximum local correlation integral* approach of Chen *et al.* (2010) has been omitted because it does not provide automatically a measure of association in the interval $(0, 1)$, but only provides a test for checking the non-llinear association.

Acknowledgments: De Bastiani acknowledges L'oréal Brasil, UNESCO and ABC, the Brazilian agencies CAPES, CNPq and FACEPE and UFPE.

References

Chen, Y.A., Almeida, J.S., Richards, A.J., Muller, P., Carroll, R.J., Rohrer, B. (2010). A nonparametric approach to detect nonlinear correlation in gene expression. *Journal of Computational and Graphical Statistics*, 19 (3), pp.552–568.

Cover, T.M., J. A. Thomas (1999). *Elements of information theory*. John Wiley & Sons.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models*. Routledge.

Huang, S.Y., Lee, M.H. and Hsiao, C.K., (2009). Nonlinear measures of association with kernel canonical correlation analysis and applications. *Journal of Statistical Planning and Inference*, 139(7), pp.2162-2174.

Eilers, P.H. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 11 (2), pp.89-121.

Ranjan, C. and Najari, V. (2022). nlcor: Compute Nonlinear Correlations. R package

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M. and Sabeti, P.C., (2011). Detecting novel associations in large data sets. *Science*, 334(6062), pp.1518-1524.

Rovelli, C. (2021) *Helgoland*, Flammarion.

Stasinopoulos, D.M., Rigby, R.A., Heller, G.Z., Voudouris, V. and De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in* R. Chapman and Hall.

van der Laken, P.(2021) ppsr: Predictive Power Score. R package

Wood, S.N. (2017) *Generalized Additive Models: An Introduction with R* (2nd edition). Chapman and Hall/CRC.

Flexible joint model for time-to-event and non-Gaussian longitudinal outcomes

Hortense Doms[1], Philippe Lambert[1,2], Catherine Legrand[1]

[1] Institut de Statistique, Biostatistique et Sciences Actuarielles, Université catholique de Louvain, Belgium
[2] Institut de Mathématique, Université de Liège, Belgium

E-mail for correspondence: hortense.doms@uclouvain.be

Abstract: In medical studies, repeated measurements of biomarkers and time-to-event data are often collected during the follow-up period. To assess the association between these two outcomes, joint models are frequently considered. The most common approach uses a linear mixed model for the longitudinal part and a proportional hazard model for the survival part. The latter assumes a linear relationship between the survival covariates and the log hazard. In this work, we propose an extension allowing the inclusion of non-linear covariate effects in the survival model using Bayesian penalised B-splines. Our model is valid for non-Gaussian longitudinal responses since we use a generalized linear mixed model for the longitudinal process. Data from intensive care unit are analysed to illustrate the method.

Keywords: Joint models; Survival analysis; Bayesian P-splines.

1 Introduction

In medical studies, while the primary interest is often to record the time at which a particular event occurs, information on multiple biomarkers is also collected longitudinally throughout the follow-up period. This provides a combination of survival and longitudinal information on each individual under study. Sometimes longitudinal measurements could have a predictive role in the analysis of patient survival. The joint modelling framework is therefore proposed to assess the association between longitudinal measurements and the event risk. In the most common joint model approach, repeated measurements of a biomarker are modelled using a mixed model and the risk event is modelled using a proportional hazard model. Latent

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

variables are used to capture the association between the two outcomes. The book of Rizopoulos (2012) provides an overview of the theory of these joint models. In this paper, we focus on the flexibility of the survival submodel. We propose a joint model allowing the inclusion of non-linear covariate effects in the survival model. Our proposal is illustrated using data from an intensive care unit. We aim to assess the association between patient hypotension and the risk of developing delirium, taking into account the potential non-linear effect of some delirium risk factors. We compare our results with those obtained with the `JMBayes` package. This package allows to compute a joint model under a Bayesian approach but does not enable the inclusion of non-linear covariate effects in the survival submodel.

2 The joint model

2.1 Longitudinal submodel

For every individual ($i = 1, \ldots, n$) we observe the vector of longitudinal response y_i at time points $t_i = [t_{i1}, \ldots, t_{in_i}]^{\mathrm{T}}$. The subject-specific evolution over time of the longitudinal response is model using a generalized linear mixed model (GLMM). In particular, we postulate

$$\eta_i(t) = g\{\mathbb{E}(y_i(t) \mid b_i)\} = {x_i}^{\mathrm{T}}(t)\,\beta + {z_i}^{\mathrm{T}}(t)\,b_i \tag{1}$$

where $x_i(t)$ and $z_i(t)$ are the time-dependent design vectors associated respectively with the vector of fixed effect β and the vector of Gaussian random effects b_i with mean 0 and variance-covariance matrix D. Conditionally on the random effects b_i, the distribution of y_i is assumed to be a member of the exponential family and $g(.)$ denotes the associated canonical link function.

2.2 Survival submodel

To model the event risk, the most standard approaches in the joint modelling framework rely on the proportional hazard models for the conditional hazard:

$$h_i(t|\mathcal{N}_i(t), \omega_i) = h_0(t)\exp\{{\gamma_\omega}^{\mathrm{T}}\omega_i + \alpha\eta_i(t)\}, \quad t > 0 \tag{2}$$

where $h_0(t)$ is the baseline hazard function, $\omega_i = [\omega_{i1}, \ldots, \omega_{ip}]^{\mathrm{T}}$ is the vector of baseline covariates with the corresponding vector of coefficients γ_ω. Parameter α measures the strength of the association between the linear predictor at time t, $\eta_i(t)$, and the event risk at the same time. $\mathcal{N}_i(t) = \{\eta_i(s), 0 \leq s < t\}$ denotes the history of the true unobserved longitudinal process up to time t. To specify $h_0(t)$, Rizopoulos (2012) proposes a B-spline approach that allows flexibility and expresses the logarithm of the baseline hazard function as follows

$$\log\{h_0(t)\} = \sum_{u=1}^{U} \gamma_{h_0,u} b_u(t)$$

where γ_{h_0} is the B-spline coefficients associated to the uth cubic basis function $b_u(t)$.

2.3 Flexible survival submodel

We propose an extended version of (2) that allows for possible non-linear effects of some survival covariates. Specifically, we have

$$h_i(t|\mathcal{N}_i(t), \omega_i, v_i) = h_0(t) \exp\{{\gamma_\omega}^{\mathrm{T}} \omega_i + \sum_{j=1}^{q} f_j(v_{ij}) + \alpha \eta_i(t)\}, \quad t > 0 \quad (3)$$

where $v_i = [v_{i1}, \dots v_{iq}]^{\mathrm{T}}$ is a vector of continuous covariates whose effect is potentially non-linear. The additive terms f_j are estimated using a B-spline approach and are expressed as

$$f_j(v_{ij}) = \sum_{k=1}^{K} \gamma_{v,jk} b_{jk}(v_{ij})$$

where $\gamma_{v,jk}$ represents the B-spline coefficient associated to the cubic B-spline basis function $b_{jk}(.)$ at the respective covariate value v_{ij}. As suggested by Eilers and Marx (1996), we choose a large number of equally spaced knots and we counterbalance the flexibility of the fit by adding a roughness penalty based on differences of adjacent B-spline coefficients.

3 Bayesian Estimation

We estimate the parameters of the joint model (1) and (3) using MCMC algorithms. Assuming independence between the longitudinal and survival processes conditionally on the random effects b_i, the likelihood contribution for the ith subject is given by

$$p(T_i, \delta_i, y_i; \theta) = \int p(T_i, \delta_i | b_i; \theta_s, \beta) \big[\prod_{j=1}^{n_i} p\big\{ y_i(t_{ij}) | \, b_i; \theta_y \big\} \big] p(b_i; \theta_b) \, db_i$$

A Gauss-Kronrod quadrature is used to approximated the integral involded in $p(T_i, \delta_i | b_i; \theta_s, \beta)$.

The roughness penalty introduced in Section 2.3 is translated into a multivariate prior for the B-spline paramaters γ_v as

$$p(\gamma_{h_v} | \tau_v) \propto \tau_v^{\rho(K)/2} \exp\big(-\frac{\tau_v}{2} \gamma_v^T \, K \, \gamma_v \big) \quad \text{and} \quad \tau_v \sim \text{Gamma}(a_1, b_1)$$

where τ_v are smoothing parameter and $K = \Delta_r^{\mathrm{T}} \Delta_r$ is the penalty matrix of rank $\rho(K)$, where Δ_r denotes the rth difference penalty matrix. We consider the same specification for γ_{h_0} with the respective smoothing parameter τ_{h_0}. Univariate diffuse normal priors are used for γ_w and α.

4 Simulation

We perform a simulation study to evaluate the statistical performance of our approach ('`FlexibleJM`'). We extract N=100 samples of n=500 individuals, on which a longitudinal counting response is recorded. Each individual is assumed to be observed for a maximum number of repeated measurements equal to 10. The specification of the joint model considered is defined as follows :

$$\begin{cases} \eta_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + \beta_2 w_{ci} + \beta_3 w_{li} \\ h_i(t) = h_0(t) \exp\left\{\gamma_1 w_{ci} + \gamma_2 w_{hi} + f_1(v_{i1}) + f_2(v_{2i}) + \alpha \eta_i(t)\right\} \end{cases}$$

where $w_{ci}, w_{li}, w_{hi} \overset{i.i.d.}{\sim} \text{Bin}(1, 0.4)$. We generate the observed longitudinal measurements from $y_{ij}|b_{0i} \sim \text{Pois}(\lambda_{ij})$ where $\eta_{ij} = \log(\lambda_{ij})$. The true baseline hazard function is simulated from a Weibull distribution. The smooth additive terms are defined with the functions:

$$f_1(v_{i1}) = -0.07 v_{1i} + 1.6 \quad ; \quad f_2(v_{i2}) = -\sin((v_{i2} + 30)/20) - 4$$

where $v_{i1} \sim \text{U}(16, 35)$, $v_{i2} \sim \text{U}(18, 89)$. Survival times are generated using an algorithm involving numerical integration and root-finding techniques described by Lambert and Crowther (2013). On average, 445 (89%) events occur.

In Figure 1, we show the estimation of the smooth additive terms (gray curves) across all replications. The estimated curves are close to their target (red curves).

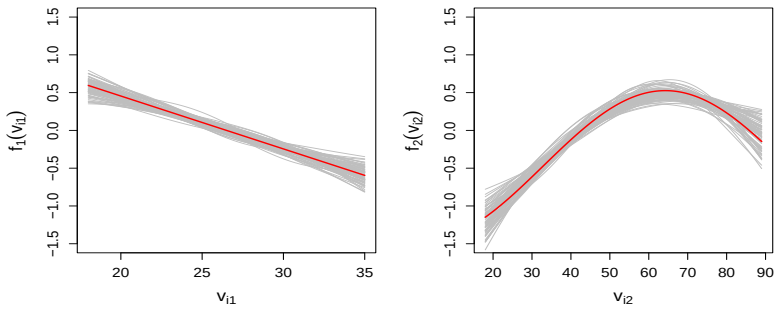


FIGURE 1. True (red) and estimated (gray) smooth functions f_1 and f_2.

In Table 1, we compare the simulation results for the estimated parameters obtained with the `FlexibleJM` and `JMBayes` methods, the latter approach wrongly assuming linear forms for functions f_1 and f_2 (Rizopoulos, 2016). While both approaches show similar performance for the regression parameters $\beta_0, \ldots, \beta_3$ in the longitudinal submodel, the `FlexibleJM` method

outperforms `JMBayes` in estimating the regression parameters γ_1 and γ_2 and, to some extent, the association parameter α, due to the presence of non-linear effects in the survival part.

TABLE 1. Simulation results for N=100 replicates of sample size $n = 500$ with the `FlexibleJM` and `JMBayes` methods : bias and root mean square error (RMSE) for the estimators of the regression parameters.

		`FlexibleJM`		`JMBayes`	
	True	Bias	RMSE	Bias	RMSE
β_0	1.1	0.010	0.026	0.010	0.025
β_1	-0.1	-0.005	0.006	-0.005	0.006
β_2	-0.6	0.020	0.049	0.020	0.049
β_3	0.3	-0.010	0.038	-0.010	0.039
γ_1	1.3	-0.314	0.353	-1.364	1.364
γ_2	0.4	-0.039	0.104	-0.348	0.384
α	-1.0	0.504	0.529	0.529	0.563

5 Application

Patients admitted to an intensive care unit (ICU) are likely to develop cerebral dysfunction and, in particular, delirium. Brain dysfunction can be caused by a prolonged state of low blood pressure, called hypotension. We use a joint model to investiguate whether there is an association between hypotension and the risk of developing delirium. The measured longitudinal response is the number of hypotensive episodes for a patient in one day. Possible risk factors for the occurence of delirium are body mass index (BMI), age and gender. The data set contains 243 patients. Follow-up times varied between 1 and 22 days and 114 patients (46,9%) were right censored. The `FlexibleJM` and `JMBayes` approches consider the following survival sub-models respectively:

$$\begin{aligned} &\texttt{FlexibleJM}: \quad h_i(t) = h_0(t)\exp\{f_1(\texttt{BMI}) + f_2(\texttt{Age}) + \gamma_1\texttt{Gender} + \alpha\eta_i(t)\} \\ &\texttt{JMBayes}: \quad h_i(t) = h_0(t)\exp\{\gamma_1\texttt{BMI} + \gamma_2\texttt{Age} + \gamma_3\texttt{Gender} + \alpha\eta_i(t)\} \end{aligned}$$

where f_1 and f_2 are two unknown smooth functions. While the traditional analysis assuming linearity and relying on `JMBayes` does not confirm the role of BMI as a risk factor, our approach `FlexibleJM` might suggest a nonlinear effect of BMI on the risk of delirium (conditionally on Age and Gender), see Figure 2, with Age playing an important predictive role.

TABLE 2. Estimation results of the survival submodel parameters : estimates, 95% credible intervals, posterior standard deviations are presented.

	`FlexibleJM`		`JMBayes`	
	Est.	CI 95%	Est.	CI 95%
BMI	(see Figure 2)		-0.002	[-0.034; 0.030]
Age	(see Figure 2)		0.028	[0.013; 0.043]
GenderM	0.542	[0.169; 0.835]	0.497	[0.105; 0.863]
Assoc (α)	0.032	[-0.024; 0.094]	0.038	[-0.020; 0.098]

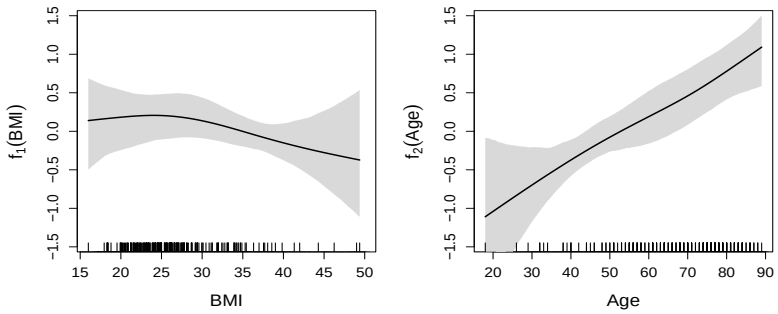


FIGURE 2. Estimation of the non-linear effect of BMI and Age on the log hazard function. The gray surface corresponds to a 95% pointwise credible envelope.

Acknowledgments: The authors acknowledge the support of the ARC project IMAL (grant 20/25-107) financed by the Wallonia-Brussels Federation and granted by the Académie universitaire Louvain.

References

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

Lambert P.C. Crowther M.J. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, **32**(23), 4118 – 4134.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R.* Boca Raton : Chapman & Hall.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC.*Journal of Statistical Software*, **72**(7), 1 – 45.

Smoothing on Networks with P-splines

Paul H.C. Eilers[1]

[1] Erasmus University Medical Centre, The Netherlands

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

Abstract: Smoothing with P-splines on networks with linear edges is a challenging and interesting task, because of equality conditions on fitted values (and possibly derivatives) at nodes where edges join. I propose a model with explicit constraints and Lagrange multipliers that gives complete freedom in the choice of the degree of the splines, the number of knots and order of the penalty.

Keywords: Constraints, Lagrange multipliers, intensity estimation.

1 Introduction

Observations on networks generate interesting statistical challenges. One of them is estimation of the intensities of events along edges of the network. Example are traffic accidents on roads, occupation of parking bays and spines on the dendritic network of a neuron. The network can be simplified as a collection of linear edges, connecting at the nodes. Points on the edges locate the events.
It is of interest to model the intensity of the events. A simple approach is to consider each edge in isolation and apply a smoothing algorithm. That would neglect the connections between the edges, leading to jumps in intensity when moving from one edge to a connected neighbor.
Schneble and Kauermann (2021) proposed to use P-splines. They combine linear splines and a first order penalty with a special construction to join edges, by sharing knots at the nodes. Their algorithm is implemented in the R package `geonet` (Schneble 2021). This approach is hard to extend to higher degrees of the splines. Also it does not allow constraints on derivatives, at the nodes or elsewhere. I propose an alternative algorithm, allowing free choice of knot positions, spline degree and order of the penalty. The key point is tht it puts constraints on the fitted curves (and their derivatives, if needed), using Lagrange multipliers.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Theory

Consider a very simple network, with three nodes and two edges. Edge 1 connects nodes 1 and 2, while edge 2 connects nodes 2 and 3. The edges are straight lines and the relative position on an edge, generally indicated by x, has a value between 0 and 1. For edge 1, $x = 0$ coincides with node 1 and for edge 2, $x = 0$ coincides with node 2. We have two data sets, one the pair of vectors x_1 and y_1 of length m_1 and the other the pair of vectors x_2 and y_2, of length m_2. The x vectors indicates positions on the edges. The y vectors contain arbitrary observed values.
We wish to compute two smooth curves, one based on x_1 and y_1, the other on x_2 and y_2. We use P-splines, combining a B-spline basis with a discrete difference penalty (Eilers and Marx, 1996). The degree of the B-splines and the order of the penalty can be chosen freely. We want the fitted curves to connect without a jump at node 2. If we indicate the first curve by $z_1(x)$ and the second one by $z_2(x)$, then the condition is that $z_1(1) = z_2(0)$.
Let B_1 be a B-spline basis, based on x_1, and let $P_1 = \lambda_1 D_1^{\mathrm{T}} D_1$ be the penalty matrix. To fit P-splines to only the first data set, we solve

$$(B_1^{\mathrm{T}} B_1 + P_1)\hat{a}_1 = B_1^{\mathrm{T}} y_1 \tag{1}$$

for the coefficient vector $\hat{a}_1$. A fitted curve on an arbitrary grid $\breve{x}$ can then be computed as $z_1 = \breve{B}_1 \hat{a}_1$, where $\breve{B}_1$ is a B-spline basis on that fine grid. The choice of knots is free, as long as the domain of x_1 is covered by the B-spline basis. In an analogous way, we can fit the second data set separately. If we form

$$B = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}, \tag{2}$$

we can fit both sets at the same time, by solving $(B'B + P)a = B'y$, with $a = [a_1^{\mathrm{T}} \,|\, a_2^{\mathrm{T}}]^{\mathrm{T}}$, $y = [y_1^{\mathrm{T}} \,|\, y_2^{\mathrm{T}}]^{\mathrm{T}}$, and P a block-diagonal matrix, with P_1 and P_2 on the diagonal. This has no advantages when there are no constraints, but it is a convenient vehicle for adding them.
Currie (2013) showed that fitting P-splines with constraints $Ca = 0$ leads to the system of equations

$$\begin{bmatrix} B^{\mathrm{T}} B + P & C^{\mathrm{T}} \\ C & 0 \end{bmatrix} \begin{bmatrix} a \\ \kappa \end{bmatrix} = \begin{bmatrix} B^{\mathrm{T}} y \\ 0 \end{bmatrix}. \tag{3}$$

Here κ is a vector of Lagrange multipliers. Suppose we evaluate the first basis at a chosen point x to get the row vector b_1 and we evaluate the second basis at the same x to get the row vector b_2, then $C = [-b_1 \,|\, b_2]$ forces a fit with $b_1 a_1 = b_2 a_2$. Note that the degree of the B-splines and the number of knots play no explicit role. The constraint works directly on the fitted curves, and indirectly on their coefficients. At some nodes three (or more) edges may join, which all should have the same value of their smooth

curves. This can be arranged with two (or more) lines in C that constrain different pairs.
The derivative of a B-spline fit can be expressed as Fa, with $F = \bar{B}D_1/h$, with $\bar{B}$ the B-spline basis matrix of degree one lower than that of B, D_1 the matrix that forms first differences, and h the distance between knots. That allows us to put a constraint on derivatives at arbitrary locations with $[-f_1 \,|\, f_2]$ as a row of C. Multiple simultaneous constraints, on values and derivatives, lead to multiple rows of C.
To estimate an intensity curve on an isolated edge, we count events in bins, to form a histograms y and model $\eta = \log \mu = Ba$, with $\mu = E(y)$ the expected values of the counts (Eilers and Marx, 1996). We repeatedly solve $(B^{\mathrm{T}}\tilde{M}B + P)a = B^{\mathrm{T}}(y - \tilde{\mu} + \tilde{M}\tilde{\eta})$, where a tilde indicates a current approximation and $M = \mathrm{diag}(\mu)$.
With two edges we have histograms y_1 and y_2 with expected values μ_1 and μ_2, We join y_1 and y_2 to form y and we form M as the block-diagonal matrix with $M_1 = \mathrm{diag}(\mu_1)$ and $M_2 = \mathrm{diag}(\mu_2)$ on the diagonal and repeatedly solve

$$\begin{bmatrix} B^{\mathrm{T}}\tilde{M}B + P & C^{\mathrm{T}} \\ C & 0 \end{bmatrix} \begin{bmatrix} a \\ \kappa \end{bmatrix} = \begin{bmatrix} B^{\mathrm{T}}(y - \tilde{\mu} + \tilde{M}\tilde{\eta}) \\ 0 \end{bmatrix}. \tag{4}$$

Here a tilde indicates the current approximation.
It is convenient to put constraints on η (and its derivative), not on μ, to have a similar construction as in the case of linear smoothing.
To achieve the same amount of smoothing on edges of different lengths, the penalties should have the form $\lambda||Da_k||/n_k$, where a_k is the coefficient vector, with length n_k of segment k.
Automatic smoothing is attractive. With $G = (B^{\mathrm{T}}\hat{M}B + P)^{-1}B^{\mathrm{T}}\hat{M}B$, we have that the effective model dimension is $\mathrm{ED} = \mathrm{trace}(G)$, which can be used for the computation of AIC. Instead of searching for the minimum of AIC for a series of values of λ, mixed model ideas can be used; the Harville-Fellner-Schall (HFS) algorithm is attractive (Eilers and Marx, 2021). Assuming that the conditional distributions of the histogram counts are Poisson with expectations μ, it can be shown that $||D\alpha||^2/\mathrm{ED} = 1/\lambda$. This gives an easy way to update λ in each iteration.

3 A small example

The R package `geonet` contains data from a part of the highway network in Montgomery county (Maryland, USA). Figure 1 shows a very small subset of these data. There are 5 edges and 6 nodes. On each edge accidents are located as dots; the goal is to estimate accident intensities.
One way to analyze these data is to interpret them as five coupled segments. At node 2 the segments 1 and 4 connect, and at node 1 the four edges 1, 2, 3, and 5. All other nodes are end points of edges. Alternatively, we can

interpret edges 4-1-5 and 3-2 as single linear segments and only have one condition on the fitted values in node 1 where these segments connect.
The value of λ was determined by the HFS algorithm.

4 Discussion

By implementing constraints on fitted curves, instead of on the spline knots, we get a lot of freedom in the specification of the P-splines we use. We can also constrain derivatives, although the possibility to smooth connected paths consisting of multiple edges reduces the need for this option.
Smoothing on paths can simplify the amount of work. The documentation of the package `geonet` (Schneble 2021) describes a part of the road network in Chicago. In typical American style much of it is a rectangular grid, which can be modeled as a a combination of crossing vertical and horizontal paths. However, in more general cases, it is often not automatically obvious how to construct paths. In Figure 1 I connected the edged 4-1-5 and 3-2, but 4-1-2, connected to the two individual edges 3 and 5 would also make sense. In a network with different types of roads, one might prefer to form paths with road sections of equal type.
Intensity values have no sense of direction, but derivatives have. Intuitively it feels that if edges join at a sharp angle, equality of derivatives is less convincing.

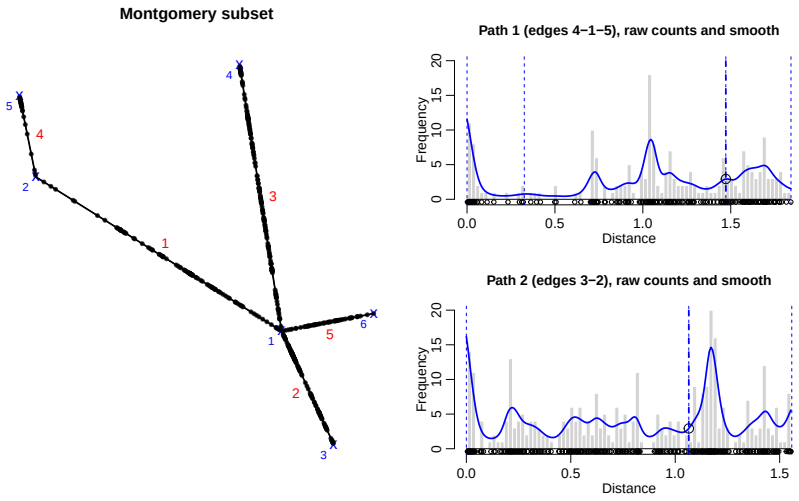


FIGURE 1. Road accidents in Montgomery county (selection). Left panel: five segments, with the locations of accidents (dots). Right panel: two paths (edges 4-1-5 and edges 3-2) with histograms and fitted smooth curves. The vertical broken lines indicate the node where the edges on a path connect. The thicker vertical lines indicate where the two paths connect (at node 1). The circles there indicate the fitted values, constrained to be equal.

References

Currie, I.D. (2013). Smoothing Constrained Generalized Linear Models with an Application to the Lee-Carter Model. *Statistical Modelling*, **13**, 69 – 93.

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

Eilers, P.H.C. and Marx, B.D. (2021). *Practical Smoothing. The Joys of P-splines*. Cambridge University Press.

Schneble, M. (2021) geonet: Intensity Estimation on Geometric Networks with Penalized Splines. R package version 0.6.0.

`https://CRAN.R-project.org/package=geonet`.

Schneble, M. and Kauermann G. (2021) Intensity Estimation on Geometric Networks with Penalized Splines. arXiv:2002.10270

Detection of relevant covariates involved in casual rentals in the Capital bikeshare program of Washington, D.C., by means of new nonparametric specification tests for additive concurrent models formulation

Laura Freijeiro-González[1], Manuel Febrero-Bande[1], Wenceslao González-Manteiga[1]

[1] Centre for Mathematical Research and Technology Transfer of Galicia (CITMAga). Department of Statistics, Mathematical Analysis and Optimization. Universidade de Santiago de Compostela, Santiago de Compostela, Spain.

E-mail for correspondence: laura.freijeiro.gonzalez@usc.es

Abstract: Motivated by the detection of relevant covariates to explain casual rentals in the Washington, D.C., bike sharing system, new nonparametric specification tests for additive concurrent models are proposed. These are based on the martingale difference divergence coefficient and exhibit the novelty that neither it is necessary to estimate smoothing/penalizing parameters or model structure compared to existing literature.

Keywords: Specification tests; covariates selection; functional concurrent model; MDD.

1 Problem motivation

The capital of the United States, Washington, D.C., is one of the cities most visited in the U.S. As an example, the number of visitors per year exceeds 20 million since 2014. In consequence, the Capital bikeshare system has interest in predicting casual bike rentals, defined as rentals to cyclists without membership in the program, to properly face the demand.
Making use of the Washington, D.C., bike sharing dataset of Fanaee-T and Gama (2014), collected from 1 January 2011 to 31 December 2012, we try to determine which meteorological covariates influence casual bike rentals

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

on Saturdays. For this purpose, four covariates are considered: daily temperature (temp), feeling temperature (atemp), humidity (hum) and wind speed (wind). Complete data is displayed in Figure 1.

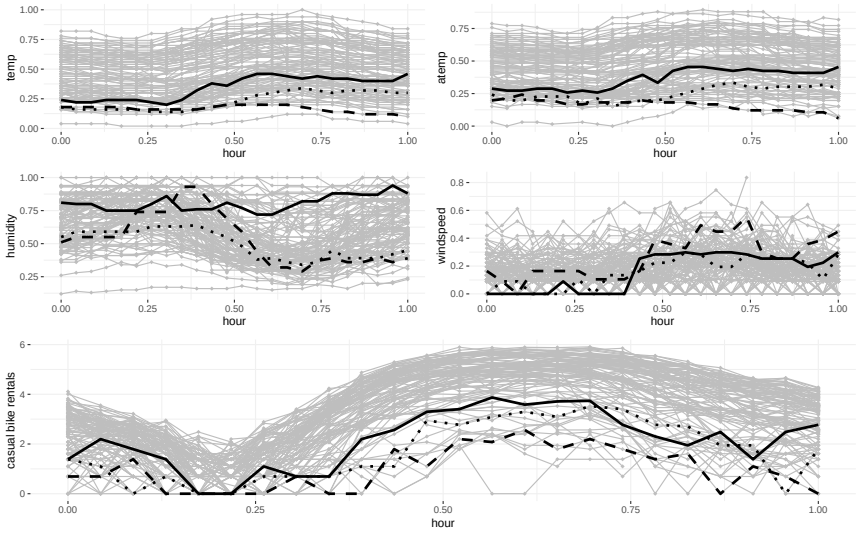


FIGURE 1. Daily temperature (temp), feeling temperature (atemp), humidity (hum), wind speed (wind) and casual bike rentals on an hourly basis in Washington, D.C., on Saturdays.

How it is expected that meteorological variables have real-time dynamic effects on the number of bike rentals, we can relate them by means of a quite flexible additive concurrent model. This is given by

$$\begin{aligned} Y(t) =& F_1(t, X_{\text{temp}}(t)) + F_2(t, X_{\text{atemp}}(t)) \\ &+ F_3(t, X_{\text{hum}}(t)) + F_4(t, X_{\text{wind}}(t)) + \varepsilon(t) \end{aligned} \tag{1}$$

where $F_j(\cdot)$ are the additive effects for $j = 1, 2, 3, 4$ covariates and $\varepsilon(\cdot)$ is the unknown model error.
Thus, as preliminary steps, we need to determine 1) if the considered meteorological covariates support useful information to model the bikes rentals and, if so, 2) if all of them are relevant or some can be excluded from the model. This last would result in a problem dimension reduction.

2 Novel MDD global specification tests for additive concurrent models

An additive concurrent model is a regression model where the response $Y \in \mathbb{R}$ and $p \geq 1$ covariates $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ are all functions of

the same argument $t \in \mathcal{D}_t$, and the influence is concurrent, simultaneous or point-wise in the sense that X is assumed to only influence $Y(t)$ through its value $X(t) = (X_1(t), \ldots, X_p(t)) \in \mathbb{R}^p$ at time t by means of the relation

$$Y(t) = F_1(t, X_1(t)) + \cdots + F_p(t, X_p(t)) + \varepsilon(t),$$

where $F_j(\cdot)$ are unknown functions collecting the $\mathbb{E}\left[Y(t)|_{X_j(t)}\right]$ information and $\varepsilon(t)$ is the error of the model, which is assumed to have mean zero, independent of X and with covariance function $\Omega(s,t) = \operatorname{cov}(\varepsilon(s), \varepsilon(t))$. Thus, to assure the veracity of the model structure, nullity of main effects have to be rejected. For this purpose, taking $D \subset \{1, \ldots, p\}$, we are interested in elucidating whether the selected covariates are relevant in the $Y(t)$ explanation. Notice that we can consider the particular cases of $D = \{1, \ldots, p\}$, which translates in testing if all p covariates are relevant, or $D = \{j\}$ for some $j = 1, \ldots, p$, allowing us to implement covariates screening. As a result, a specification test can be performed by means of

$$\begin{aligned} H_0 &: \mathbb{E}\left[Y(t)|_{X_j(t)}\right] = \mathbb{E}\left[Y(t)\right] \text{ a.s. } \forall t \in \mathcal{D}_t \setminus \mathcal{N}_t \text{ and every } j \in D \\ H_a &: \mathbb{P}\left(\mathbb{E}\left[Y(t)|_{X_j(t)}\right] \neq \mathbb{E}\left[Y(t)\right]\right) > 0 \text{ for some } t \in \mathcal{V}_t \text{ and } j \in D \end{aligned} \tag{2}$$

where $\mathcal{D}_t \setminus \mathcal{N}_t$ is the domain of t minus a null set $\mathcal{N}_t \subset \mathcal{D}_t$ and $\mathcal{V}_t \subset \mathcal{D}_t$ is a positive measure set.
Making use of the innovative Martingale Divergence Difference (MDD) coefficient of Shao and Zhang (2014), the specification test displayed in (2) can be rewritten in terms of the MDD as

$$\begin{aligned} H_0 &: \int_{\mathcal{D}_t} MDD^2(Y(t)|_{X_j(t)})dt = 0 \text{ a.s. for every } j \in D \\ H_a &: \mathbb{P}\left(\int_{\mathcal{D}_t} MDD^2(Y(t)|_{X_j(t)})dt \neq 0\right) > 0 \text{ for some } j \in D \end{aligned} \tag{3}$$

In order to implement the previous test, an integrated statistic is consider estimating $MDD^2(Y(t)|_{X_j(t)})$ for $j = 1, \ldots, p$. This is based on

$$T_D = \sqrt{\binom{n}{2}} \frac{\sum_{j \in D} \int_{\mathcal{D}_t} MDD_n^2(Y(t)|_{X_j(t)})dt}{\widehat{\widetilde{\mathcal{S}}}_D}$$

where $\widehat{\widetilde{\mathcal{S}}}_D$ is a suitable variance estimator of $\sum_{j \in D} \int_{\mathcal{D}_t} MDD_n^2(Y(t)|_{X_j(t)})dt$. We refer the reader to Zhang et al. (2018) for more details.
Verifying some assumptions, the asymptotic normality of the T_D statistic is guaranteed under the null hypothesis and a consistent wild bootstrap procedure is proposed to estimate its p-value in practice.

It is relevant to highlight that other approaches have been proposed to implement specification tests for the concurrent model in literature. Examples of these are the works of Wang et al. (2017) and Ghosal and Maity (2022a) in the linear concurrent model formulation and the one of Kim et al. (2018), which extends Ghosal and Maity (2022a) ideas to additive effects. Nevertheless, all of them depend on some smoothing parameters: selection of a proper bandwidth value or the number of considered basis terms for effects representations, jointly with the error model structure estimation in Ghosal and Maity (2022a) and Kim et al. (2018). In contrast, our procedure has the novelty that no smoothing parameters as well as no structure assumption are required, resulting in a nonparametric approach easier to implement in practice.

3 Conclusions

Assuming the model formulation displayed in (1) we carry out global and partial versions of test (3) to determine which meteorological covariates are relevant. The global test obtains a p-value = 0, which rejects the null hypothesis of independence for usual significance levels. Then, considered covariates have an effect in causal rentals formulation. Regarding to partial tests we obtain p-values of 0, 0, 0.007 and 0.001 for temperature (temp), feels-like temperature (atemp), relative humidity (humidity) and wind speed (windspeed), respectively. Thus, we can claim that all of them have an impact on the number of casual rentals at significance levels as the 1%. This last agrees with other studies as the one of Ghosal and Maity (2022), where different covariates are selected by their considered penalties. In an overview of their results, each covariate is selected at least two times over the five considered procedures. As a result, all of them seem to play a relevant role separately. The difference between both approaches may be because our test is able to detect causality, whereas the Ghosal and Maity (2022b) study focuses on covariates selection in terms of minimizing model residuals in the estimation process. Then, some relevant covariates may be excluded from the model because collinearity effects or due to the fact that their inclusion does not contribute too much in residuals reduction. As a result, our tests provide a different covariates screening point of view.

Acknowledgments: We want to thank the economical support to: Consellería de Cultura, Educación e Ordenación Universitaria, Consellería de Economía, Emprego e Industria of the Xunta de Galicia (project ED481A-2018/264), Project PID2020-116587GB-I00 funded by MCIN/AEI/10.13039 /501100011033 along with "ERDF A way of making Europe" and the Competitive Reference Groups 2021–2024 (ED431C 2021/24) from the Xunta de Galicia through the ERDF. Besides, we acknowledge to the Centro de Supercomputación de Galicia (CESGA) for

computational resources.

References

Fanaee-T, H. and Gama, J. (2014). (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, **2**, 113 – 127.

Ghosal, R. and Maity, A. (2022a). A score based test for functional linear concurrent regression. *Econometrics and Statistics*, **21**, 114 – 130.

Ghosal, R. and Maity, A. (2022b). Variable selection in nonparametric functional concurrent regression. *Canadian Journal of Statistics*, **50(1)**, 142 – 161.

Kim, J. S., Staicu, A.-M., Maity, A., Carroll, R. J., and Ruppert, D. (2018). Additive function-on-function regression. *Journal of Computational and Graphical Statistics*, **27**, 234 – 244.

Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, **109(507)**, 1302 – 1318.

Wang, H., Zhong, P.-S., Cui, Y., and Li, Y. (2017). Unified empirical likelihood ratio tests for functional concurrent linear models and the phase transition from sparse to dense functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**.

Zhang, X., Yao, S., and Shao, X. (2018). Conditional mean and quantile dependence testing in high dimension. *The Annals of Statistics*, **46(1)**, 219 – 246.

A comparison of unconstrained parameterisations for additive mean and covariance matrix modelling

Vincenzo Gioia[1], Matteo Fasiolo[2], Ruggero Bellio[1]

[1] Department of Economics and Statistics, University of Udine, Italy
[2] School of Mathematics, University of Bristol, UK

E-mail for correspondence: gioiavincenzo.25790@gmail.com

Abstract: Covariance models for multivariate normal data must ensure the positive definiteness of the covariance matrix. Computational scalability for handling large samples is further desirable. We propose flexible covariance modelling by reparameterising the covariance matrix according to two different approaches, namely the matrix logarithm and the modified Cholesky decomposition. The performances of the proposed additive covariance models (ACM) are compared on an electricity load modelling application.

Keywords: Matrix logarithm; Modified Cholesky decomposition; Multivariate electricity load forecasting; Penalised likelihood; Smoothing covariance modelling.

1 Introduction

Covariance models for multivariate normal data assume a specific form of the $d \times d$ covariance matrix $\mathbf{\Sigma}$ of the response vector of interest, considering a functional dependence on some available covariates. The positive definiteness of $\mathbf{\Sigma}$ must be ensured, and it has been tackled in various ways in the literature. Two important methods are the matrix logarithm (logM) of $\mathbf{\Sigma}$ (Chiu et al., 1996) and the modified Cholesky decomposition (MCD) of the precision matrix $\mathbf{\Sigma}^{-1}$(Pourahmadi, 1999). Here we propose to extend these two approaches, allowing for the elements of the covariance matrix to vary smoothly with the covariates. Computational performances and model comparisons are illustrated on electricity load data.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Additive mean and covariance matrix models

Consider independent d-dimensional response vectors $\mathbf{y}_i = (y_i^1, \ldots, y_i^d)^\mathrm{T}$, $i = 1, \ldots, n$, normally distributed with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Let $\boldsymbol{\eta}_i = (\eta_i^1, \ldots, \eta_i^q)^\mathrm{T}$, $q = d + d(d+1)/2$, be the i-th linear predictor vector, used to model both the mean vector and the covariance matrix. The mean vector $\boldsymbol{\mu}_i$ has elements $\mu_i^k = \eta_i^k, k = 1, \ldots, d$, and the unconstrained elements of a suitable parameterisation of $\boldsymbol{\Sigma}_i$, or of $\boldsymbol{\Sigma}_i^{-1}$, are $\eta_i^k, k = d+1, \ldots, q$. That is, the non-redundant elements of $\log \boldsymbol{\Sigma}_i$, under the logM approach, and the diagonal elements of $\log \mathbf{D}_i^2$ and the off-diagonal nonzero elements of $\mathbf{T}_i$, under the MCD approach, are modelled via the linear predictors. In the latter approach

$$\boldsymbol{\Sigma}_i^{-1} = \mathbf{T}_i^\top \mathbf{D}_i^{-2} \mathbf{T}_i \,,$$

where $\mathbf{T}_i = \mathbf{D}_i \mathbf{C}_i^{-1}$, with $\mathbf{C}_i$ from the Cholesky decomposition $\boldsymbol{\Sigma}_i = \mathbf{C}_i \mathbf{C}_i^\top$, and $\mathbf{D}_i$ is the diagonal matrix containing the diagonal elements of $\mathbf{C}_i$.
The log-likelihood for the i-th observation, up to an additive constant, is

$$\ell(\boldsymbol{\eta}_i) = -\frac{1}{2} \log |\boldsymbol{\Sigma}_i| - \frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) \,.$$

The k-th element of the i-th linear predictor is

$$\eta_i^k = \mathbf{Z}_i^k \boldsymbol{\gamma}^k + \sum_j g_j^k(x_i^j) \,, \tag{1}$$

with $\mathbf{Z}_i^k$ the i-th row of a parametric model matrix, $\boldsymbol{\gamma}^k$ a regression parameter vector, and g_j^k a smooth function of covariate x^j. Each g_j^k can be represented using a reduced rank spline basis, with an associated quadratic penalty. Hence, each linear predictor can be written as $\eta_i^k = \mathbf{X}_i^k \boldsymbol{\beta}^k$, with $\mathbf{X}_i^k = (X_{i1}^k, \ldots, X_{ip_k}^k)$ and $\boldsymbol{\beta}^k = (\beta_1^k, \ldots, \beta_{p_k}^k)^\mathrm{T}$; here X_{i1}^k is equal to one.
The estimation of $\boldsymbol{\beta} = ({\boldsymbol{\beta}^1}^\mathrm{T}, \ldots, {\boldsymbol{\beta}^q}^\mathrm{T})^\mathrm{T}$ is carried out by Newton optimisation of the penalised log-likelihood, that is

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{\beta}) - \frac{1}{2} \sum_{j=1}^{M} \lambda_j \boldsymbol{\beta}^\top \mathbf{S}^j \boldsymbol{\beta} \right\}, \tag{2}$$

for fixed M-dimensional smoothing parameter vector $\boldsymbol{\lambda}$, with $\lambda_j \in \mathbb{R}^+$ and where the $\mathbf{S}^j$ are matrix of known coefficients. The generalised Fellner-Schall method (Wood and Fasiolo, 2017) is used to select $\boldsymbol{\lambda}$ by maximising a Laplace approximation to the marginal likelihood. See also Wood (2017). An approximate Bayesian framework can be used for inference. Indeed, note that $\hat{\boldsymbol{\beta}}$ is the posterior mode for $\boldsymbol{\beta}$, and that the generalized ridge penalty in (2) corresponds to the improper prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{S}_{\boldsymbol{\lambda}}^-)$, with $\mathbf{S}_{\boldsymbol{\lambda}}^-$ being a suitable pseudo-inverse of $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_j \lambda_j \mathbf{S}^j$. For fixed $\boldsymbol{\lambda}$ and in the large

sample limit we have that $\boldsymbol{\beta}|\mathbf{y} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, (\hat{\boldsymbol{\mathcal{H}}} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1})$, where $\hat{\boldsymbol{\mathcal{H}}}$ is the observed information matrix of the log-likelihood at $\hat{\boldsymbol{\beta}}$. With this asymptotic approximation we can obtain (possibly by simulation) Bayesian credible intervals for any function of $\boldsymbol{\beta}$.

3 Multivariate electricity load modelling

Data from the electricity load forecasting track of the GEFCom2014 challenge (Hong et al., 2016) are considered for illustrative purposes. The hourly load (in MW) for an undisclosed US utility and the temperature data span the period from 2005/01/02 to 2011/11/27. Hourly loads from 5 p.m. to 10 p.m on a daily basis are used as response variables. Hence, the number of variables is $d = 6$, while the number of observations is $n = 2513$. The covariates include the day of the year (x^1), the day of the week (x^2), the temperature (x^3), and the hourly loads of the previous day (x^4). The mean model is specified as in (1), with two smooth terms for x^1 and x^3, where g_1^k and g_3^k are smooth functions, built using ten basis functions. Parametric linear effects are used for x^2 and x^4.
The covariance models are specified by considering a fixed covariance structure (`Fixed`) and an additive covariance model (`ACM`). In the latter, all the elements of the logM and MCD parameterisations are modelled as in (1), with a smooth term for x^1 and a parametric component for x^2; here the total number of linear predictors is $q = 27$. Under the MCD approach, an order-1 ante-dependence model (`ACM AD-1`) is also considered (see Pourahmadi, 1999), achieving a reduction in the number of parameters.
The times (in minutes) taken for one fit, by varying d and according to the `ACM` specification, are reported in Table 1. It is clear that MCD outperforms logM, which is due to the analytical expression of the likelihood quantities, including the Hessian of the log-likelihood function with respect to the linear predictors. Indeed, such Hessian involves multiple summations in the logM approach, due to the computation of the matrix exponential derivatives. The model fit was carried out using a laptop computer with an Intel Core i5-10210U processor and 16 GB of RAM.

TABLE 1. Estimation times considering the `ACM` specification (in minutes).

d	dim($\boldsymbol{\beta}$)	MCD	logM
2	100	0.06	0.14
3	174	0.26	1.38
4	264	0.41	6.77
5	370	1.39	32.84
6	492	2.98	81.08

The models are evaluated by considering the energy and logarithmic scores and the forecasting metrics are compared by using the skill score (see Gneiting and Raftery, 2007). A block bootstrap resampling of skill scores is used to determine whether the forecasting metrics differ significantly at 0.05 level. A cross-validation procedure, known as evaluation on a rolling forecasting origin (see Hyndman and Athanasopoulos, 2021), is implemented. In this respect, the origin of forecast rolls forward in time of 1 week starting from 2011/01/01. Forecasting metrics and the significance of the difference between scores are reported in Table 2. The `ACM` specification is preferable to the fixed covariance structure model and the `ACM - AD1` form is comparable to the `ACM` one.

TABLE 2. Results for electricity load forecasting. Underline indicates that the skill score relative to the `ACM` one is not significantly different from zero.

	MCD		logM	
Model	Energy	Logarithmic	Energy	Logarithmic
`Fixed`	23.49	16.03	23.49	16.03
`ACM AD - 1`	22.97	15.73	–	–
`ACM`	22.89	15.64	22.84	15.60

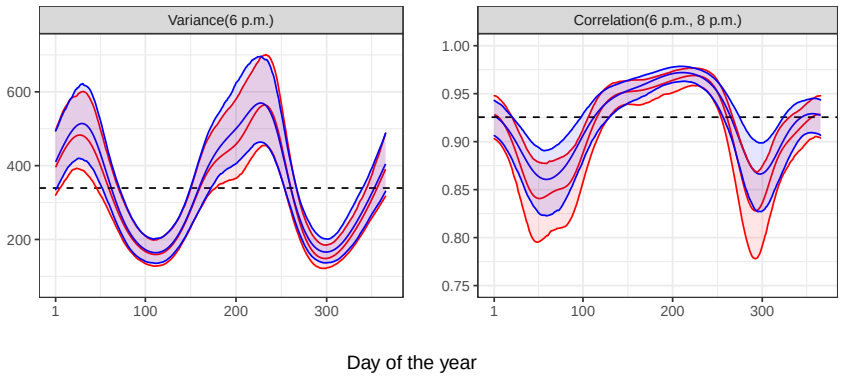


FIGURE 1. Results for the `ACM` specification using MCD (red) and logM (blue). Fitted variances (in MW^2) and correlations, with 95% credible intervals, by day of the year and for Mondays. The black dashed lines correspond to the `Fixed` model fit.

A comparison between the unconstrained MCD and logM parameterisations on the fitted variances (in MW^2) and correlations is shown in Figure 1. The slight differences between the MCD and logM fits are due to the

different parameterisations used. It seems appropriate to consider a varying effect of the day of the year on the variances and the correlations, as the deviations from the fixed covariance model fit are apparent.
Figure 2 reports the fitted variances and correlations for the `ACM` specification using the MCD approach and by considering three different hours (6 p.m., 7 p.m., and 8 p.m.). The variances show seasonal peaks due to the heating and cooling effects related to the winter and summer periods, respectively. In addition, it is apparent the drop in the variances of the summer peak as the time approaches the night hours. The troughs of the correlations correspond to the late winter and autumn periods, characterized by mild temperatures. The correlations decrease as the lag time increase, especially during the correlation troughs. The patterns in Figure 2 are shifted vertically depending on the day of the week.

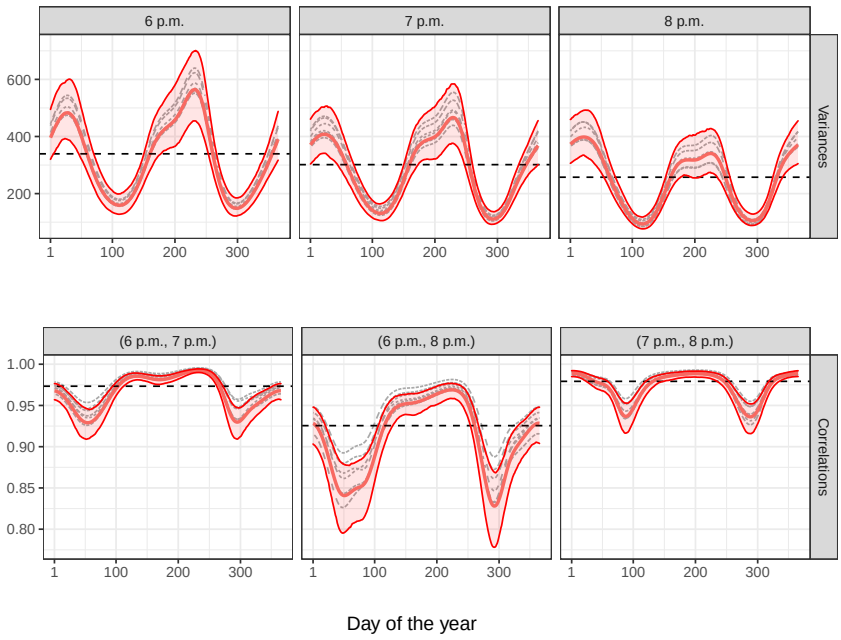


FIGURE 2. Results for the `ACM` specification using MCD. Fitted variances (in MW^2) and correlations by day of the year (red lines for Monday, with 95% credible intervals; grey dashed lines for the other days of the week). The black dashed lines correspond to the `Fixed` model fit.

4 Conclusion and ongoing work

Additive covariance models represent a useful class of statistical models, rarely used in practice due to the intrinsic challenges involved in their adoption. The results illustrated here suggest that scalable implementation of the MCD approach might be more promising than that of the logM approach. For a more widespread usage of such tools in applications, some progress in public available software is needed. To this end, a `R` package implementing the two methods is currently under development.

References

Chiu, T.Y.M., Leonard, T., and Tsui, K.W. (1996). The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, **91**, 198 – 210.

Gneiting, T., and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359 – 378.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R.J. (2016). Probabilistic energy forecasting: global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, **32**, 896 – 9̇13.

Hyndman, R.J., and Athanasopoulos, G. (2021). *Forecasting: principles and practice,* 3rd ed. Melbourne: OTexts.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, **86**, 677 – 690.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R,* 2nd ed. Boca Raton FL: Chapman and Hall/CRC.

Wood, S.N., and Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*, **73**, 1071 – 1081

Personalised effect of discontinuing treatment in Heart Failure patients through multi-state modelling

Caterina Gregorio,[1 2], Giulia Barbati[2], Francesca Ieva[1 3]

[1] MOX - Modelling and Scientific Computing, Department of Mathematics Politecnico di Milan, Piazza Leonardo Da Vinci 32, Milan 20123, Italy
[2] Biostatistic Unit, Department of Medical Sciences, University of Trieste, via Valerio 4\1,Trieste 34100, Italy
[3] CHDS, Center for Health Data Science, Human Technopole, Viale Rita Levi Montalcini 1, Milan 20157, Italy

E-mail for correspondence: caterina.gregorio@units.it

Abstract: In heart failure, medical decisions with regard to pharmacological therapies are very complex due to the heterogeneity in subjects' profiles. In this work, we propose a flexible parametric multi-state model to represent the health and treatment path of heart failure patients over time. From this model, we are able to obtain an estimate of the effect of discontinuing the treatment according to different subjects' health status over time.

Keywords: Multi-state model; Real-world data; Heart failure

1 Introduction

Treatment of Heart Failure relies on several life-saving pharmacological therapies. Among them, mineralocorticoid receptor antagonists (MRAs) are one of the cornerstone of therapy in heart failure, yet is one that is most often discontinued by cardiologists out of fear of adverse events, e.g. alteration of potassium (Komajda et al. (2016); Maggioni et al. (2013)). However, there is no clear evidence regarding when side effects overcome benefits in terms of the risk of hospitalisation or death. Entangling the relationship between the discontinuation of MRAs and the risk of adverse events is a non-trivial statistical problem: the longitudinal nature of these processes can't be uncoupled with patient's disease progression as well as potassium behaviour over time. Therefore, in light of the above, we propose

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

a multi-state approach to model patients' health and treatment status over time jointly with the time-to-event processes i.e. repeated hospitalisations and death. The aim of this work is to extend the model by Ieva et al.(2017) considering different out-of-hospital states to represent possible combinations of patients' clinical phenotypes, potassium dynamics, and treatment with MRAs over time. From this model, we are able to obtain an estimate of the effect of discontinuing vs. continuing treatment with MRAs for each different health status with respect to the risk of hospitalisation.

2 Dataset

Data was obtained by the interrogation of the administrative regional health database of the Friuli Venezia Giulia Region in the Northern part of Italy, integrated with the Outpatient and Inpatient Clinic E-chart (Cardionet) (Iorio et al., (2019)). The cohort was followed from the date of first purchase of MRAs, until the time of death or the administrative study closure date, i.e. 31/12/20. For this specific study, data concerning demographic, clinical and instrumental variables, repeated blood tests containing the potassium measurements, all drug purchases and all dates of hospitalisation collected during the observation period have been considered.

3 Methods

3.1 Multi-state model structure

States considered in the model consists of out-of-hospital states, state in-hospital and an absorbing state i.e. death. Out-of-hospital states, $\mathcal{O}$, are defined from the clinical phenotype, the treatment status and the potassium status. Two possible treatment status, $\mathcal{T} = \{'On','Off'\}$, have been defined from subjects' drug purchases data. The clinical phenotype have been defined using subjects' clinical characteristics collected during the follow-up. We denote possible phenotypes with $\mathcal{P} = \{p_1, ..., p_{r_p}\}$. Finally, the potassium status has been defined according to whether the biomarker shows a stable or fluctuating behaviour i.e. $\mathcal{K} = \{'Stable','Fluctuating'\}$. The presence of oscillations from longitudinal potassium measurements data can be determined according to the methodology based on wavelet filtering proposed by Gregorio et al. (2022). Thus, $\mathcal{O} = \mathcal{P} \times \mathcal{T} \times \mathcal{K}$ and there are a total of $r_o = 2 \times 2 \times r_p$ out-of hospital states. We denote by $\lambda_{qs}, q, s = 1, ...r$ the set of transition intensities i.e. the instantaneous risk of moving from state q to state s. We use a semi-Markov model so that the intensities depends on the process history only through the time spent in the current state and other patient's characteristics contained in the vector of covariates $\mathbf{x}$.

3.2 Estimation of the IN-hospital transitions intensities

We are interested in studying the effect of discontinuing treatment according to subjects' different health status on the risk of hospitalisation. For this reason, we model the transition intensities from the out-of-hospital states to the in-hospital state i.e. *IN-hospital transitions*:

$$\lambda_i^{qIN}(u) = \lambda_0^{qIN}(u) exp\{\beta^T \mathbf{x}_i\} \tag{1}$$

where $q : 1, ..., r_o$ and u is the time since the entry in the current state. The model is a flexible parametric model (Royston and Parmar (2017)) in which baseline transition intensities are modelled with natural cubic splines with k internal knots.

3.3 Personalised effect of the discontinuation of the treatment

To compare the hazard between the two treatment status, among patients belonging to the same phenotype and potassium status, we can define the ratio between the IN-hospital transitions intensities as follows:

$$HR_{p\times k}^{\mathcal{T}}(u) = \frac{\lambda_0^{(p\times k\times' Off')IN}(u) exp\{\beta^T \mathbf{x}_i\}}{\lambda_0^{(p\times k\times' On')IN}(u) exp\{\beta^T \mathbf{x}_i\}} = \frac{\lambda_0^{(p\times k\times' Off')IN}(u)}{\lambda_0^{(p\times k\times' On')IN}(u)} \tag{2}$$

where $(p \times k \times' Off')$ and $(p \times k \times' On')$ denote the out-of-hospital states with $\mathcal{T} =' Off'$ and $\mathcal{T} =' On'$ respectively.

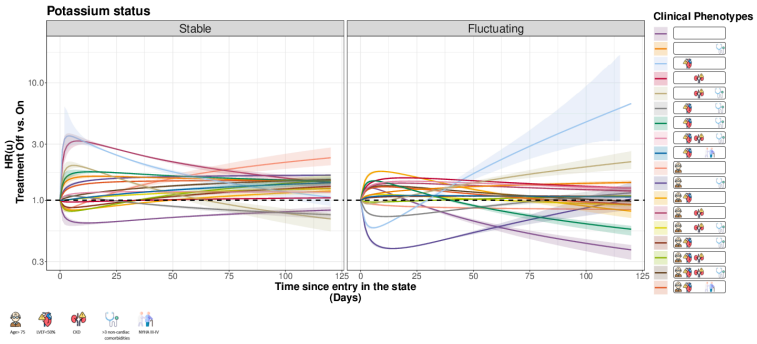


FIGURE 1. Estimated hazard ratio curves ($\widehat{HR}_{p\times k}^{\mathcal{T}}(u)$) for the risk of hospitalization according to clinical phenotypes and potassium status. Colored areas represent 95% CIs.

4 Results

In the analysis, were considered 1695 subjects with a median observation period of 3 years (IQR: 1.5-5). Age, left-ventricular ejection fraction (LVEF), New York Heart Class (N.Y.H.A.), number of non-cardiac comorbidities and Chronic Kidney Disease (CKD) were used to define 18 different clinical phenotypes. The R package `AdhereR` (Dima and Dendiu (2017)) was used to derive the treatment status over time from subjects' drug purchase data. The model for the IN-hospital transition intensities was estimated using the R package `flexsurv` (Jackson (2017)). Using the AIC criteria, 1 was chosen as the best number of internal knots for the baseline intensities of the model and sex was considered as covariate. The estimated hazard ratios of discontinuing the treatment with MRAs are shown in Figure 1.

5 Conclusions

In this work, we propose a novel method to estimate the personalized effect of discontinuing a treatment according to subjects' time-varying information from real-world data. From the results obtained, we observe that the effect of discontinuing the treatment varies greatly according to different clinical phenotypes and potassium status suggesting that new quantitative decision tools may be useful to guide medical decisions in heart failure.

Acknowledgments: This work was supported by VIFOR Pharma. The authors also thank the Trieste Observatory of Cardiovascular Disease, Dr. Andrea Di Lenarda and Dr.Arjuna Scagnetto.

References

Dima, L. A., Dendiu D. (2017). Computation of adherence to medication and visualization of medication histories in R with AdhereR: Towards transparent and reproducible use of electronic healthcare data. *PLos ONE*, **12.4**, e0174426.

Gregorio, C., Barbati, G., Ieva, F. (2022). A wavelet-mixed landmark survival model for the effect of short-term oscillations in longitudinal biomarker's profiles. https://arxiv.org/abs/2204.05870

Ieva, F., Jackson, C. H., Sharples, L. D. (2017). Multi-State modelling of repeated hospitalisation and death in patients with Heart Failure: the use of large administrative databases in clinical epidemiology. *Statistical methods in medical research*, **26.3**, 1350 – 1372.

Iorio, A., Sinagra, G., Di Lenarda, A. (2019). Administrative database, observational research and the Tower of Babel. *International Journal of Cardiology*, **284**, 118 – 119.

Jackson, C. (2017). flexsurv: A Platform for Parametric Survival Modeling in R *Journal of Statistical Software*, **70.8**, 1 – 33.

Komajda, MAnker, S. D., Cowie, M. R., Filippatos, ..., QUALIFY Investigators. (2016). Physicians' adherence to guideline-recommended medications in heart failure with reduced ejection fraction: data from the QUALIFY global survey. *European journal of heart failure*, **18(5)**, 514 – 522.

Maggioni, A. P., Anker, S. D., ... , Heart Failure Association of the ESC (HFA). (2013). Are hospitalized or ambulatory patients with heart failure treated in accordance with European Society of Cardiology guidelines? Evidence from 12 440 patients of the ESC Heart Failure Long-Term Registry. *European journal of heart failure*, **15(10)**, 1173 – 1184.

The power of Laplacian-P-splines for inference in epidemiological and survival models

Oswaldo Gressani[1], Niel Hens[1,2], Christel Faes[1]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Belgium
[2] Centre for Health Economics Research and Modelling Infectious Diseases, Vaxinfectio, University of Antwerp, Belgium

E-mail for correspondence: oswaldo.gressani@uhasselt.be

Abstract: In Bayesian statistics one typically relies on Markov chain Monte Carlo (MCMC) methods to extract information from (potentially complex) posterior distributions. Existing MCMC samplers are straightforward to implement in practical applications and are therefore often considered as the best companion for Bayesian inference. The stochastic and iterative nature of these algorithms makes them computationally expensive and even with modern multi-core machines, the waiting time for drawing posterior samples required for inference can span several hours or even days depending on the model complexity. Laplacian-P-splines (LPS) is a new "sampling-free" methodology for approximate Bayesian inference that provides a lightning fast alternative to costly MCMC samplers. Laplace approximations to selected conditional posterior distributions and Bayesian penalized B-splines are the two main forces making LPS a fast and flexible modeling tool. The presentation is split in two parts in order to highlight the strength of LPS in recent implementations within the framework of epidemiological models (Part I) and cure survival models (Part II).

Part I: Estimation of the time-varying reproduction number

The instantaneous reproduction number $\mathcal{R}_t$ (defined as the expected number of secondary cases generated by an infected individual at time t) is a statistic that plays a central role in infectious disease epidemiology. It characterizes the global transmission potential of a pathogen during an epidemic outbreak in a single number and is therefore an interesting summary measure to assist policy makers in the management of a public health crisis. EpiLPS (**E**pidemiological

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

modeling with **L**aplacian-**P**-**S**plines) is a new flexible Bayesian tool developed by Gressani et al. (2021a) that provides smooth estimates of the epidemic curve and $\mathcal{R}_t$ based on case incidence data and the serial interval distribution (the time elapsed between the onset of symptoms in an infector and the onset of symptoms in the secondary cases generated by that infector). The proposed methodology builds upon P-splines smoothers (Eilers and Marx, 1996) to approximate the mean number of incidence cases by assuming a negative binomial distribution for the daily case counts and relies on Laplace approximations to the conditional posterior of the spline parameters to speed up the inference process. The estimated spline parameters are then nested within a (discrete) renewal equation model (Fraser, 2007; Wallinga and Lipsitch, 2007) to derive daily estimates of the reproduction number and associated credible intervals. A key argument that makes EpiLPS an attractive tool to quantify the time-varying reproduction number is the absence of any sliding window assumption. The popular benchmark method of Cori et al. (2013) requires the specification of a time window to estimate $\mathcal{R}_t$ that implies a trade-off between potential oversmoothing (with a "large" time window) and undersmoothing (with a "narrow" time window). EpiLPS is not facing such a trade-off, as P-splines deal with smoothing internally, i.e. within the model. Furthermore, EpiLPS is designed to give the user a choice between (1) a fully sampling-free approach to infer $\mathcal{R}_t$ based on a maximum *a posteriori* calibration of the model hyperparameters and (2) a gradient-based MCMC algorithm based on the Langevin diffusion for efficient sampling of the posterior distribution. Extensive simulation results show that the $\mathcal{R}_t$ estimate computed by EpiLPS exhibits excellent statistical performance with a relatively low computational cost. To conclude the first part of the talk, EpiLPS is applied on historical outbreak datasets and on the recent SARS-CoV-2 pandemic. Strengths and limitations of EpiLPS are briefly summarized and the use of LPS to model $\mathcal{R}_t$ under misreported data (Gressani et al. 2021b) is discussed. Finally, we give an overview of the associated R package (Gressani 2021c) and the user-friendly website https://www.epilps.com dedicated to the EpiLPS tool.

Part II: Laplacian-P-splines in mixture cure survival models

The mixture cure model for analyzing survival data is characterized by the assumption that the population under study is divided into a group of subjects who will experience the event of interest over some finite time horizon and another group of cured subjects who will never experience the event irrespective of the duration of follow-up. When using the Bayesian paradigm for inference in survival models with a cure fraction, it is common practice to rely on MCMC methods to sample from posterior distributions. Although computationally feasible, the iterative nature of MCMC often implies long sampling times to explore the target space with chains that may suffer from slow convergence and poor mixing. Furthermore, extra efforts have to be invested in diagnostic checks to monitor the reliability of the generated posterior samples. A sampling-free strategy for fast and flexible Bayesian inference in the mixture cure model is suggested (see Figure 1) by combining Laplace approximations and penalized B-splines (Gressani et al. 2022). A logistic regression model is assumed for the cure proportion and a Cox proportional hazards model with a P-spline approximated baseline hazard is used to specify the conditional survival function

of susceptible subjects. Laplace approximations to the posterior conditional latent vector are based on analytical formulas for the gradient and Hessian of the log-likelihood, resulting in a substantial speed up in approximating posterior distributions. The spline specification yields smooth estimates of survival curves and functions of latent variables together with their associated credible interval are estimated in seconds. A fully stochastic algorithm based on a Metropolis-Langevin-within-Gibbs sampler is also suggested as an alternative to the proposed Laplacian-P-splines mixture cure (LPSMC) methodology. The statistical performance and computational efficiency of LPSMC is assessed in a simulation study. Results show that LPSMC is an appealing alternative to MCMC for approximate Bayesian inference in standard mixture cure models. Finally, the novel LPSMC approach is illustrated on real survival data.

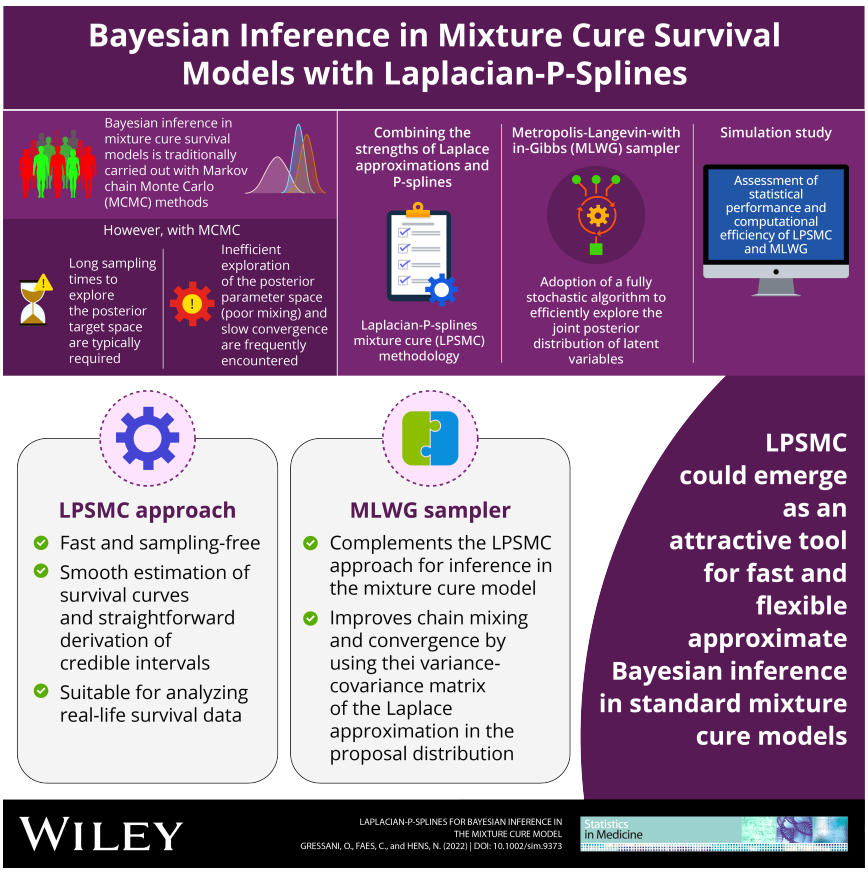


FIGURE 1. The Laplacian-P-splines methodology for Bayesian inference in mixture cure survival models.

Keywords: Laplace approximation; Approximate Bayesian inference; Epidemi-

ology; Survival analysis; Computational methods.

Acknowledgments: This research is funded by the European Union's Research and Innovation Action under the H2020 work programme, EpiPose (grant number 101003688).

References

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9):1505–1512. https://doi.org/10.1093/aje/kwt133

Eilers, P.H.C., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Fraser, C. (2007). Estimating individual and household reproduction numbers in an emerging epidemic. *PloS One*, 2(8):e758.

Gressani, O., Wallinga, J., Althaus, C. L., Hens, N. and Faes, C. (2021a). EpiLPS: a fast and flexible Bayesian tool for near real-time estimation of the time-varying reproduction number. *MedRxiv preprint.* Available at: https://doi.org/10.1101/2021.12.02.21267189

Gressani, O., Faes, C., and Hens, N. (2021b). An approximate Bayesian approach for estimation of the reproduction number under misreported epidemic data. *MedRxiv preprint.* https://doi.org/10.1101/2021.05.19.21257438

Gressani, O. (2021c). EpiLPS: a fast and flexible Bayesian tool for estimation of the time-varying reproduction number. [Computer Software]. *CRAN.* https://cran.r-project.org/package=EpiLPS

Gressani, O., Faes, C. and Hens, N. (2022). Laplacian-P-splines for Bayesian inference in the mixture cure model. *Statistics in Medicine (Early View).* https://doi.org/10.1002/sim.9373

Wallinga, J. and Lipsitch, M. (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society B: Biological Sciences*, 274(1609):599–604.

Variable Selection and Allocation in Joint Models via Gradient Boosting Techniques

Colin Griesbach[1], Andreas Mayr[2], Elisabeth Bergherr[1]

[1] Chair of Spatial Data Science and Statistical Learning, University of Göttingen, Germany
[2] Department of Medical Biometrics, Informatics and Epidemiology, University of Bonn, Germany

E-mail for correspondence: colin.griesbach@uni-goettingen.de

Abstract: In this work we construct a data-driven allocation algorithm for basic joint models for longitudinal and time-to-event data by applying recent developments from gradient boosting for distributional regression. Instead of specifying beforehand which covariate has an influence on which part of the joint model, the algorithm allocates the covariates to the appropriate sub-model.

Keywords: Joint Modelling, Statistical Learning, Variable Selection

1 Motivation

Modelling longitudinal data and risk for events separately, even though the underlying processes are related to each other, leads to loss of information and bias. Hence, the popularity of joint models for longitudinal and time-to-event data (Wulfsohn and Tsiatis, 1997) has grown rapidly in the last few decades. The basic idea of joint modelling is to formulate sub-models for each of the outcomes and estimate them jointly in one single framework. This rises the important question to which of the given sub-models a candidate variable should be assigned to, which researchers usually have to decide based on background knowledge. Gradient boosting, on the other hand, is a statistical learning method that has the inherent ability to select variables and estimate them simultaneously. We use gradient boosting techniques for distributional regression (Thomas et al., 2018) in order to extend existing boosting approaches for joint models (Waldmann et al., 2017) to an allocation routine which is capable of assigning possibly high numbers of predictor effects to the given sub-models specified for one joint model. The

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

newly designed routine is further equipped with recent developments from the field of statistical boosting including a novel correction for the random effects estimates (Griesbach et al., 2021), adaptive step-lengths (Zhang et al., 2022) and a fast tuning procedure based on probing (Thomas et al., 2017).

2 Methods

2.1 Joint Modelling

For longitudinal outcome y, a joint model consists of one longitudinal sub-model

$$y = \eta_{\text{long}}(\boldsymbol{x}_{\text{long}}, t) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

where y is modelled by some longitudinal predictor η_{long} depending on a set of longitudinal covariates $\boldsymbol{x}_{\text{long}}$ and time t itself. On the other hand, the time-to-event outcome (T, δ) is modelled by

$$\lambda(t) = \lambda_0(t) \exp\left\{\eta_{\text{surv}}(\boldsymbol{x}_{\text{surv}}) + \alpha \eta_{\text{long}}(\boldsymbol{x}_{\text{long}}, t)\right\},$$

where each individuals hazard $\lambda(t)$ consists of a baseline hazard $\lambda_0(t)$, a survival predictor η_{surv} depending on additional baseline survival covariates $\boldsymbol{x}_{\text{surv}}$ and, most importantly, the longitudinal predictor reappearing in the formulation of the hazard function. This time scaled by the association parameter α which quantifies the impact of the longitudinal model on the time-to-event outcome. Let $\boldsymbol{\vartheta}$ denote the collection of parameters specifying the two sub-models. Given some necessary independency assumptions, one can derive the joint likelihood $\mathcal{L}(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{T}, \boldsymbol{\delta})$ based on the sub-models which then is to be maximized with respect to $\boldsymbol{\vartheta}$ for regular likelihood inference.

2.2 Boosting Joint Models

Model-based gradient boosting is an iterative and component-wise fitting procedure for various classes of regression models. Key concept is that in each iteration the current gradient $\boldsymbol{u}$, often resembling some kind of residuals, of a pre-chosen loss function ρ is fitted to every single candidate variable. Only the best performing variable of each iteration gets selected into the model by a tiny amount enabling variable selection and the possibility to estimate models with potentially high numbers of covariates. This procedure can be adapted to joint models in the following way: First, choose the loss $\rho = -\mathcal{L}$ as the negative joint likelihood and a total number of iterations m_{stop} and then let the boosting procedure cycle through the single predictors of the sub-models in each iteration where only the overall best-performing covariate gets allocated to its corresponding sub-model. In a very simple manner, the mechanism can be summarized as follows:

- initialize predictors $\hat{\eta}^{[0]}_{\text{long}}$, $\hat{\eta}^{[0]}_{\text{surv}}$ and $\hat{\alpha}^{[0]}$

  `for` $m = 1, \ldots, m_{\text{stop}}$ `do`

- longitudinal boosting step
  - Fit all variables $\boldsymbol{x}_r$, $r = 1, \ldots, p$, to the longitudinal gradient $\boldsymbol{u}^{[m]}_{\text{long}} = y - \hat{\eta}^{[m-1]}_{\text{long}}$
  - Find the best performing variable $\boldsymbol{x}^*_{\text{long}}$
- survival boosting step
  - Fit all variables $\boldsymbol{x}_r$, $r = 1, \ldots, p$, to the survival gradient $\boldsymbol{u}^{[m]}_{\text{surv}} = \delta - \int_0^T \hat{\lambda}^{[m-1]}(t)dt$
  - Find the best performing variable $\boldsymbol{x}^*_{\text{surv}}$
- allocation step
  - find $\boldsymbol{x}^* \in \{\boldsymbol{x}^*_{\text{long}}, \boldsymbol{x}^*_{\text{surv}}\}$ yielding the best improvement of ρ.
  - allocate $\boldsymbol{x}^*$ to its sub-model obtaining $\hat{\eta}^{[m]}_{\text{long}}$ and $\hat{\eta}^{[m]}_{\text{surv}}$
  - update $\hat{\alpha}^{[m]}$ by optimizing ρ w.r.t. the current fit

  `end for`

3 Showcase

We showcase the algorithm briefly sketched in the previous section by simulating and estimating a joint model with linear effects in both sub-models as well as a real world appilcation.

3.1 Simulation

Set $\eta_{\text{long}}(\boldsymbol{x}_{\text{long}}) = \boldsymbol{x}_{\text{long}}\boldsymbol{\beta}_{\text{long}}$ and $\eta_{\text{surv}}(\boldsymbol{x}_{\text{surv}}) = \boldsymbol{x}_{\text{surv}}\boldsymbol{\beta}_{\text{surv}}$ with coefficient vectors

$$\boldsymbol{\beta}^T_{\text{long}} = (1, 2, 1, 2), \quad \boldsymbol{\beta}^T_{\text{surv}} = (1, 2, 1, 2), \quad \boldsymbol{\beta}^T_{\text{noise}} = (0, 0, 0).$$

This means there are eleven candidate variables in total. Four have a linear influence on the longitudinal outcome and four have a linear influence on the time-to-event outcome. The remaining three variables are noise variables which are non-informative for any of the two predictors. In this case, variable allocation simplifies to the question in which coefficient vector the effect of a given covariate should appear in: either in $\boldsymbol{\beta}_{\text{long}}$ or in $\boldsymbol{\beta}_{\text{surv}}$ (or neither of the two). Figure 1 depicts the coefficient progression in each sub-model of the described scenario. As one can see, all informative variables get assigned to the correct sub-models. Due to the tuning of the algorithm, the procedure stops shortly after 160 iterations leading to variable selection. However, one noise variable is falsely selected into the longitudinal sub-model after 60 iterations.

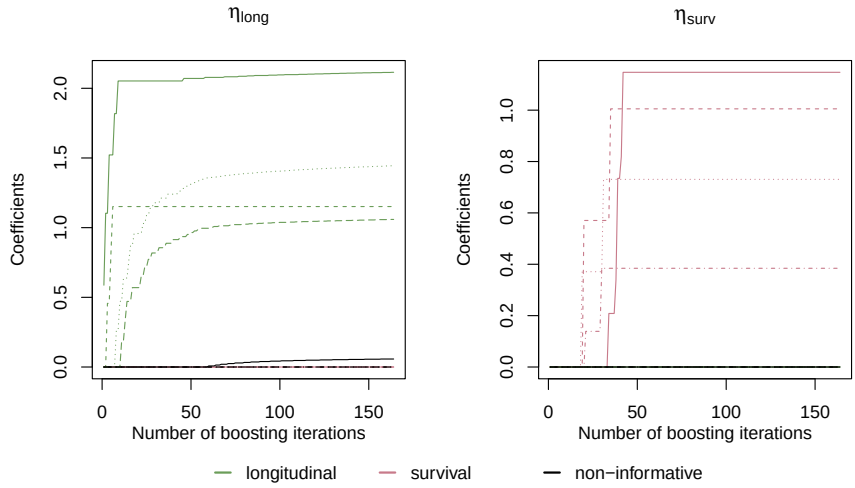


FIGURE 1. Progression of longitudinal and survival coefficient paths for correlated covariates. All informative variables have been correctly allocated. One noise variable was falsely selected into η_{long}.

3.2 AIDS Data

The 1994 AIDS data (Abrams et al., 1994) aimed at comparing two antiretroviral drugs based on a collective of HIV positive patients. It includes 1405 longitudinal observations of 467 individuals from which 188 died during the course of the study. Apart from the CD4 cell count as longitudinal outcome, death as time-to-event outcome and time t itself, the four additional baseline variables `drug` (treatment group), `gender`, `AZT` (indicator whether a previous AZT therapy failed) and `AIDS` (indicator whether AIDS is diagnosed) are observed. Figure 2 depicts the coefficient paths computed by the `JMalct` algorithm and the corresponding allocation process. The variable `AIDS` is selected into the longitudinal sub-model right away and frequently updated. This is not surprising, as diagnosis of AIDS is by definition linked to the CD4 cell count. `drug` and `gender` are also allocated to the longitudinal sub-model by a smaller amount whereas `AZT` is selected into the survival predictor indicating an increased risk of death for patients with failed AZT therapy.

4 Discussion

By applying established methods from the field of gradient boosting for distributional regression, it is possible to construct a fast-performing and data-driven allocation routine for joint models which selects and allocates

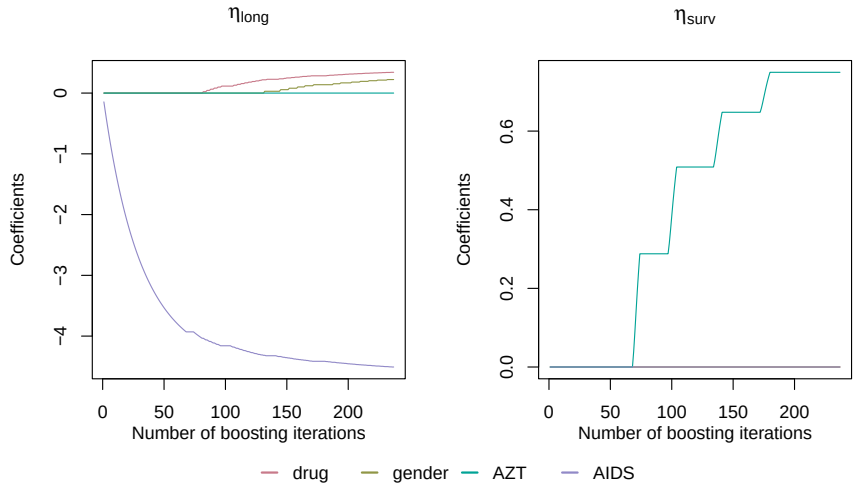


FIGURE 2. Coefficient progression in both sub-models for AIDS data. The variable `AZT` has been assigned to η_{surv}, the rest to η_{long}.

predictor effects into the single sub-models. While the allocation works reasonably well, estimates experience some shrinkage, which is well-known for regularization methods. Possible tasks for future research include extension of the concept to more general and flexible joint models in order to make it applicable to a broader variety of real world situations.

Acknowledgments: The work on this contribution was supported by the DFG (Deutsche Forschungsgemeinschaft - Projektnummer 426493614).

References

Abrams, D. I., Goldman, A. I., Launer, C., et al. (1994). A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *New England Journal of Medicine*, **330(10)**, 657 – 662.

Griesbach, C., Säfken, B., and Waldmann, E. (2021). Gradient Boosting for Linear Mixed Models. *The International Journal of Biostatistics*, **17(2)**, 317 – 329.

Thomas, J., Hepp, T., Mayr, A., and Bischl B. (2017). Probing for Sparse and Fast Variable Selection with Model-Based Boosting. *Computational and Mathematical Methods in Medicine*. Article ID 1421409.

Thomas, J., Mayr, A., Bischl, B., et al. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, **28**, 673 – 687.

Waldmann, E., Taylor-Robinson, D., Klein, N., et al. (2017). Boosting joint models for longitudinal and time-to-event data. *Biometrical Journal*, **59(6)**, 1104 – 1121.

Wulfsohn, M., and Tsiatis, A. (1997). A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, **53(1)**, 330 – 339.

Zhang, B., Hepp, T., Greven, S., and Bergherr E. (2022). Adaptive steplength selection in gradient boosting for Gaussian location and scale models. *Computational Statistics*.

Categorizing the World into Local Climate Zones - Towards Quantifying Labelling Uncertainty for Machine Learning Models

Katharina Hechinger[1], Xiao Xiang Zhu[2], Göran Kauermann[1]

[1] Ludwig-Maximilians-University Munich, Germany
[2] Technical University of Munich, Germany

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

Abstract: Many applications of image classification are prone to labeling uncertainty. To generate suitable datasets, images are often assessed by human experts and labeled according to their evaluation. This can result in ambiguities and errors, which affect any kind of subsequent machine learning model. In this work, we aim to quantify the uncertainty of the labelers in the context of remote sensing and classification of satellite images. We do so by applying classical statistical technology and exhibit different sources of ambiguity during the labeling process.

Keywords: Uncertainty; Multinomial Mixture Models; EM Algorithm.

1 Problem Description and Data

Today, machine learning is increasingly used for the classification of images, with applications in medical image analysis, face recognition and many more. In this work, we focus on satellite images and their use to classify the world into Local Climate Zones (LCZ), see Figure 1. This concept proposed in Steward (2011) is a general standard in remote sensing and assumes that the structure of landscape influences the local climate. Massive effort has been spent in developing algorithms that transform satellite images into a LCZ map (Qiu et al. 2019). Thereby the algorithms are based on and require labeled data for training.
We here focus on this additional layer of uncertainty, which is often omitted, namely that the ground truth remains unknown. In our case, this uncertainty is analysed based on the earth observation benchmark data set

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

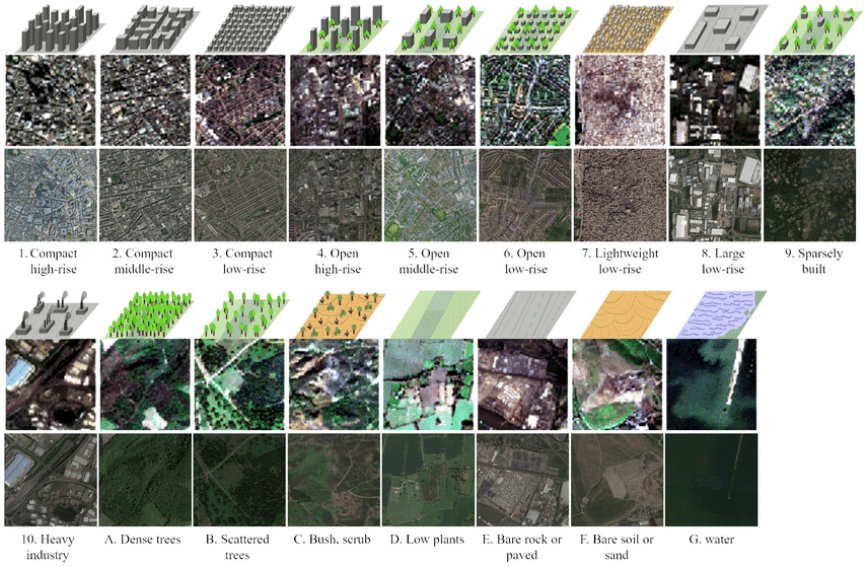


FIGURE 1. Representation of the 17 LCZs as Sentinel-1 (first row), Sentinel-2 (second row) and Google Maps images (third row).

So2Sat LCZ42, see Zhu et al. (2020). It comprises satellite images that were manually labeled by experts into LCZs. All in all, 159581 images from 9 citys were categorised into 17 classes by 11 experts. This process is time consuming and not without ambiguities. We aim to model the uncertainty of the experts about the images by applying classical statistical technology.

2 Modelling Annotation Uncertainty

To achieve the goal of exploring labeling uncertainty, we look at the experts' votes. Each image patch $i, i = 1, ...n$ is assessed by a set of experts indexed with $j, j = 1, ..., J$ and classified into the LCZ k where $k = 1, ..., K$. The corresponding vote of the expert is denoted by $V_j^{(i)} \in \{1, \ldots, K\}$. Rewriting the vote information as indicator vector allows to accumulate the votes into the data points $\boldsymbol{Y}^{(i)} = (Y_1^{(i)}, ..., Y_K^{(i)})$. We assume further that each image comes from a single true class (=ground truth), which is a reasonable assumption based on the clustered data structure described above. Hence we assume that there is no ambiguity of the image class, but apparently there are ambiguities in the voters' opinion about this class. We denote with $Z^{(i)} \in \{1, \ldots, K\}$ the true class of image i, which apparently remains unknown. Our intention is now to get information on $Z^{(i)}$ given the voters' distribution $\boldsymbol{Y}^{(i)}$. We will therefore apply Bayesian reasoning, which requires to formulate a distribution framework. For the true

classes, we assume a multinomial prior, i.e. $\boldsymbol{Z}^{(i)} \sim Multi(\boldsymbol{\pi}, 1)$, i.i.d. with $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$. Given the true class we further assume that the labelers' vote also follows a multinomial distribution, i.e.

$$\boldsymbol{Y}^{(i)} | Z^{(i)} \sim Multi(\boldsymbol{\theta}_{Z^{(i)}}, J). \tag{1}$$

We collect the coefficients into the confusion matrix $\Theta = (\theta_{pk}, p, k = 1, \dots K)$. Given that the true classes are unobserved, we are in the framework of mixture models and we obtain the likelihood contribution of the ith image by summing over all classes. Apparently, this is getting clumsy, so that we apply the EM algorithm, or to be more precise, a stochastic version of it as proposed by Celeux et al. (1996). A welcome advantage of SEM is that we can directly quantify uncertainty of the estimates, in the form of estimation variance of the parameters, primarily of the (mis)classification matrix Θ, see Rubin (1976).
Like in every mixture model, the numbering of the resulting classes does not match the original numbering of the LCZs, also referred to as label switching. We construct the permutation $\sigma()$ of the cluster labels C to the voter labels L such that its inverse fulfills $\sigma^{-1}(k) = \arg\max_l \big(P(Z^{(i)} = l | V^{(i)} = k)\big)$. This rule is applied repeatedly to get a unique definition.

3 Sources of uncertainty in the votes

We are now in the position to approach different questions related to human annotation of satellite images which cover various aspects of label uncertainty.
The first question we want to answer is: how distinguishable are the LCZs in general? Due to the definition of the LCZs, it is obvious that some classes are harder to distinguish than others. The parameter of main interest is the estimated confusion matrix $\hat{\Theta}$ shown in Figure 2. Looking at the diagonal entries, most classes seem to be well separable, whereas 1 and 7 could rarely be detected correctly.
Generally, our results depend on the input votes as the algorithm can only detect classes where the data basis is sufficient. Furthermore, it should be mentioned here our true confusion matrix is subject to the implemented label switching process. As the multinomial mixture model produces "meaningless" clusters, that have to be assigned to LCZs afterwards, the resulting estimates and their interpretation are based on the assignment strategy, which might not be unambiguous. Generally, however, we obtain interpretable insight in the inevitable ambiguity when classifying LCZs.
Second, we pose the question, whether experts are biased or homogeneous in terms of their voting behaviour. The described model allows to assess the impact of each individual experts and their heterogeneity. If experts were homogeneous, their voting behaviour does not differ, and dropping the votes of one expert at a time should not change the final estimated

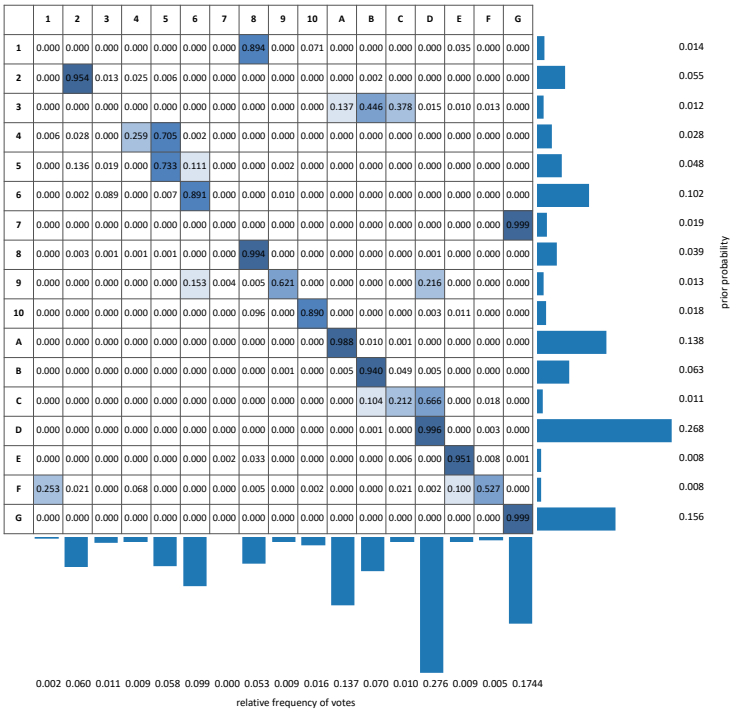


FIGURE 2. The matrix shows the true confusion of the voted (columns) with the true classes (rows), along with the estimated prior probabilities (right) and the relative vote frequency (bottom).

distribution. The parameter of interest here are therefore $\hat{\boldsymbol{\tau}}_l^{(i)}$ or $\hat{\boldsymbol{\tau}}_{(-j)l}^{(i)}$, where the bracketed index $-j$ refers to the excluded voter j. We use the modified Chi-squared measure and compare the observed votes of expert j against the expected counts. The latter is thereby given by $\hat{\tau}_{(-j)k}^{(i)}$. To assess the magnitudes of the resulting measures, we make use of a simulation based procedure, that resembles a parametric bootstrap approach. We conclude that for 6 out of 11 experts the observed votes match the expected votings, while 5 voters seem to be more heterogeneous.

Depending on the application at hand, expert heterogeneity is often not only accepted but also desired. In this particular case, experts received the same training on how to conduct classification of satellite images and are generally assumed to produce homogeneous votings.

Third, we want to know whether the voting behaviour is influenced by geographic differences. The polygons used for the voting procedure come from

9 different European cities, which are known to be quite diverse in terms of structure and architecture. While this might be intended to cover all LCZs as good as possible, it complicates the assessment of the images. The question is whether earth observation experts have difficulties in assigning certain images to certain climate zones, depending on the respective region. The crucial aspects here are the misclassification probability matrices $\Theta(s)$ for $s = \{1, ..., S\}$ denoting the region/city. However, one has to note here that the voting distribution shown in the bottom plot of Figure 2 depends on the initial draw of images or polygons in each city. The pursued strategy might lead to imbalanced labels and therefore a bias in the voting probabilities. We here are however interested in the confusion matrix and whether this matrix differs in the different cities. We construct a test statistic and test the null hypothesis of equal confusion matrices for citys s and s'. We repeat this procedure and construct pairwise tests for all cities. On a significant level of 0.05, we can assume that the confusion matrices are different for 22 of 34 city pairs.

4 Conclusion

The paper demonstrates that labeling of images is subject to error, misclassification and heterogeneity of labelers. The results are relevant for all applications where image classification is pursued and image labeling is subject to humans. It is important to note that error and uncertainty in the labeling process might stem from different sources and is multi-dimensional, as we showed in Section 3. In the context of classifying satellite images into climate zones, we were able to detect three main sources of label uncertainty. These can be analyzed based on the assumption that a latent ground truth label exists, on which the labelers condition their assessment. As a next step, it would be useful to include the uncertainty into the machine learning process as well. The labeling process only serves as a preprocessing step of the data at hand and produces a labeled training data. The results obtained by analyzing the sources of labeling uncertainty and being able to quantify them could create possibilities to improve and stabilize machine learning processes in terms of overall uncertainty.

Acknowledgments: The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS-006 - Munich School for Data Science@Helmholtz, TUM&LMU

References

Steward, I. D. (2011). Local climate zones: Origins, development, and application to urban heat island studies. *Proceedings of the Annual*

Meeting of the American Association of Geographers, Seattle, WA, 12 – 16.

Qiu, C. and Mou, L. and Schmitt, M. and Zhu, X. X. (2009). Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network. *ISPRS Journal of Photogrammetry and Remote Sensing*, **154**, 151 – 162.

Celeux, G. and Chauveau, D. and Diebolt, J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, **55**, 287 – 314.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581 – 592.

Zhu, X. X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., ... & Wang, Y. (2020). So2Sat LCZ42: A benchmark data set for the classification of global local climate zones. *IEEE Geoscience and Remote Sensing Magazine*, **8(3)**, 76 – 89.

Functional regression on variable domains: a penalized approach.

Pavel Hernández-Amaro[1], María Durbán [1], M. Carmen Aguilera-Morillo[2], Cristobal Esteban Gonzalez[3], Inma Arostegui[4]

[1] University Carlos III of Madrid, Spain
[2] Universitat Politècnica de València, Spain
[3] Osakidetza Basque Health Service, Spain
[4] University of the Basque Country UPV/EHU, Spain

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

Abstract: In this work we present a novel methodology to estimate a variable domain functional regression model assuming the basis representation of both, the functional coefficient and the functional data. The model's coefficients are estimated via Penalized Quasi-likelihood using the mixed model representation of a penalized spline. We test our methodology in a simulation study and apply it to a data set of COPD patients.

Keywords: Variable domain functional regression model, basis representation, P-spline, COPD.

1 Introduction

Variable domain functional regression models are an extension of scalar-on-function models where the functional predictor is observed on a grid of different length for each subject. Such type of data have become very common recently, specially due to the wide-spread use of wearable devices and their ability to collect data that can improve, for example, health diagnostics. Our proposal is inspired by the functional regression model proposed in Gellar et al. (2014) and it is motivated by the assumption that the functional covariate is smooth but observed with error in practice. To deal with this problem, a two-dimensional anisotropic penalty is considered to provide a good estimation of the proposed model. The good performance

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

of the proposed approach has been shown on a simulation study and a real data set.

2 Methodology

Given the following sample data: $\{Y_i, \boldsymbol{C}_i, X_i(t)\}$, $i = 1, \ldots, N, t \in [1, T_i]$, where $\boldsymbol{C}_i$ are the non-functional covariates, $X_i(t)$ is the functional covariate and T_i is the length of the domain of the variable t for the subject i that satisfies $T_i \leq T_{i+1}$. The response variable Y_i follows an exponential family distribution with mean μ_i. Gellar et al. (2014) proposed the following generalized variable domain functional regression model:

$$\eta_i = g(\mu_i) = \alpha + \boldsymbol{C}_i\boldsymbol{\gamma} + \frac{1}{T_i}\int_1^{T_i} X_i(t)\beta(t, T_i)\,\mathrm{d}t, \quad t \in [1, T_i]. \tag{1}$$

Their approach was based on a varying coefficient model, we propose an alternative approach for the estimation of the model assuming the basis representation of both, the functional coefficient and the functional data. Therefore, our first step is to make a basis representation of the functional variable $X_i(t)$ and the bi-dimensional functional coefficient $\beta(t, T_i)$:

$$X_i(t) = \sum_{j=1}^{p} a_{ij}\phi_{ij}(t) = \boldsymbol{\phi}_i\boldsymbol{a}_i,$$

$$\beta(t, T_i) = \sum_{l=1}^{q}\sum_{k=1}^{r} b_{lk}\varphi_{il}(t)\psi_{ik}(T_i) = (\boldsymbol{\varphi}_i \otimes \boldsymbol{\psi}_i)\boldsymbol{b} = \boldsymbol{M}_i\boldsymbol{b},$$

with $(\boldsymbol{\phi}_i)_{T_i\times p}$, $(\boldsymbol{\varphi}_i)_{T_i\times q}$, $(\boldsymbol{\psi}_i)_{1\times r}$ the basis used in the representation of the functional data $X_i(t)$ and the functional coefficient $\beta(t, T_i)$ respectively with $\boldsymbol{a}_i$ and $\boldsymbol{b}$ their respective coefficients and $M_i = \boldsymbol{\varphi}_i \otimes \boldsymbol{\psi}_i$ where $\otimes$ represents the Kronecker product.
The basis $\boldsymbol{\phi}_i$ can be described as the T_i first rows of the basis $(\boldsymbol{\phi}_N)_{T_N\times p}$ which is the basis of the subject with more observations, and $\boldsymbol{\psi}_i$ is the i-th row of a basis $\boldsymbol{\psi}_{N\times r}$. This basis is associated to the number of observations of all the subjects: $\boldsymbol{T} = [T_1, \ldots, T_N]$, hence for subject i we only use the i-th row. We use B-splines for all our basis representations.
With these representations, our model is transform into a multivariate regression model:

$$\begin{aligned}\boldsymbol{\eta} &= \boldsymbol{\alpha} + \boldsymbol{C}\boldsymbol{\gamma} + \tfrac{1}{\boldsymbol{T}}\textstyle\int_{\boldsymbol{T}} X(t)\beta(t, T)\,\mathrm{d}t \\ &= \boldsymbol{\alpha} + \boldsymbol{C}\boldsymbol{\gamma} + \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{b} = \boldsymbol{B}\boldsymbol{\theta},\end{aligned}$$

with $(\boldsymbol{A})_{N\times Np}$ a block diagonal matrix which i-th block of the diagonal is $\boldsymbol{a}_i^T$ and $(\boldsymbol{\Psi})_{Np\times qr} = (\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_N)^T$ where $\boldsymbol{\Psi}_i = \frac{1}{\boldsymbol{T_i}}\int_{T_i} \boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{M_i}\,\mathrm{d}t$. The calculation of the matrix of inner products $\boldsymbol{\Psi}$ is one of the novelties of our work. One of the mayor difficulties for its calculation is the fact that the product is between a one-dimensional base and a two-dimensional one.

We perform the integration calculations only in the t dimension but perform the product of the two whole basis while maintaining the proper two-dimensional structure. As far as we know this type of inner product has not been done before. Then for the estimation of the model coefficients we will use Penalized Maximum Likelihood, particularly an anisotropic two dimensional penalization will be used, allowing us to control the smoothness of the functional coefficient independently for each dimension

$$L_p(\boldsymbol{\theta}, \boldsymbol{y}) = L(\boldsymbol{\theta}, \boldsymbol{y}) - \tfrac{1}{2}\boldsymbol{\theta}^{\mathbf{T}}\boldsymbol{P}\boldsymbol{\theta},$$

where $L(\boldsymbol{\theta}, \boldsymbol{y})$ is the likelihood of Y, $\boldsymbol{P} = \lambda_t(\boldsymbol{I}_r \otimes \boldsymbol{D}_2^{qT}\boldsymbol{D}_2^q) + \lambda_T(\boldsymbol{I}_q \otimes \boldsymbol{D}_2^{rT}\boldsymbol{D}_2^r)$ is the penalty matrix and $\boldsymbol{D}_2^{qT}$,$\boldsymbol{D}_2^{rT}$ are matrices of second order difference of adjacent coefficients of $\boldsymbol{\theta}$.
In order to efficiently estimate the smoothing parameters λ_t and λ_T jointly with the rest of the parameters in the model we reparametrize our model as a mixed model. Therefore, we are in the context of Generalized Linear Mixed Models (GLMMs) and we use Penalized Quasi-Likelihood (Breslow N. E. et al, 1993) for parameter estimation. To speed up computations we make use of the SOP(Separation of Overlapping Penalties) algorithm (Rodríguez-Álvarez, M. et al., 2019). In the next sections we will show the good performance of our methodology via a simulation study.

3 Simulation study

Our methodology is tested in a simulation study based on the one presented in Gellar et al. (2014) and we compare the obtained results of using three different methodologies: Our approach named "*New approach*", the Gellar approach and the Goldsmith approach (Goldsmith et al., 2011); this approach perform a preregistration of the curves to a common domain and then estimate the usual functional regression model. Our simulation model is the following:

$$Y_i = \eta_i + \epsilon_i, \quad \epsilon_i \sim N(0,1),\ i = 1, \dots N = 100$$
$$\eta_i = \tfrac{1}{T_i}\sum_{t_j=1}^{T_i} X_i(t_j)\beta_b(t_j, T_i),$$

where the domain of the sample curves T_i is simulated using an uniform distribution: $T_i \sim \text{Uniform}(10, 100)$. The true functional coefficients $\boldsymbol{\beta}_b(t, T_i)$ are simulated by the following functions:

$$\beta_1(t,T_i) = 10\frac{t}{T_i} - 5\,;\ \beta_2(t,T_i) = \left(1 - \frac{2T_i}{J}\right) \times \left(5 - 40\left(\frac{t}{T_i} - 0.5\right)^2\right);$$

$$\beta_3(t,T_i) = 5{-}10\left(\frac{T_i - t}{J}\right);\ \beta_4(t,T_i) = sen\left(\frac{2\pi T_i}{J}\right) \times \left(5 - 10\left(\frac{T_i - t}{J}\right)\right).$$

The functional data is simulated from two different scenarios.
Scenario 1: The functional data is generated as smooth curves

$$X_i(t_j) = u_i + \sum_{k=1}^{10} \left\{ v_{ik1} \cdot sen\left(\frac{2\pi k}{J} t_j\right) + v_{ik2} \cdot \cos\left(\frac{2\pi k}{J} t_j\right) \right\}$$
$$u_i \sim N(0,1), \quad v_{ik1}, v_{ik2} \sim N(0, \tfrac{4}{k^2}), \quad t_j = 1, \ldots, J = 100.$$

Scenario 2: The functional data is generated as noisy curves by adding noise to the previous smooth curves

$$X_i(t_j) = u_i + \sum_{k=1}^{10} \left\{ v_{ik1} \cdot sen\left(\frac{2\pi k}{J} t_j\right) + v_{ik2} \cdot \cos\left(\frac{2\pi k}{J} t_j\right) \right\} + \delta_i(t_j)\,,$$

$u_i \sim N(0,1)$, $v_{ik1}, v_{ik2} \sim N(0, \frac{4}{k^2})$, $\delta_i(t_j) \sim N(0,1)$, $t_j = 1, \ldots, J = 100$.

Finally for all the possible scenarios our sample data set is the response variable Y_i and the noisy functional data X_i from **scenario 2**: $S = \{Y_i, X_i\}$, this will allow us to evaluate the ability to filter the noise of our methodology through the basis representation of the functional data.

We evaluate the performance of the three methodologies in the simulation study using the following evaluation criteria for the prediction and estimation errors respectively:

$$\begin{aligned} RMSE^r &= \sqrt{\frac{\left(\sum_{i}^{K} (Y_i^r - \hat{Y}_i^r)^2\right)}{K}} \\ AMSE^r &= \frac{1}{J(J+1)} \sum_{k=1}^{J} \sum_{j=1}^{k} \left\{ \beta_b^r(t_j, k) - \hat{\beta}_b^r(t_j, k) \right\}^2, \end{aligned}$$

where $\hat{Y}_i^r$ and $\hat{\beta}_b^r(t_j, k)$ are the estimation of the response variable and the functional coefficient respectively. For the estimation of the response variable we use a cross-validation K-fold approach considering $K = 10$. Finally we repeat our simulation $r = 100$ times. The obtained results are shown in Figure 1 and Figure 2 for the RMSE and AMSE respectively.

We can see that our methodology outperform the others with a great improvement in the estimation errors. This will be of great importance for studies in medical research where a good estimation of the functional coefficient directly imply a good feedback for patients on what to do for improving their health.

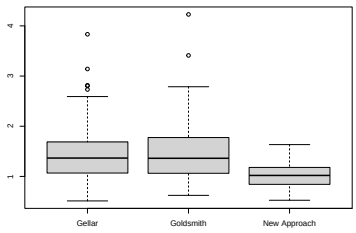


FIGURE 1. RMSE for the $Beta_1$ in **scenario 2**.

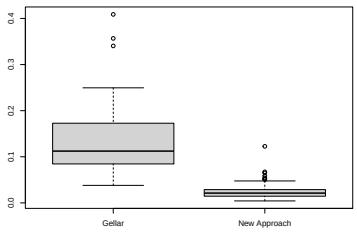


FIGURE 2. AMSE for the $Beta_1$ in **scenario 2**.

4 Real data application

TELEPOC data set (Esteban, C. et al, 2016) collects a wide range of data from 110 patients suffering Chronic Obstructive Pulmonary Disease(COPD). The aim is to study the relationship between physical activity, measured as daily steps given by each patient (functional covariate) and the annual ratio of hospitalizations due to COPD (response variable, $Y \sim Poi(\mu)$). The number of days where steps are collected is different from patient to patient; therefore we are in presence of variable domain functional data. Moreover, non-functional covariates have been considered providing the following variable-domain functional regression model:

$$\boldsymbol{\eta} = \log(\boldsymbol{\mu}) = \boldsymbol{\alpha} + \boldsymbol{C\gamma} + \frac{1}{\boldsymbol{T}} \int_{\boldsymbol{T}} X(t)\beta(t,T)\, \mathrm{d}t + \log\left(\frac{\boldsymbol{T}}{365}\right), \quad t \in [1,T].$$

4.1 Results

The functional parameter $\beta(t,T)$ represents the relationship between physical activity and the annual ratio of hospitalizations due to COPD. The correct estimation of this coefficient is of great importance because it permit us to determine how to improve the patient health. Figure 1 shows the estimated functional parameters for patients that carried out physical activity for different length periods. The curves present a common feature: doing physical activity regularly during more than 8 months help to reduce the mean number of hospitalizations due to COPD. From the rest of the curves we see how, regardless of the length, their values end up being negative and their slope decreasing, indicating a positive influence of physical activity in the reduction of the annual rate of hospitalizations. From the non-functional covariates included in the model we conclude: people that had been hospitalized before are more prone to suffer new hospitalizations, women are more susceptible to be hospitalized than men, depressive symptomatology rise the annual rate of hospitalizations and anxious symptomatology reduces it.

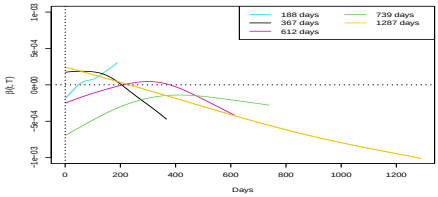


FIGURE 3. Curve $\beta(t, T_i)$ for patients with T_i days in the study.

5 Conclusions and future work

We have proposed a new methodology for the estimation of a variable domain functional regression model based on penalized B-splines approximation and their mixed model representation, using the SOP method for the estimation of the functional coefficients and the penalization parameters. This methodology is tested with a simulation study where results show our method outperforming other approaches, and the TELEPOC data set, where we conclude that regular physical activity helps to reduce the mean number of hospitalizations of COPD patients. We are currently working on the extension of this methodology for the case of more than one functional covariate.

Acknowledgments: This work is supported by the grant ID2019-104901RB-I00 from the Spanish Ministry of Science, Innovation and Universities, MCIN/AEI/10.13039/501100011033.

References

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, **88**, 9 – 25.

Esteban, C., et al. (2016). Outcomes of a telemonitoring-based program (telEPOC) in frequently hospitalized COPD patients. *Int J Chron Obstruct Pulmon DIS*, **11**, 1919.

Gellar, J. E., Colantuoni, E., Needham, D. M. and Crainiceanu, C. M. (2014). Variable-domain functional regression for modeling ICU data. *J. Am. Stat. Assoc.*, **109**, 1425 – 1439.

Goldsmith, Jeff, Bobb, Jennifer, Crainiceanu, Ciprian M., Caffo, Brian, & Reich, Daniel. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics*, **20(4)**, 830 – 851.

Rodríguez-Álvarez, M. X., Durban, M. and Lee, D-J. and Eilers, P. (2019). On the estimation of variance parameters in non-standard generalised linear mixed models: Application to penalised smoothing. *Statistics and Computing*, **29** , 483 – 500.

Using saturated models for data synthesis

James Jackson[1], Robin Mitra[2], Brian Francis[1], Iain Dove[3]

[1] Lancaster University, UK
[2] Cardiff University, UK
[3] Office for National Statistics, UK

E-mail for correspondence: j.jackson3@lancaster.ac.uk

Abstract: The use of synthetic data sets are becoming ever more prevalent, as regulations such as the General Data Protection Regulation (GDPR), which place greater demands on the protection of individuals' personal data, are coupled with the conflicting demand to make more data available to researchers. This paper discusses the approach of synthesizing categorical data at the aggregated (contingency table) level using a saturated count model, which adds noise - and hence protection - to cell counts. The paper also discusses how distributional properties of synthesis models are intrinsic to generating synthetic data with suitable risk and utility profiles.

Keywords: synthetic data; saturated count models; data privacy

1 Introduction

As organisations have both a legal and ethical obligation to protect individuals' personal data, data sets pertaining to individuals cannot be released directly to researchers. Thus prior to release, statistical disclosure control methods, such as the use of synthetic data sets, need to be applied.
Synthetic data sets (Rubin 1993, Little 1993), which are generated by simulating from a model fit to the original data, can be released to researchers in place of the original data. The notion is that, as the synthetic data sets are inherently artificial, individuals' privacy should be protected; while, as synthetic values are based on original values, researchers' ability to obtain valid inferences should remain undiminished. The method relies on the synthesizer – he or she responsible for generating the synthetic data – accurately modelling the data's underlying distribution.
The theory of synthetic data evolved from the multiple imputation of missing data theory (Rubin, 1987). The synthesizer either imputes values for in-

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

dividuals not included in the original data (resulting in fully synthetic data; Raghunathan 2003) or generates replacement values for those individuals who were included in the original data (resulting in partially synthetic data; Reiter 2003). As with imputation, it is typical to release multiple ($m > 1$) data sets to allow analysts – through combining rules; see Drechsler (2011) - to average point estimates and properly account for the extra uncertainty arising from synthesis when calculating estimates' variances.
When synthesizing a data set comprising p variables $Y_1, Y_2, \ldots, Y_p$, the underlying distribution of the data can be captured through a product of conditional models, that is,

$$f(Y_1, Y_2, \ldots, Y_p \mid X) = f_1(Y_1 \mid X) \prod_{j=2}^{p} f_j(Y_j \mid Y_{j-1} \ldots, Y_2, Y_1, X),$$

where X denotes any other data available to the synthesizer, such as other relevant data sets, census tables or administrative data.
The synthesis models for $Y_1, Y_2, \ldots, Y_p$ can take a variety of forms - parametric or non-parametric - ranging from generalised linear models (GLMs), to tree-based methods such as CART, to complex machine learning algorithms. The aim of all these methods, though, is the same: to model the underlying distribution governing the original data.
A categorical data set comprises categorical variables only. Its discrete nature allows the data to be aggregated into a contingency table, such that cell counts give the frequencies with which the various combination of categories (cells) are observed; a given set of categories may not be observed, in which case the cell count would be zero. Synthesis can take place by fitting a *count* model to this table, which is more convenient as the response is univariate rather than multivariate.

2 The motivation for using saturated models

The purpose of synthesis models, then, is for neither inference nor prediction, but to reproduce the structure of the original data. Therefore, unlike when estimating a population parameter, modelling assumptions are not intrinsic to obtaining meaningful estimates and standard errors. For this reason, Jackson et al. (2022) proposed the use of saturated count models for synthesis.
Let $f_1, f_2, \ldots, f_K$ denote the observed counts in the original data's contingency table (the original counts). Then the corresponding counts in the synthetic data's contingency table (the synthetic counts) $f_1^{\text{syn}}, f_2^{\text{syn}}, \ldots, f_K^{\text{syn}}$ are generated by simulating from:

$$f_i^{\text{syn}} \sim X_i \quad i = 1, 2, \ldots, K \tag{1}$$

where X_i is a count distribution with mean f_i. Section 3 considers the best distribution to use for X_i.

The advantage of using a saturated count model is three-fold. Firstly, saturated models require no model selection - which in categorical data involves deciding which interactions to include in the model - as *all* interactions are included. This ensures all relationships are preserved in the resulting synthetic data, thus avoiding the scenario where a researcher's analysis subsequently performed on the synthetic data is more complex than - and hence unsupported by - the synthesis model (Meng, 1994).
Secondly, the time taken to undertake the synthesis, computationally, is substantially reduced because the model-fitting time is null: the model's fitted values are just the original counts.
Thirdly, synthetic counts are unbiased with expectations equal to the original counts. In turn, this gives the synthesizer an insight *a priori* (prior to synthesis) into the likely risk and utility profiles of the synthetic data. To illustrate, original counts of one are usually those at greatest risk of disclosure in a categorical data set because they relate to statistically unique individuals. Therefore, a suitable risk metric for synthetic data is $\tau_3(1)$ (Jackson et al. 2022): the probability that an original count of one is synthesized to one, which relates to a unique in the original data remaining unique in the synthetic data. Now, this unbiasedness property means that if, say, the Poisson is used for synthesis - that is, if the Poisson is chosen as X in (1) - then $\tau_3(1)$ is fixed and equal to exp(-1)=0.37; those familiar with R will recognize this as the quantity *dpois(1,1)*.
This third advantage opens up a new approach in relation to generating synthetic data sets. As synthetic data generation is typically an iterative process, involving extensive post-synthesis evaluations to establish risk and utility, gaining an insight into properties of the synthetic data *a priori* improves the efficiency of the synthesis and invites a more formal approach.

3 The use of multi-parameter count distributions

The most obvious choice of distribution for modelling categorical data is the Poisson. Besides, models often assume that individuals' observations are independent. While for data sets in microdata format this translates into assuming the rows of the data set are independent, for a contingency table it translates into assuming cell counts are Poisson distributed.
The problem with using the Poisson, though, is that each synthetic count's variance is always equal to the mean (the original count). Therefore, the variance of each synthetic count is fixed, and this uncertainty may be insufficient to mask - and hence protect - the underlying original count.
There are benefits, therefore, to using more flexible count distributions instead of the Poisson. The flexibility of the GAMLSS (Generalized Additive Models for Location, Scale and Shape) framework developed by Rigby and Stasinopoulos (2005) is particularly useful here. For more about the distributions mentioned henceforth and their parameterizations, see Rigby et al. (2019), the book written by the creators of the GAMLSS approach.

A two-parameter count distribution such as the negative binomial (NBI) provides the synthesizer with control over the scale (the variance) in addition to the location (the mean), thereby allowing more uncertainty to be applied to original counts. The metric $\tau_3(1)$, for example, then depends on σ the NBI's shape parameter. The intention is that the synthesizer treats σ as a tuning parameter in the synthesis; after all, as the model is saturated, σ could not be estimated anyway through maximum likelihood.
However, increasing the variance of the NBI through increasing σ increases the heaviness of the tails, resulting in a substantial probability point mass at zero. This produces synthetic data with an inflated number of zeros, which is exacerbated by the fact that, as saturated models are used, zero counts in the original data are not synthesized to non-zero counts.
This calls for further flexibility and motivates the use of three-parameter count distributions, which allow the synthesizer to control the shape in addition to the location and scale. One such example is the Delaporte distribution. For a given mean and variance, the shape of the Delaporte can be adjusted to reduce the heaviness of the tails, resulting in fewer zero synthetic counts as well as fewer unnecessarily large synthetic counts. This can be seen in Figure 1, which gives three Delaporte distributions, with the same means and variances but different shapes; for example, the probability of obtaining a zero is much greater in the distribution given by the red (solid) line than in the other two.
The problem in general, though, with distributions that arise through Poisson mixtures (such as the NBI and Delaporte), is that their variances are *increasing* functions of the mean, hence relatively more noise is applied to larger counts than smaller counts. However, as larger counts tend to be lower risk than smaller counts, it is preferable if the variance is a *decreasing* function of the mean, so that larger counts are perturbed less.
Rather than using a standard count distribution, an alternative it to use discretization to produce a more bespoke count distribution, by discretizing a continuous distribution defined on the interval $(0, \infty)$ - an "underused" method according to Rigby et al. (2019). A candidate for discretization is the gamma family (GAF) distribution, which has three-parameters μ, σ and ν, and where ν controls the variance-mean relationship. The mean is μ and the variance $\sigma^2\mu^\nu$; thus, when $\nu < 0$, the variance is a decreasing function of the mean, and larger counts *are* perturbed less than smaller counts - the desired behaviour. Figure 2 displays the variance-mean relationship for three GAF distributions, which is one of exponential decay, where ν controls the rate at which the variance falls away.

4 Conclusion

To briefly conclude, while saturated models are uninformative from an inferential perspective and too rigid from a predictive perspective, they have

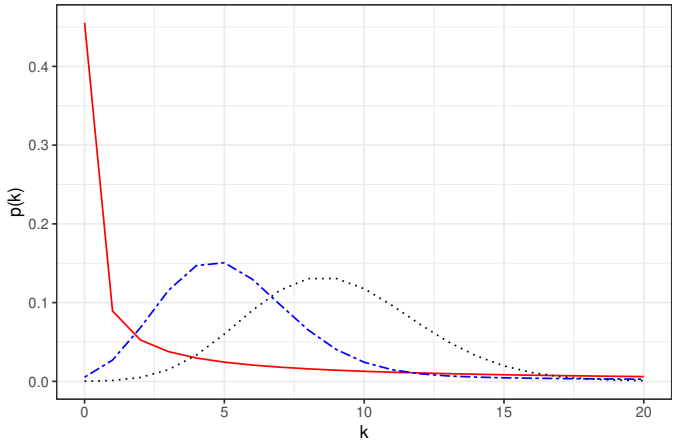


FIGURE 1. The probability mass functions of three Delaporte distributions with the same mean and variance, 10 and 510, respectively. The flexibility afforded by a three-parameter count distribution allows the shape of the distribution to be adjusted. Incidentally, the red (solid) line is an NBI distribution, which is a special case of the Delaporte.

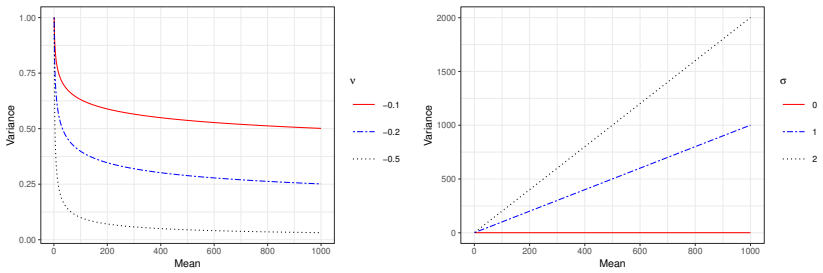


FIGURE 2. The variance-mean relationship for three GAF distributions with different ν values (with $\sigma = 1$). For comparison, the variance-mean relationship for three NBI distributions are placed alongside (with different σ values).

a practical use in data synthesis, where it suffices to obtain a noisy version of the original data. Coupled with the use of a flexible multi-parameter count distribution - for which it can be equally difficult to justify the use of in practice - saturated models allow properties of the synthetic data to be derived analytically *a priori*, thus facilitating a more efficient and transparent synthesis.

References

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Lecture Notes in Statistics. New York: Springer.

Jackson, J. E., Mitra, R., Francis, B. J., Dove, I. (2022). Using saturated count models for user-friendly synthesis of categorical data. *Forthcoming in Journal of the Royal Statistical Society: Series A (Statistics in Society)*. Preprint available at https://arxiv.org/abs/2107.08062

Little, R. J. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, **9**, 407 – 426.

Meng, X.-L. (1994). Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, **9**, 538 – 558.

Raghunathan, T.E., Reiter, J.P., Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation *Journal of Official Statistics*, **19(1)**, 1 – 16.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, **29(2)**, 181 – 188.

Rigby, R. A., Stasinopoulos, M. D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54(3)**, 507 – 554.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z. and De Bastiani, F. (2019) *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. CRC Press.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1993). Statistical Disclosure Limitation. *Journal of Official Statistics*, **9**, 461 – 468.

Robust regression and adaptive filtering

Michele Lambardi di San Miniato[1], Ruggero Bellio[1], Luca Grassetti[1], Paolo Vidoni[1]

[1] University of Udine, Italy

E-mail for correspondence: michele.lambardi@uniud.it

Abstract: Adaptive filters are local estimators from control theory that can track time-varying parameters in a compute-efficient fashion. Robust implementations are not frequent, but are necessary for local estimation. We motivate our claim based on a challenging example of sensor data stream, publicly available.

Keywords: Adaptive filtering; Spatio-temporal statistics.

1 Introduction

Sensor data analysis blends together regression with spatio-temporal statistics. Robustness concerns may arise in the case of noisy data. Computational issues will likely add up if the analyses are to be carried out repeatedly or even in real time. Online processing makes the analysis even more challenging, in this respect.
Control theory is a methodology that can help to turn such analyses into automation. In particular, here we focus on adaptive filters as a versatile tool to keep system information up-to-date even in the case of frequent regime shifts. Adaptive filtering approaches assume time-varying parameters, which can be tracked efficiently via recursive estimators (Haykin, 2014). These estimators will approximately maximize a suitable objective function, e.g., a log-likelihood, based on stochastic gradients and Hessians or their approximations.
Inspired by Cressie (1993), we present a Spatio-Temporal Conditional Auto-Regressive model (STCAR) that can be useful in sensor data analysis. An example based on open data illustrates how the model can be tracked. The filter becomes more robust when stated in terms of median regression. As a result, all the parameters are tracked robustly.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Model formulation

We propose a STCAR model that is suitable for prediction, as it involves only past observations. The style of formulation is closely related to Politis' transformative approach (2015), where some whitening transformations are devised to remove the correlation structure from the data.
Let $s = 1, \dots, S$ and $t = 1, \dots, T$ be space and time indices, respectively. Here, $(Y_{st})_{s,t}$ denotes the response process, $(\epsilon_{st})_{s,t}$ a Gaussian white noise with variance σ^2, and $(x_{st})_{s,t}$ are p-dimensional covariate vectors. The parameter vector is $\theta = (\psi^{\mathrm{T}}, \sigma)^{\mathrm{T}}$ with $\psi = (\rho, \phi, \beta^{\mathrm{T}})^{\mathrm{T}}$, $\rho, \phi \in \mathbb{R}$, $\beta \in \mathbb{R}^p$, $\sigma > 0$. Also let $w_{ss'} \in \mathbb{R}$, for $s, s' \in \mathcal{S}$, be spatial weights based on proximity, customly chosen, under the constrains $w_{ss} = 0$ and $\sum_{s'} w_{ss'} = 1$. Let $\Delta \in \{1, 2, \dots\}$ be the prediction horizon.
The STCAR model is formulated as

$$W_S[W_T\{W_V(Y_{st})\}] = \epsilon_{st}\,,$$

where W_S, W_T, W_V are whitening transformations, defined as

$$W_S(z_{st}) = z_{st} - \rho\hat{z}_{st'}\,, \quad W_T(z_{st}) = z_{st} - \phi z_{st'}\,, \quad W_V(z_{st}) = z_{st} - \beta^{\mathrm{T}} x_{st'}\,,$$

for all spatio-temporally referenced quantities z_{st}. Here, $\hat{z}_{st} = \sum_{s'} w_{ss'} z_{s't}$ and $t' = t - \Delta$.
The model can be stated alternatively as

$$Y_{st} = \mu_{st} + \epsilon_{st}\,,$$

with

$$\mu_{st} = \mu_{st}(\psi) = \rho\hat{Y}_{st'} + \phi Y_{st'} - \rho\phi\hat{Y}_{st''} + \beta^{\mathrm{T}}(x_{st'} - \rho\hat{x}_{st''} - \phi x_{st''} + \rho\phi\hat{x}_{st'''})\,.$$

This is a regression model, though nonlinear in the parameters, in a state-space form, where Y_{st} is a function of the past up to an innovation term ϵ_{st}. Such a formulation helps in both estimation and prediction.

3 Filtering

The parameter θ is assumed to be time-varying, thus denoted by θ_t at time t, and is tracked using a local mean regression estimator $\hat{\theta}_t$, ruled exchangeably by a bandwidth parameter $\delta > 1$ or by a learning rate $\lambda \in]0, 1[$, related to each other as

$$\lambda = 1/\delta\,.$$

These quantities are crucial in trading off bias with variance, as often is the case with tuning constants of semi-parametric methods. A large value of λ, close to unit, makes the filter more efficient, as it will follow more

closely the true trajectory of θ, but also leads to noise fitting and oscillatory behavior. On the contrary, a low value of λ makes the filter more stable, but only capable to estimate a coarse approximation to the true parameter trajectories. An in-between solution is required in practice.
One can track σ_t via exponential smoothing, as $\hat{\sigma}_t^2 = (1-\lambda)\,\hat{\sigma}_{t-1}^2 + \lambda \sum_s \hat{\epsilon}_{st}^2/S$, with $\hat{\epsilon}_{st} = Y_{st} - \hat{\mu}_{st}$, $\hat{\mu}_{st} = \mu_{st}(\hat{\psi}_{t-1})$. Similarly, ψ_t can be tracked via a quasi-Newton Stochastic Gradient Descent (SGD) algorithm with constant learning rate λ (Haykin, 2014), as

$$\hat{\psi}_t = \hat{\psi}_{t-1} + \lambda\, C_t \sum_s \frac{g_{st}}{S}\,,\; C_t = \left\{ (1-\lambda)\, C_{t-1}^{-1} + \lambda\, \frac{1}{S} \sum_s g_{st}\, g_{st}^{\mathrm{T}} \right\}^{-1}, \quad (1)$$

where C_t is a condition matrix, tracked recursively, and g_{st} is defined as

$$g_{st} = \left(W_T\{W_V(\hat{y}_{st'})\}\,,\, W_S\{W_V(y_{st'})\}\,,\, W_S\{W_T(x_{st'})\}^{\mathrm{T}}\right)^{\mathrm{T}} \cdot e_{st}\,, \quad (2)$$

with $e_{st} = \hat{\epsilon}_{st}/\hat{\sigma}_{t-1}^2$. The more robust median regression (Koenker, 2005) is attained by replacing e_{st} with $\bar{e}_{st}$, defined as

$$\bar{e}_{st} = \frac{\mathrm{sign}(\hat{\epsilon}_{st})}{\hat{\gamma}_{t-1}}\,. \quad (3)$$

The filter formed by (1)–(3) is also known as the sign algorithm in the literature, see Shao et al. (2010). Here, $\gamma_t > 0$ is a mean absolute deviation parameter, tracked by means of $\hat{\gamma}_t$ defined as

$$\hat{\gamma}_t = (1-\lambda)\,\hat{\gamma}_{t-1} + \lambda\, \frac{1}{S} \sum_s |\hat{\epsilon}_{st}|\,, \quad (4)$$

The formulation made up of (1)–(4) underlies a Laplace likelihood instead of a Gaussian one. The error distribution is just a working assumption that be wrong, but under suitable misspecified scenarios it may suffice to multiply $\bar{e}_{st}$ by a constant that depends on the true error distribution. In adaptive filtering, such a constant can be absorbed into the learning rate, which is thus the only hyperparameter to be tuned.
Often, in practice, it may happen that the variability in the covariates drop and the design matrix becomes near-singular. In such cases, it is not advisable to formulate C_t as in (1). It can be safer in this sense to formulate C_t in a diagonal fashion, which is also easier to invert in high-dimensional problems. In the next example, a diagonal version of C_t was deemed necessary to avoid singularity. Whenever the original version of C_t does not break down, the filter with diagonal C_t looks to be a good approximation, thus little efficiency was lost, while benefiting from increased stability.

4 Example: KETI data

The data on focus were collected by Hong et al. (2017), via sensors provided by the Korea Electronics Technology Institute (KETI), and

are available at `https://cseweb.ucsd.edu/~dehong/data/keti.html`. The data provide information on the environmental conditions of 51 rooms, scattered across four floors, in Sutardja Dai Hall, an office building at UC Berkeley. Floor plans can be found at `https://citris-uc.org/about/sutardja-dai-hall`.
The sensors provided readings asynchronously, roughly every five seconds, for one week in August 2013. We project the data so the readings are synchronous and available every second, so 1 hour provides 3600 data rows. The variables were: room temperature (in °C), relative air humidity (%), CO2 concentration (ppm), light (lux) and passive infrared motion data (made binary here). In Room 419, temperatures of 500°C and above were clearly off-scale, so we imputed them using the last available observation. These were trivial outliers that can be easily ruled out in practice based on sensor specs and common sense. Luminosity typically revolved around 250 lux, but in Room 668 it was 2000 lux, which is still legitimate, thus we retain it in the analysis. For the generic location s, its Markov neighborhood was defined as the cluster of rooms it belongs to, based on floor plans; the spatial weights $w_{ss'}$ imply simple averages of such neighborhoods.
Predicting humidity is the aim of the present analysis, which can be repeated for other variables. Both local mean and median regressions were carried out, with bandwidth δ ranging in $1, \ldots, 12$ hours. The prediction horizon is chosen to be $\Delta = 1$ hour. In Figure 1, the optimal bandwidth can be found, based on either mean absolute prediction error (L1) and root mean square error (L2). It looks optimal to set $\delta = 6$ hours for mean regression and $\delta = 3$ for median regression. Moreover, in the case of mean regression, the optimum is not sharp, but a slightly longer bandwidth may perform similarly. It seems thus that the robust filter can leverage more local information, while the non-robust filter needs a tuning more oriented to stable behavior.
In Figure 2, L2 error is reported for each location, for both regression types. The tuning constants were set equal to their optimal values, as inferred from Figure 1. To prove our point, we compare the two regression types in terms of L2 error to slightly favor mean regression, only to show that even in such a favorable case it will be outperformed by the median regression in some relevant sense. Mean regression is designed to minimize L2 error globally, but not within each location. Even so, one would expect that mean regression will outperform median regression in the L2 sense. However, it seems that most locations are better predicted in this sense by median regression. Mean regression looks driven mostly by outliers, see the lone room in the top right. Similar results can be obtained based on L1 error, with greater edge for median regression.

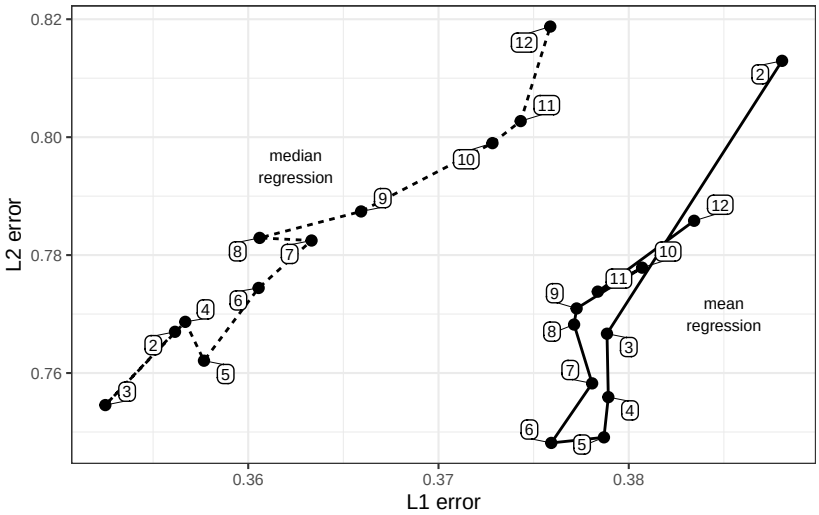


FIGURE 1. L1 and L2 errors depending on bandwidth δ (in hours, within boxes) and regression types (mean and median).

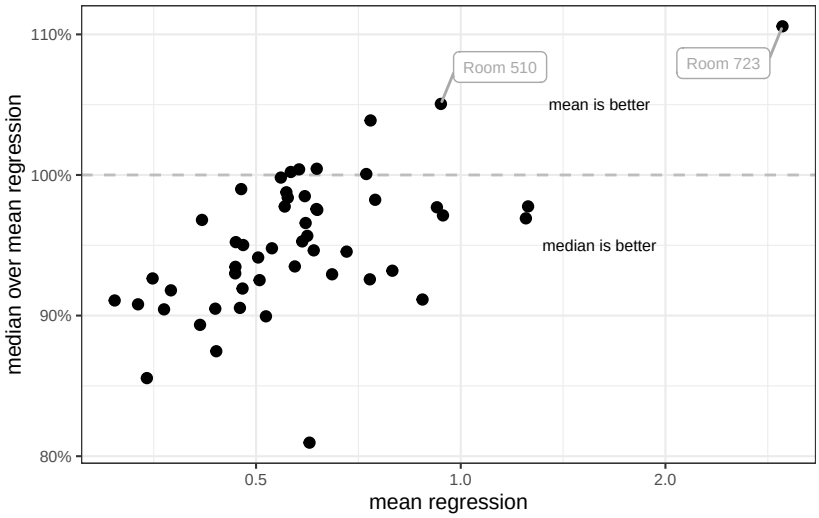


FIGURE 2. L2 error for mean regression (absolute, x-axis) and median regression (relative, y-axis). One dot for each location, logarithmic scale on the x-axis.

5 Closing remarks

Local estimators can accommodate for time-varying dynamics in an algorithmic fashion. Here, robust methods were shown to work even more

locally than their non-robust counterparts. Mean regression may demand larger bandwidths and sacrifice locality, just to smooth the outliers out. Median regression supports smaller bandwidths and can thus better approximate nonlinear systems.
Ongoing research may involve a deeper study of the theoretical properties of the proposed adaptive filter. Indeed, its properties can be deduced within control theory, but a more statistical assessment may be provided within the related kernel estimation framework. Furthermore, some other real data examples will likely provide some further insights into the behavior of the filter. A software implementation in `R` is being developed.

Acknowledgments: This work was supported by the Competence Centre ASSIC - Austrian Smart Systems Integration Research Center, cofunded by the Austrian Federal Ministry for Transport Innovation and Technology (BMVIT), the Austrian Federal Ministry for Education, Science and Research (BMWFW) and the Austrian Federal Provinces of Carinthia and Styria within the Competence Centres for Excellent Technologies Programme (COMET). The research of Michele Lambardi di San Miniato was supported by the European Social Fund (Investimenti in favore della crescita e dell'occupazione, Programma Operativo del Friuli Venezia Giulia 2014/2020) - Programma specifico 89/2019 - Sostegno alla realizzazione di dottorati e assegni di ricerca, operazione PS 89/2019 ASSEGNI DI RICERCA - UNIUD (FP1956292002, canale di finanziamento 1420_SRDAR8919).

References

Cressie, N.A.C. (1993). *Statistics for Spatial Data*. Wiley.

Haykin, S.O. (2014). *Adaptive Filter Theory*. Pearson.

Hong, D., Gu, Q., and Whitehouse, K. (2017). *High-dimensional time series clustering via cross-predictability*. In: *Proceedings of the 20th AISTATS International Conference*, PMLR, **54**, 642 – 651.

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

Politis, D.N. (2015) *Model-Free Prediction and Regression: A Transformation-Based Approach to Inference*. Springer.

R Core Team. (2022) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.r-project.org/.

Shao, T., Zheng, Y.R., and Benesty, J. (2010). An affine projection sign algorithm robust against impulsive interferences. *IEEE Signal Processing Letters*, **17**, 327–330.

Laplace approximation for penalty selection in double additive cure survival model with exogenous time-varying covariates

Philippe Lambert[1], Michaela Kreyenfeld[2]

[1] Université de Liège & Université catholique de Louvain, Belgium
[2] Hertie School Berlin, Germany

E-mail for correspondence: p.lambert@uliege.be

Abstract: Extended cure survival models enable to separate covariates that affect the probability of an event (or *long-term* survival) from those only affecting the event *timing* (or *short-term* survival). We propose to generalize the bounded cumulative hazard model to handle additive terms for time-varying (exogenous) covariates jointly impacting long- and short term survival. The selection of the penalty parameters is a challenge in that framework. A fast algorithm based on Laplace approximations in Bayesian P-spline models is proposed. The methodology is illustrated with the analysis of pension register data enabling to explore the association between time-varying women's earnings and fertility transitions in Germany.

Keywords: Cure survival model; Time-varying covariates; Additive sub-models ; P-splines; Laplace approximations; Fertility transitions.

1 The extended cure survival model

Cure survival models explicitly acknowledge that a proportion of the studied population will never experience the event of interest whatever the duration of the follow-up. Our starting point is the *promotion time* (cure) survival model, also named the *bounded cumulative hazard model* (Yakovlev and Tsodikov, 1996) and its extension in Bremhorst and Lambert (2016). Let $\mathbf{v} = (\mathbf{z}, \mathbf{x})$ and $\tilde{\mathbf{v}} = (\tilde{\mathbf{z}}, \tilde{\mathbf{x}})$ denote vectors of categorical and quantitative covariates (with $\mathbf{0}$ denoting their reference values). If $S_p(t|\mathbf{v}, \tilde{\mathbf{v}})$ is the conditional population survival function, then

$$S_p(t|\mathbf{v}, \tilde{\mathbf{v}}) = \exp\{-\vartheta(\mathbf{v})F(t|\tilde{\mathbf{v}})\}$$

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

where $t > 0$, $\vartheta(\mathbf{v}) > 0$ and where $F(t|\tilde{\mathbf{v}})$ is a cumulative distribution function such that $F(0|\tilde{\mathbf{v}}) = 0$ and $F(T|\tilde{\mathbf{v}})) = 1$ with T denoting the minimal time after which a subject can be declared cured. The corresponding population hazard is $h_p(t|\mathbf{v}, \tilde{\mathbf{v}}) = \vartheta(\mathbf{v}) f(t|\tilde{\mathbf{v}})$. The proportion of cured subjects in the sub-population defined by $\mathbf{v}$ is $\pi(\mathbf{v}) = \exp\{-\vartheta(\mathbf{v})\} > 0$. A log-linear model is taken for $\vartheta(\mathbf{v}) = \exp\{\eta_\vartheta(\mathbf{v})\}$ and a proportional hazards model is considered for $F(t|\tilde{\mathbf{v}}) = 1 - S_0(t)^{\exp(\eta_F(\tilde{\mathbf{v}}))}$ where $S_0(0) = 1$, $S_0(T) = 0$ and $\eta_\vartheta(\cdot)$, $\eta_F(\cdot)$ are (possibly non-linear) function of the covariates with identification constraint $\eta_F(\mathbf{0}) = 0$ for the latter.

1.1 Reference survival and additive sub-models

A flexible form based on P-splines (Eilers and Marx, 1996) is taken for $f_0(t) = -dS_0(t)/dt$, $f_0(t) = \exp\left(\sum_k b_k(t)\phi_k\right) / \int_0^T \exp\left(\sum_k b_k(u)\phi_k\right) du$ with $t \in (0, T)$, $\{b_k(\cdot)\}_{k=1}^K$ a large B-splines basis associated to equidistant knots on $(0, T)$ and $\boldsymbol{\phi} = (\phi_k)_{k=1}^K$ a vector of spline parameters with $\phi_{\lfloor K/2 \rfloor} = 0$ (for identification purposes). It is directly connected to the reference population hazard, $h_p(t|\mathbf{0}, \mathbf{0}) = \mathrm{e}^{\beta_0} f_0(t)$. Smoothness is forced on $f_0(\cdot)$ by penalizing changes in the spline coefficients. Additive models are suggested to describe the effects of quantitative covariates on $\eta_\vartheta(\mathbf{v})$ and $\eta_F(\tilde{\mathbf{v}})$:

$$\left(\eta_\vartheta(\mathbf{v}_i)\right)_{i=1}^n = \mathbf{Z}\boldsymbol{\beta} + \sum_{j=1}^J \mathbf{f}_j = \mathcal{X}\boldsymbol{\psi} \quad ; \quad \left(\eta_F(\tilde{\mathbf{v}}_i)\right)_{i=1}^n = \tilde{\mathbf{Z}}\boldsymbol{\gamma} + \sum_{j=1}^{\tilde{J}} \tilde{\mathbf{f}}_j = \tilde{\mathcal{X}}\tilde{\boldsymbol{\psi}}$$

where $[\mathbf{f}_j]_i = \sum_{\ell=1}^L s_{j\ell}(x_{ij})\theta_{\ell j} = [\mathbf{S}_j\boldsymbol{\theta}_j]_i$, $[\tilde{\mathbf{f}}_j]_i = \sum_{\ell=1}^L \tilde{s}_{j\ell}(\tilde{x}_{ij})\tilde{\theta}_{\ell j} = [\tilde{\mathbf{S}}_j\tilde{\boldsymbol{\theta}}_j]_i$, denote smooth additive terms quantifying the effect of quantitative covariates on long- and short-term survival, respectively, with matrices of recentered B-splines $\mathbf{S}_j$, $\tilde{\mathbf{S}}_j$, design matrices $\mathcal{X} = [\mathbf{Z}, \mathbf{S}_1 \dots \mathbf{S}_J]$, $\tilde{\mathcal{X}} = [\tilde{\mathbf{Z}}, \tilde{\mathbf{S}}_1 \dots \tilde{\mathbf{S}}_J]$, regression and spline parameter vectors $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J)$, $\tilde{\boldsymbol{\psi}} = (\boldsymbol{\gamma}, \tilde{\boldsymbol{\theta}}_1, \dots, \tilde{\boldsymbol{\theta}}_J)$.

1.2 Extension to time-varying covariates

Assume now that the covariates are exogenous and can change values over time. Our proposal is to model the hazard rate at the population level using

$$h_p(t|\mathbf{v}(t), \tilde{\mathbf{v}}(t)) = \vartheta(\mathbf{v}(t)) f(t|\tilde{\mathbf{v}}(t)) = \mathrm{e}^{\eta_\vartheta(\mathbf{v}(t)) + \eta_F(\tilde{\mathbf{v}}(t))} f_0(t) S_0(t)^{\exp(\eta_F(\tilde{\mathbf{v}}(t)) - 1}$$

yielding a (promotion time) cure survival model with time-varying covariates, shortly named the TVcure model. The associated cumulative hazard function can be obtained numerically using quadrature. In the special case where covariates are constant, we recover the expressions for the population hazard and cumulative hazard functions in Bremhorst and Lambert (2016). In the general case, the linear predictors η_ϑ and η_F not only change over units, but also potentially over time. Therefore, the associated design matrices also depend on time with $\left(\eta_\vartheta(\mathbf{v}_i(t))\right)_{i=1}^n = \mathcal{X}_t\boldsymbol{\psi}$, $\left(\eta_F(\tilde{\mathbf{v}}_i(t))\right)_{i=1}^n = \tilde{\mathcal{X}}_t\tilde{\boldsymbol{\psi}}$.

2 Inference and penalty parameter selection

Assume now that data for each unit can be reported in a regular manner over time (measured in dt units of time). Then, the data for the ith unit would be $\mathcal{D}_i = \{(d_{it}, \mathbf{v}_i(t), \tilde{\mathbf{v}}_i(t)) : t = 1, \ldots, t_i\}$ where d_{it} is the event indicator identically equal to 0 for all t, except perhaps the last value d_{it_i} equal to one if $\delta_i = 1$ when an event is observed within $(t_i - dt, t_i)$, and zero otherwise. Then, on can show that the log-likelihood contribution for unit i is $\ell_i(\boldsymbol{\phi}, \boldsymbol{\psi}, \tilde{\boldsymbol{\psi}}|\mathcal{D}_i) = -\sum_{t=1}^{t_i} \mu_{it} + d_{it_i} \log \mu_{it_i}$, with $\mu_{it} = h_p(t|\mathbf{v}_i(t), \tilde{\mathbf{v}}_i(t))\, dt$. If $\ell = \sum_{i=1}^{n} \ell_i$ denotes the log-likelihood, then the joint posterior for the model parameters, $p(\boldsymbol{\phi}, \boldsymbol{\psi}, \tilde{\boldsymbol{\psi}}, \tau_0, \boldsymbol{\tau}, \tilde{\boldsymbol{\tau}}|\mathcal{D})$, follows from Bayes's theorem, with smoothness priors for the spline parameters defining $f_0(t)$ and the additive terms in long- and short-term survival, $p(\boldsymbol{\phi}|\tau_0) \propto \exp\left(-.5\, \boldsymbol{\phi}^{\mathrm{T}}(\tau_0 \mathbf{P}_0)\, \boldsymbol{\phi}\right)$, $p(\boldsymbol{\theta}_j|\tau_j) \propto \exp\left(-.5\, \boldsymbol{\theta}_j^{\mathrm{T}}(\tau_j \mathbf{P})\, \boldsymbol{\theta}_j\right)$ and $p(\tilde{\boldsymbol{\theta}}_j|\tilde{\tau}_j) \propto \exp\left(-.5\, \tilde{\boldsymbol{\theta}}_j^{\mathrm{T}}(\tilde{\tau}_j \tilde{\mathbf{P}})\, \tilde{\boldsymbol{\theta}}_j\right)$.
The selection of the penalty parameters $\boldsymbol{\lambda} = (\tau_0, \boldsymbol{\tau}, \tilde{\boldsymbol{\tau}})$ is particularly challenging. It is based on the maximization of their marginal posterior. Starting from the following identity $p(\boldsymbol{\lambda}|\mathcal{D}) = p(\boldsymbol{\phi}, \boldsymbol{\psi}, \tilde{\boldsymbol{\psi}}, \boldsymbol{\lambda}|\mathcal{D}) / p(\boldsymbol{\phi}, \boldsymbol{\psi}, \tilde{\boldsymbol{\psi}}|\boldsymbol{\lambda}, \mathcal{D})$ with a Laplace's approximation substituted to the conditional posterior of the spline parameters in the denominator, one can approximate the marginal posterior of the penalty parameters by

$$p(\boldsymbol{\lambda}|\mathcal{D}) \dot{\propto} p(\hat{\boldsymbol{\phi}}_{\tau_0}, \hat{\boldsymbol{\psi}}_{\tau}, \hat{\tilde{\boldsymbol{\psi}}}_{\tilde{\tau}}, \boldsymbol{\lambda}|\mathcal{D}) \left|\Sigma_{\tau}^{-1}\right|^{-1/2}$$

with explicit mathematical expressions for the precision matrix Σ_{τ}^{-1}, see Lambert (2021) for a similar strategy in nonparametric double additive location-scale models. The maximization of the marginal posteriors for the penalty parameters can be made using the fixed point iteration method, while the MAP for the regression and spline parameters can be obtained using Newton-Raphson algorithms (for given penalty parameters). The procedure is iterative and fast thanks to explicit forms for first and second derivatives. It is crucial given the amount of data in the application.

3 Application

The data of interest is a random sample from German Pension registers. We focus here on cohorts of West German women that were childfree at their 20 years anniversary. The registers provide individual information on employment status and gross earnings on a monthly basis. Earnings (expressed here as a percentage of the average gross earnings in a given year) vary frequently over time, unless a woman is studying or is unemployed, in which case it is zero. The starting month of the pregnancy was calculated as the birth date of the 1st baby minus 9 months.
Analyses based on the TVcure model described in the previous section were made separately for four consecutive 5-year cohorts between 1950 and

1969, see Table 1 for descriptive statistics. The follow-up considered for each woman started at the age of 20 until at most 45 with a possible interruption at the 1st pregnancy or due to a loss of follow-up (i.e. right censoring) with, in the latter case, an uncertainty on the 'cure' (i.e. childfree) final status of the person. Thus, in this fertility context, a woman will be considered 'cured' if she doesn't have a child by age 45.
The analysis suggests that earnings have a non-significant effect on the probability of becoming a mother (see 2nd row on Figure 1), except in the 1950-54 cohort where low-income women are more likely to have a child. Although this effect tends to diminish over the years (see 3rd row in Figure 1), women with lower income tend to have their first child earlier. As an illustration, the estimated hazard for women with half the average gross earnings has been computed for the different cohorts (see the 1st row in Figure 1) with a modal value at about age 24 for the 1950-54 cohort moving fifteen years later to age 28 with the 1965-69 cohort.

TABLE 1. Summary statistics on cohort data with monthly individual follow-up.

		Person	Mother by age 45		
Cohort	n	months	Yes	No	Right-cens.
1950-54	1628	196419	1209 (74.3%)	346 (21.2%)	73 (4.5%)
1955-59	2379	315789	1667 (70.1%)	555 (23.3%)	157 (6.6%)
1960-64	3002	412876	2088 (69.5%)	650 (21.7%)	264 (8.8%)
1965-69	3356	485773	2216 (66.0%)	763 (22.8%)	377 (11.2%)

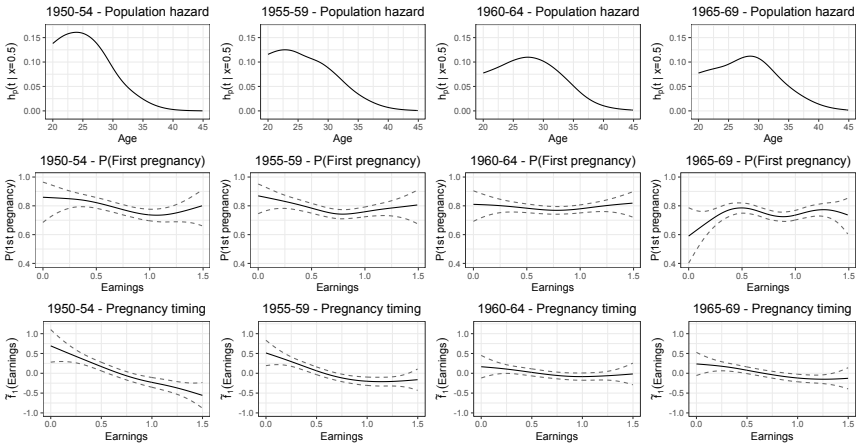


FIGURE 1. Estimated 1st pregnancy hazard $h_p(t|x = .5)$ for women with half the average gross earnings (Row 1) ; Effects of gross earnings on the probability (Row 2) and the timing (Row 3) of a first pregnancy.

Acknowledgments: The first author acknowledges the support of the ARC project IMAL (grant 20/25-107) financed by the Wallonia-Brussels Federation and granted by the Académie universitaire Louvain.

References

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

Lambert, P. (2021). Fast Bayesian inference using Laplace approximations in nonparametric double additive location-scale models with right- and interval-censored data. *Computational Statistics and Data Analysis*, **161**, 107250.

Yakovlev, A. and Tsodikov, A. (1996). *Stochastic Models for Tumor of Latency and Their Biostatistical Applications*. World Scientific Publishing, Singapore.

A modified dividing local Gaussian processes algorithm for theoretical particle physics applications

Timo Lohrmann[1], Anders Kvellestad[2], Riccardo De Bin[1]

[1] Department of Mathematics, University of Oslo, Norway
[2] Department of Physics, University of Oslo, Norway

E-mail for correspondence: timolo@math.uio.no

Abstract: We explore the use of dividing local Gaussian processes in the context of theoretical particle physics. We adapt an existing algorithm to the specific problem, trading some speed for a better precision. An intensive sensitivity analysis is performed.

Keywords: Gaussian processes; dividing local Gaussian processes; theoretical particle physics.

1 Substantive problem and approach

Physicists have proposed a wide range of new theories that extend the Standard Model of particle physics with new types of fundamental particles and new interactions. Such new theories are collectively known as Beyond-the-Standard-Model (BSM) theories. A BSM global fit refers to a large-scale parameter estimation study, in which the preferred values or ranges for the parameters of a BSM theory are determined by simultaneously comparing the theory's predictions to the results from all relevant experiments. The basis for this parameter estimation is a joint likelihood function, which is a function of the parameters of the BSM theory. The evaluation of the likelihood, however, involves many time-consuming physics calculations and simulations, which limit the scope of current BSM global fits. Our goal is to introduce an algorithm that can provide fast, per-point surrogate models for these time-consuming physics computations, or for the likelihood function directly. For simplicity, we will focus on the latter case. Since the set of relevant experimental results that enter the likelihood function is usually

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

updated between every new BSM global fit, a key requirement for the algorithm is that it is able to train the surrogates *on-the-fly* during the execution of the BSM global fit.

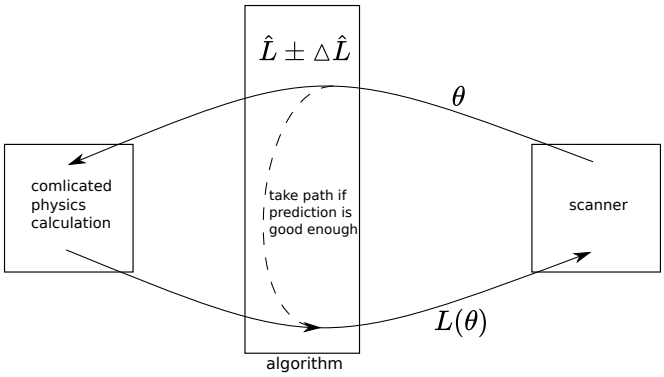


FIGURE 1. Conceptual sketch of how the algorithm fits into the computational framework of a BSM global fit.

Figure 1 provides a sketch of how the algorithm fits into the computational framework of a BSM global fit. An adaptive sampling algorithm (the "scanner") is responsible for sampling parameter points θ from the BSM parameter space. Each such point is passed to a code responsible for carrying out the expensive physics calculations that results in a single likelihood value $\mathrm{L}(\theta)$ for the given point. This likelihood is then passed back to the scanner, which uses it to determine how to pick the next θ point. Our algorithm will operate as a middle layer, that can intercept the communication between the scanner and the true likelihood computation. By learning from the continuous stream of θ and $\mathrm{L}(\theta)$ values, the algorithm can gradually learn the likelihood function in the regions of θ space explored by the scanner. Once the algorithm has obtained a good enough estimate $\hat{\mathrm{L}}(\theta)$ for $\mathrm{L}(\theta)$ in a certain region, it can short-circuit the likelihood calculation for any future θ points in this region by passing the fast estimate $\hat{\mathrm{L}}$ directly back to the scanner. However, for θ points where the estimate $\hat{\mathrm{L}}$ is still highly uncertain, the algorithm will pass the point on to the true likelihood calculation and learn from the $\mathrm{L}(\theta)$ value being passed back.

2 Data

We use data from a recent BSM global fit by the GAMBIT Collaboration (2019). Here the target likelihood is a function of four free BSM input parameters. To emulate the global fit we pass the data points through our algorithm one by one, in the order they were sampled by the differential evolution scanner used in the GAMBIT study. For comparison with earlier

work, discussed below, we also perform tests with the SARCOS dataset available from `www.gaussianprocess.org/gpml/data`.

3 Model(s)

We start from the dividing local Gaussian process (DLGP) algorithm of Lederer et al (2020). The algorithm uses the stream of input data to dynamically divide the input space into sub-spaces and train a Gaussian process (GP) on each of them. This dynamic splitting can be viewed as a growing tree structure, where the outermost nodes, called *leaves*, each contain a part of the data and a corresponding GP. Many decisions have been made to implement the algorithm: A leaf splits into two child leaves when a certain number of input points $\bar{N}$ is reached. The split is performed along the input dimension with the largest variance. A smoothing effect is created by randomly assigning some points to the sibling leaf. The probability of assigning points decays linearly with the distance from the splitting value. The predictive distribution of the tree is obtained by computing the probability of assigning the point $\mathbf{x}$ to leaf j, $\tilde{p}_j(\mathbf{x}) = \prod_{i=1}^{\nu_j} p_{\lfloor \frac{j+1}{2^i} \rfloor - 1}(\mathbf{x})$, with depth ν_j, that leads to the predictive distribution

$$p_{\mathrm{DLGP}}(f(\mathbf{x})|\mathbf{x}, X, y) = \sum_j \tilde{p}_j(\mathbf{x}) p_{\mathrm{GP}_j}(f(\mathbf{x})|\mathbf{x}, D_j),$$

where p_{GP_j} is the prediction of the GP of the j-th leaf on the data set D_j. X and y are the data set and the target variable, respectively. The predictive distribution follows a normal distribution with mean μ_* and variance σ_*^2,

$$\mu_*(\mathbf{x}) = \sum_{j \in \mathcal{I}} \tilde{p}_j(\mathbf{x}) \mu_j(\mathbf{x}) \tag{1}$$

and

$$\sigma_*^2(\mathbf{x}) = \sum_j \tilde{p}_j(\mathbf{x}) \left(\sigma_j^2(\mathbf{x}) + \mu_j(\mathbf{x})\right) - \mu_*(\mathbf{x}), \tag{2}$$

where $\mu_j(\mathbf{x})$ and $\sigma_j^2(\mathbf{x})$ are the mean and variance of leaf j.

3.1 Extension of previous work

Our work extends this base-line implementation by offering additional options for the aforementioned decisions, regulating them by hyperparameters. We offer the chance to perform the splitting along the first principal component. We include a Gaussian "decay shape". We allow varying the covariance function from the squared exponential originally used in Lederer et al (2020) to Matérn kernels. And most importantly, we update the parameters of the GPs each time a new data point is added. This provides a slower, but more precise and flexible algorithm than the original approach of fixing the GP parameters based on the first 100 data points.

4 Results

4.1 Target quantities for model selection

We aim at getting as close as possible to the truth both in terms of the expected prediction (RMSE) and in terms of variability. For the later, we compute the mean difference between predicted uncertainty and standard deviation of the input noise σ_ϵ,

$$\Delta_\sigma = \frac{1}{N}\sum_i \sigma_*(\mathbf{x}_i) - \sigma_{\epsilon,i}. \tag{3}$$

4.2 DLGP performance

As we can see in Fig. 2, the RMSE steadily decreases with the number of input points. Moreover, we observe that the covariance function is crucial to the performance of the algorithm. Over the whole range of input points, the Matérn kernel with $\nu = \frac{3}{2}$ performs equally or better than a Gaussian kernel while keeping all other parameters the same.

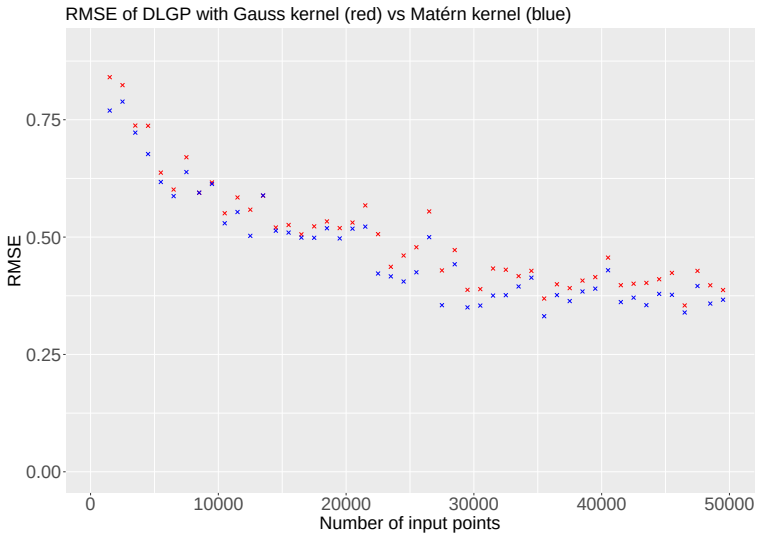


FIGURE 2. Example of the performance of our version of the DLGP on the GAMBIT data. The crosses are the RMSE aggregated over 1000 data points. The parameters of both DLGPs where identical except for the covariance function. In the case of the red crosses, a Gaussian kernel was used. Similarly, a Matérn kernel with $\nu = \frac{3}{2}$ was used for the blue crosses. We can clearly see that choosing the Matérn kernel leads to improved results. The other parameters for the DLGP are $\bar{N} = 100$, splitting along the first principal component while using the median as center. A linear overlap shape with 1% overlap was used.

In Fig. 3, we see that σ_* converges to σ_ϵ and remains in its vicinity afterwards. This shows that the DLGP has learned to estimate the variability of the training points to an adequate degree.

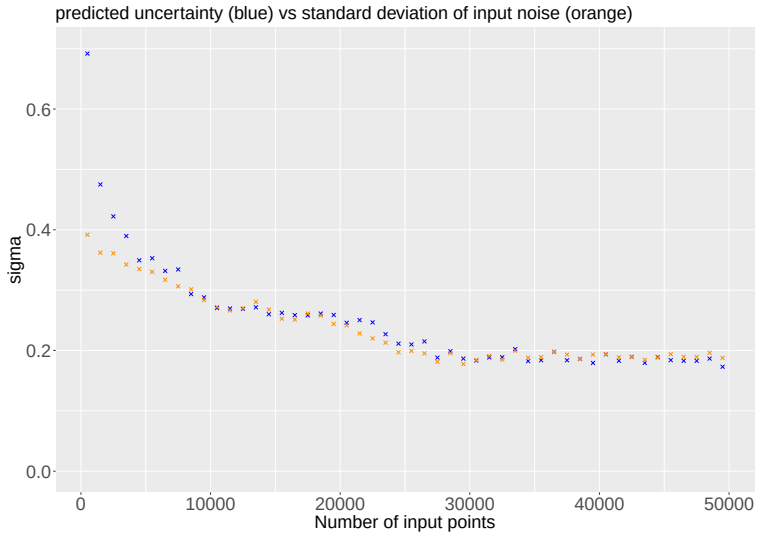


FIGURE 3. Example of the performance of our version of the DLGP on the GAMBIT data. The blue crosses are the predicted uncertainty aggregated over 1000 data points. The orange crosses are the aggregated σ_ϵ values. We observe that σ_* converges to σ_ϵ remains in its vicinity afterwards. For this DLGP, the following parameters where chosen: Matérn kernel with $\nu = \frac{3}{2}$, $\bar{N} = 100$, splitting along first principal component while using the median as center. A linear overlap shape with 1% overlap was used.

4.3 Sensitivity Analysis

A grid-search is performed to find the best combination of hyperparameters. We explore 96 different GP tree configurations and perform tests both with the data from the GAMBIT Collaboration (2019) and the SARCOS data used in Lederer et al (2020). Due to the computational expense, we use only the 50000 first data points in these initial tests. For the most promising GP tree configurations we then carry out tests using the full data sets.
When describing the performance of the tree, we refer to an improvement if a parameter improves on RMSE and Δ_σ. We observe that the choice of covariance function impacts the performance the most. The Matérn kernel function with $\nu = \frac{3}{2}$ clearly outperforms the other kernels. In agreement with the SARCOS study, we observe that for the GAMBIT data a larger maximum number of points per leaf $\bar{N}$ improves the result. Furthermore, increasing the overlap between the sibling leaves yields improvements with

regards to the RMSE, in contrast to the SARCOS results of Lederer et al (2020). However, σ_* tends to be larger and less consistent for larger overlaps as Δ_σ is significantly larger in this case. The overlapping shape has a consistently insignificant contribution to the performance. There seems to be a connection between using the median to define the center of the splitting dimension and splitting along the first principal component: Using the median instead of mean decreases the performance if all other parameters are equal, and the same is true for splitting along the first principal component instead of the dimension with maximum variance. However, the aforementioned combination performs equally well as mean and maximum variance as splitting criterion.

5 Conclusion

In this paper, an application of dividing local Gaussian processes (Lederer et al, 2020) is elaborated. In the presented theoretical physics use case, the focus lies more on high quality than fast predictions. We showed that it is worth considering improvements to the original algorithm. In particular we showed that working on the covariance function can have a large impact on the results. We believe that this online learning algorithm can also be applied in other fields where fast and continuous updating is relevant, such as finance or chemistry.

References

GAMBIT Collaboration: Athron, P., Balázs, C. et al (2019). *Combined collider constraints on neutralinos and charginos*. Eur. Phys. J. C 79, 395

Lederer, A., Conejo, A.J.O., Maier, K., et al (2020). *Real-time regression with dividing local Gaussian processes. arXiv preprint*, arXiv:2006.09446.

Carl Edward Rasmussen and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press, 2006. ISBN 0-262-18253-X

Olivier Roustant, David Ginsbourger, Yves Deville (2012). *DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization*. Journal of Statistical Software, 51(1), 1–55

Estimation and classification in semiparametric nonlinear mixed models using P-Splines and the SAEM algorithm

Maritza Márquez[1], Cristian Meza[1], Dae-Jin Lee[2], Rolando de la Cruz[3]

[1] CIMFAV-INGEMAT - Facultad de Ingeniería, Universidad de Valparaíso, Valparaíso, Chile
[2] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[3] Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Santiago, Chile

E-mail for correspondence: cristian.meza@uv.cl

Abstract: In this work, we propose an extension of a semiparametric nonlinear mixed–effects models for longitudinal data that incorporates more flexibility with penalized splines (P–splines) as smooth terms. The novelty of the proposed approach consists of the formulation of the model within the stochastic approximation version of EM (SAEM) for maximum likelihood estimation. The formulation of a P–spline as a mixed–effects model allows for the use of the the computational advantages of the existing software for the SAEM algorithm by constructing the basis functions and model matrices required in the modelling and the variance components to be estimated. We apply the proposed method to the classification of two groups of pregnant women to the risk of miscarriages. We perform the classification of these nonlinear mixed models using an adaptive importance sampling scheme. From this point of view, the proposed models improve the analysis of this type of data concerning previous studies. These improvements are reflected both in the fit of the models and in the classification of the groups.

Keywords: Nonlinear mixed models; SAEM algorithm; P-splines.

1 Semiparametric nonlinear mixed-effects model

A special semiparametric nonlinear mixed-effects (SPNLME) model proposed by Ke and Wang (2001) can be expressed as:

$$y_{ij} = \eta(\phi_i, Z_{ij}) + f(t_{ij}) + \epsilon_{ij}, \quad i = 1, \ldots, N,\ j = 1, \ldots, n_i \qquad (1)$$

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

where $y_{ij} \in \mathbb{R}$ is the j-th observation of individual i, Z_{ij} is a vector of known regressor variables including the j-th time of individual i, denoted by t_{ij}, η is a known function of Z_{ij} and individual parameters $\boldsymbol{\phi}_i$, and f is an unknown function. We assume that $\boldsymbol{\phi}_i$ can be modelled parametrically as a function of fixed effects $\boldsymbol{\beta}$, and individual random effects $\boldsymbol{\zeta}_i$, $i = 1, \ldots, n$ such as $\zeta_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. We propose to use P–splines (penalized splines with B–splines as basis functions and discrete order penalties) to model f due to its flexibility and the fact that they can be written in the form of a mixed model. Then the function f described in (1) can be written as:

$$f(t_{ij}) = \alpha_0 + \alpha_1 t_{ij} + \cdots + \alpha_S t_{ij}^S + \sum_{k=1}^{K} b_{ik} Z_k(t_{ij})^s, \tag{2}$$

where Z_k, $1 \leq k \leq K$, is an appropriate spline basis and S is the degree of the basis and b_{ik} is a random effects such as $b_{ik} \sim \mathcal{N}(\beta_{b_k}, \sigma_k^2)$ for $k = 1, \ldots, K$. The fixed effects are then $\boldsymbol{\beta} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\beta}_b)$, with $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_S)$ and $\boldsymbol{\beta}_b = (\beta_{b_1}, \ldots, \beta_{b_K})$. The vector of random effects $\boldsymbol{\psi} = (\boldsymbol{\phi}, \boldsymbol{b})$, where $\boldsymbol{\phi} = (\phi_i)_{i=1,\ldots,n}$ and $\boldsymbol{b} = (b_k)_{k=1,\ldots,K}$, follows a Multivariate Normal distribution with covariance matrix $\boldsymbol{\Gamma}\text{diag}(\boldsymbol{\Sigma}, \sigma_{b_1}^2, \ldots, \sigma_{b_K}^2)$.
We propose to obtain the maximum likelihood in the SPNLME model defined in (1)-(2) to use the stochastic approximation version of the EM (SAEM) (Delyon et al., 1999) since the random individual effects, $\boldsymbol{\Psi} = (\boldsymbol{\beta}, \boldsymbol{b})$, are treated as non-observed data. The SAEM algorithm is implemented in the `R` library `saemix` and the SPNLME model described below can be easily fitted using `R`.

Classification

In the SPNLME model defined in (1), we have $\boldsymbol{Y}_i^{(m)} \sim \mathcal{N}_{n_i}(\eta(\boldsymbol{\phi}_i, f; \boldsymbol{Z}_i), \sigma^2)$, where $f(\cdot)$ is defined in Equation 2. Let us consider an individual comes from the sub–population K_ℓ, the response vector $\boldsymbol{Y}$, taken at arbitrary times $T = (t_1, t_2, \cdots, t_n)$ has pdf $q_\ell(\boldsymbol{Y}; \theta_\ell)$, where θ_ℓ is the set of parameters associated with this distribution. If we assume that $\pi_1, \pi_2, \cdots, \pi_m$ are the prior probabilities of belonging to the sub-population, the Bayes rule to classify $\boldsymbol{Y}$ in the population K_g is

$$\log \pi_g + \log q_g(\boldsymbol{Y}; \theta_g) = \max_{\ell} \left\{ \log \pi_\ell + q_\ell(\boldsymbol{Y}; \theta_\ell) \right\}, \quad \ell = 1, 2, \cdots, m \tag{3}$$

where $\pi_\ell q_g(\boldsymbol{Y}; \theta_\ell)$ is proportional to the posterior distribution of belonging to population ℓ. Unlike the case of linear mixed-effects models, the marginal densities corresponding to SPNLME models involve integrals with no closed-form solution. We propose to use the Importance Sampling (IS) method to find the pdf ql and evaluate the classification rule described in Equation 3. The IS is based on the conditional distribution of the random

effects $p(\boldsymbol{\psi}|\boldsymbol{y};\boldsymbol{\theta})$ which is obtained using the empirically estimated mean $E(\boldsymbol{\psi}_i|\boldsymbol{y}_i;\hat{\boldsymbol{\theta}})$ and conditional variance $V(\boldsymbol{\psi}_i|\boldsymbol{y}_i;\hat{\boldsymbol{\theta}})$ of $\boldsymbol{\psi}_i$ for each subject $i = 1,\ldots,n$. These quantities can be easily obtained from the simulation-step of SAEM at convergence.

2 Application: The pregnant women dataset

The data set consists of repeated measures of β–HCG concentration levels taken over a period of two years on 173 different pregnant women divided in two groups: (i) pregnancies with a normal development that came to term without important complications (124 individuals); and (ii) a group of abnormal pregnancies with serious anomalies that ended up with the loss of the fetus (49 individuals). Measurements were recorded at different times for each woman during the first trimester of pregnancy. It is well known that the β–HCG concentration levels in the two groups follow different patterns (see Figure 1).The SPNLME considered here is

$$y_{ij} = \frac{a_i}{1+\exp\left[-(t_{ij}-b_i)/c_i\right]} + \alpha_0 + \alpha_1 t_{ij} + \sum_{d=1}^{12} b_{id} Z_{ij}^{(d)} + \epsilon_{ij} \qquad (4)$$

where y_{ij} is the j-th concentration of the β–HCG hormone of the i-th woman at time t_{ij} with $i = 1,\ldots,N$, $j = 1,\ldots,n_i$; $\epsilon_{ij} \sim \mathcal{N}(0,\sigma^2)$ for each woman i; a_i is the asymptotic level of the hormone β–HCG; b_i represents the time in which the woman reaches half the asymptotic level of the hormone; c_i is the time elapsed for the woman to reach between half and three quarters of her asymptotic hormone level; α_0 is the intercept of the model; α_1 is a linear term over time t_{ij}; $Z_{ij}^{(d)}$ are the values of the d-th regressor variables in time t_{ij} for $d = 1,\ldots,12$; $b_{id} \sim \mathcal{N}(0,\sigma_d^2)$ for all i and $d = 1,\ldots,12$.

Results

The SPNLME model with 12 nodes shows better performance than parametric NLME with one and three parameters based on for example the Visual Predictive Checks (VPC). The VPC plots allow us to graphically assess whether the model simulations can reproduce both the central tendency and the variability of the observations over time since they compare the empirical percentiles of the data (here 5th, 50th and 95th percentiles) and the theoretical percentiles of simulated data and their respective prediction intervals. We observe also better results with our model in classification rates and ROC curves (see Tables 1-2 and Figure 2). In Table 1 we summarize the global classification error for the semiparametric model (SPNLME) and the two parametric NLME model with one and three random effects, respectively denoted by $\text{NLME}_{(1)}$ and $\text{NLME}_{(3)}$. We see

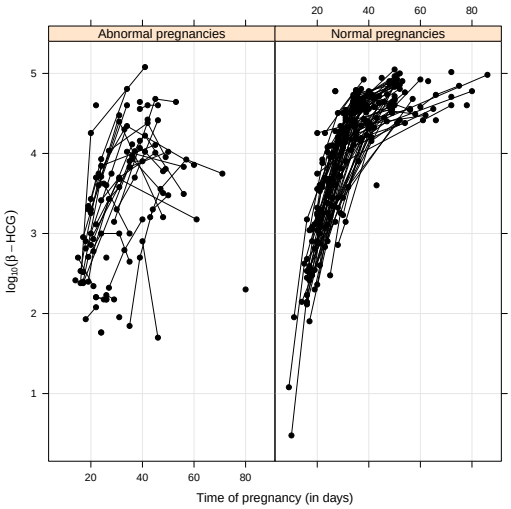


FIGURE 1. Observed profiles of $\log_{10}(\beta - HCG)$ for the abnormal group (left panel) and normal group (right panel).

that our semiparametric model improves the classification compared to the classical NLME model with one random effects proposed by Marshall and Barón (2000) since we observe a decrease of 28% in the classification error rate using Leave–one out Cross Validation (LOOCV). The decrease in the error rate is lower between our semiparametric model and the NLME model with three random effects (8% using LOOCV) but in the same way we can observe a better performance with our model. In Table 2 we compare the classification results in both groups, normal and abnormal pregnancies, for the semiparametric model (SPNLME) and the two parametric NLME models. As can be seen in Table 2 for the $\text{NLME}_{(3)}$ model, using LOOCV, 119 individuals are being well classified in the normal group and 29 individuals in the abnormal group. This classification improves when implementing the SPNLME additive model, obtaining an improvement in the classification of the normal group since 122 women are being well classified in the normal group. However, we can note that the improvement is due to better classification of the individuals of the normal group, whereas the misclassification error rate of the abnormal group remains constant. Based on this classification table and also looking at the AUC's for each model, we can see that the semiparametric additive model allows to improve the classification of pregnant women in the two groups and allows a better fit to the data (see VPC figures).

Acknowledgments: This work was funded by ANID-FONDECYT grants 1190801 and 1181662 (Chile).

TABLE 1. Error rate in the pregnancies dataset: SPNLME vs parametric NLME models using one (Marshall and Barón, 2000) and three random effects.

Error rate	$\text{NLME}_{(1)}$	$\text{NLME}_{(3)}$	SPNLME
Within sample	0.156	0.139	0.121
Leave-one out CV	0.185	0.145	0.133

TABLE 2. Classification of the $NLME_{(3)}$ and SPNLME models via Importance Sampling using the SAEM algorithm. The results are compared with the model $NLME_{(1)}$ from Marshal and Barón (2000)

Group	**$\text{NLME}_{(1)}$**		**$\text{NLME}_{(3)}$**		**SPNLME**		**Total**
	N	A	N	A	N	A	Total (173)
	Within sample						
Normal	118	6	120	4	122	2	124
Abnormal	21	28	20	29	19	30	49
	Leave-one-out CV						
Normal	113	11	119	5	121	3	124
Abnormal	21	28	20	29	20	29	49

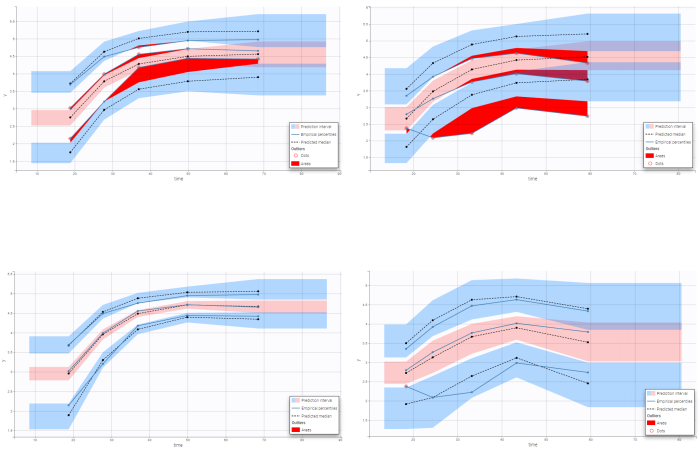
N = Normal; A = Abnormal.

References

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, 94 – 128.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *JASA*, **96**(456), 1272 – 1298.

Marshall, G. and Barón, A.E. (2000). Linear discriminant models for unbalanced longitudinal data. *Statistics in Medicine*, **19**, 1969 – 1981.



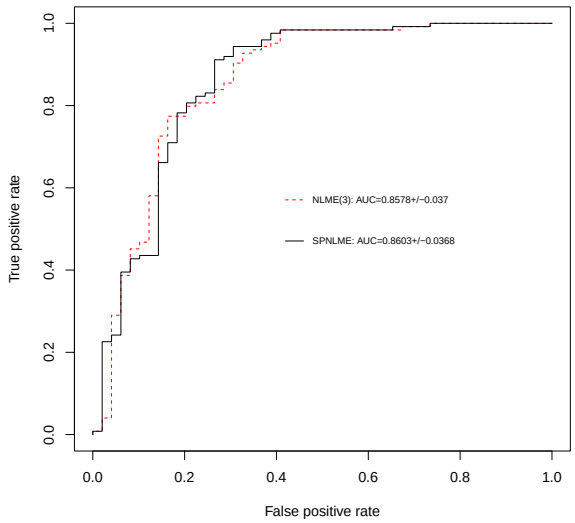


FIGURE 2. **VPC plots**. *Top*: NLME model with three random effects for pregnancies dataset. *Bottom*: SPNLME with 12 nodes. (Normal group on the left; Abnormal group on the right). **Receiver operating characteristic curves**. ROC curves for classification in the pregnancies dataset under NLME(3) (dashed line) and SPNLME (solid line), using leave-one-out CV.

Bayesian Mixtures of Envelope Models

Andrea Mascaretti[1], Antonio Canale[1]

[1] Department of Statistics, University of Padova, Padova, Italy

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

Abstract: Envelope models are linear models that assume a geometric relation between covariates and predictors to increase the efficiency of estimation, by assuming that not all predictors are influenced by covariates. In this work, we present finite mixture of envelopes, from a Bayesian perspective, to relax the hypothesis that the same subset of covariates is is not relevant in prediction across all data. We conduct a simulation study to obtain a first assessment of the impact of the choice of the number of components on model fitting.

Keywords: Envelopes; Mixtures; Bayesian methods

1 Bayesian Mixture of Envelope Models

1.1 Introduction to Envelopes

Envelope models, (Cook et al. (2010) and Cook and Zhang (2015)), are a class of models whose objective is to increase the efficiency of multivariate regression by positing a stochastic relation between responses and predictors. The framework is as follows: given a series of random variables $Y_i \in R^r$ we assume that

$$Y_i = \mu + \beta X_i + \varepsilon_i, \quad i = 1, \dots, n. \tag{1}$$

We have that $\{X_i\}_{i=1}^n$ is a sequence of non-stochastic vectors, with $X_i \in R^p$ for $i = 1, \dots, n$, the errors are independent and identically normal distributed with zero mean and covariance Σ, $\mu \in R^r$ is an unknown vector of intercepts and $\beta \in R^{(r \times p)}$ (where $R^{(a \times b)}$ denotes the space of real matrices of dimensions (a, b)) is the unknown vector of regression coefficients.
The intuition behind envelope models is that not all linear combinations of responses might be influenced by variations in the non-stochastic predictors. From a mathematical point of view, this is akin to assuming that there are two matrices Γ and Γ_0 such that $O = [\Gamma\ \Gamma_0]$ is orthogonal, that $\Gamma_0^T Y | X \sim \Gamma_0^T Y$ and $\Gamma^T Y \perp \Gamma_0^T Y | X$.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

The conditions above entail that span $(\beta) \subseteq$ span (Γ) and $\Sigma = \Sigma_1 + \Sigma_2 = P_\Gamma \Sigma P_\Gamma + Q_\Gamma \Sigma Q_\Gamma$, where $P_{(\cdot)}$ is the orthogonal projector operation on a space and $Q_{(\cdot)} = I - P_{(\cdot)}$ is the projection on the orthogonal space. In this scenario, span (Γ) is a reducing subspace of Σ (Cook et al. (2010)). The Σ-envelope of $\mathcal{B} =$ span (β), $\mathcal{E}_\Sigma(\mathcal{B})$, is the smallest reducing subspace of Σ that contains $\mathcal{B}$.
Model in Eq. (1) can be rewritten as

$$Y_i = \mu + \Gamma \eta X_i + \varepsilon,$$

where $\beta = \Gamma\eta$, $\Gamma \in R^{(r \times u)}$ is an orthogonal basis of $\mathcal{E}_\Sigma(\mathcal{B})$ and u is the dimension of the envelope $\mathcal{E}_\Sigma(\mathcal{B})$. Moreover, the variance is $\Sigma = \Sigma_1 + \Sigma_2 = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$, where $\Omega \in R^{u \times u}$ and $\Omega_0 \in R^{(r-u)\times(r-u)}$ are two diagonal matrices carrying the coordinate information with respect to the basis Γ and Γ_0.

1.2 Mixtures of Envelopes

In this work, we consider a finite mixture of envelopes. The goal is to relax the hypothesis that the combination of points with respect to which the responses are invariant be fixed across all data points.
Our proposal draws on Khare et al. (2017), whose contribution is the only Bayesian formulation of envelope models to the best of our knowledge.
Assuming to have K components, the prior distribution is defined on the parameters $\theta_k = (\mu_k, \eta_k, (\Gamma_k, \Gamma_{0,k}), \Omega_k, \Omega_{0,k})$. We constrain Ω_k and $\Omega_{0,k}$ to be diagonal matrices with entries disposed in decreasing order to preserve identifiability.
To aid in the formulation of the model, we consider a set of auxiliary variables $\{Z_i\}$. Each Z_i takes value in $\{1, \ldots, K\}$ and represent the cluster to which point i is assigned.
The likelihood is then defined as

$$Y_i | Z_i = k, x_i \sim \mathcal{N}\left(\mu_k + \Gamma_k \eta_k x_i, \Gamma_k \Omega_k \Gamma_k^T + \Gamma_{0,k} \Omega_{0,k} \Gamma_{0,k}^T\right),$$

$$Z_i \sim \mathcal{M}(\rho_1, \ldots, \rho_K).$$

For each component k, the prior is defined as follows:

1. μ_k is set to be independent from the other parameters. We endow it with a multivariate normal distribution, so that $\pi(\mu_k) = \mathcal{N}_r(\mu_0, \Sigma_0)$,
2. The conditional prior on η_k is a matrix normal:

$$\pi(\eta_k | (\Gamma_k, \Gamma_{0,k}, \Omega_k, \Omega_{0,k})) = \mathcal{N}_{(u,p)}\left(\Gamma_k^T, \Omega_k, C^{-1}\right),$$

where C^{-1} is a positive definite matrix in $R^{p \times p}$.

3. The prior on $O_k = (\Gamma_k, \Gamma_{0,k})$ is a matrix Bingham distribution with parameters G and D, where G is a positive semi-definite matrix in $R^{r\times r}$ and D has ordered positive entries. Thus, $\pi(O_k) = \mathcal{B}_{(r,r)}\left(G, D^{-1}\right)$. The density is proportional to $\exp\left\{(-1/2)\operatorname{tr}\left(D^{-1}O^TGO\right)\right\}$

4. Denoting by ω_k and $\omega_{0,k}$ the diagonal vectors of, respectively, Ω_k and $\Omega_{0,k}$, we assume that, a priori, they are distributed as order statistics of u and $r-u$ independent and identically distributed observations from Inverse-Gamma distributions of shape and rate parameters α, ψ and α_0, ψ_0.

Moreover,

$$(\rho_1, \ldots, \rho_K) \sim \mathcal{D}ir\left(\gamma_1, \ldots, \gamma_K\right).$$

As it is not possible to compute the exact posterior of the model, we rely on a Gibbs sampler to obtain estimates from a chain whose ergodic distribution is what we wish to sample. The sampler is adapted from Marin et al. (2005) and Khare et al. (2017).

TABLE 1. Parameters employed to generate the simulated dataset. Notice that for η and μ, we only reported the coefficients that premultiply matrices or vectors made up of ones for brevity. ω and ω_0 are the elements of the diagonal matrices Ω and Ω_0.

Parameter	Cluster 1	Cluster 2	Cluster 3
Γ	$(1,0,0)^T$	$(0,1,0)^T$	$(0,0,1)^T$
Γ_0	$\begin{pmatrix}0,1,0\\0,0,1\end{pmatrix}^T$	$\begin{pmatrix}1,0,0\\0,0,1\end{pmatrix}^T$	$\begin{pmatrix}1,0,0\\0,1,0\end{pmatrix}^T$
η	1.9	1.5	-6.0
μ	1.2	0.0	-2.0
ω	6.2	6.2	6.2
ω_0	$(3.2,1.4)^T$	$(3.2,1.4)^T$	$(3.2,1.4)^T$

2 Simulation Study

We tested the model on simulated datasets in order to gain a better understanding of the performance of the model. We wish to see how varying the number of components of the mixture affects the results. We set $r = 3$, $u = 1$, $p = 2$ and $n = 180$.
The dataset is made of three clusters, the values are reported in Table 1. We fit the model with $k = 2$, $k = 3$ and $k = 8$. We run three Markov chains for

TABLE 2. Posterior means of the βs fitted on the simulated data.

k	Cluster 1	Cluster 2	Cluster 3
2	$\begin{pmatrix} 2.3056, 1.6963 \\ 0.0667, 0.0492 \\ 0.0895, 0.0658 \end{pmatrix}$	$\begin{pmatrix} 0.0203, 0.0192 \\ 0.0142, 0.0135 \\ -5.983, -5.6864 \end{pmatrix}$	
3	$\begin{pmatrix} 2.3056, 1.6963 \\ 0.0667, 0.0492 \\ 0.0895, 0.0658 \end{pmatrix}$	$\begin{pmatrix} 0.0203, 0.192 \\ 0.0142, 0.0135 \\ -5.983, -5.6864 \end{pmatrix}$	$\begin{pmatrix} -0.382, -0.0464 \\ 1.4247, 1.6901 \\ -0.0859, -0.1016 \end{pmatrix}$
8	$\begin{pmatrix} -0.0425, -0.049 \\ 1.5161, 1.7249 \\ -0.0859, -0.0958 \end{pmatrix}$	$\begin{pmatrix} -0.0185, 0.0175 \\ 0.014, 0.0133 \\ -5.9804, -5.6775 \end{pmatrix}$	$\begin{pmatrix} 2.2976, 1.7204 \\ 0.0573, 0.043 \\ 0.088, 0.0661 \end{pmatrix}$

3000 iterations and consider the first 1000 to be of burn-in. Main results are reported in Table 2. When $k = 2$, we obtain a first cluster with an average number of 67 observations and a second cluster of 113 observations on average. What is interesting is the reconstruction of the βs, as the first cluster resembles the third simulated cluster and the other two are joint together. When $k = 3$, the reconstruction yields clusters that, on average, have 61, 65 and 53 members. Again, we see a specialisation of each cluster to mimic one of the three true classes. To conclude, when $k = 8$, the clusters on average have the following cardinality: 49, 0, 66, 0, 61, 1, 0, 1. We see that three main clusters are more populated than the rest and we see that those clusters do tend to resemble the original ones.

References

Cook, R.D., Bing, L., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, **60**, 927 – 960.

Cook, R.D., and Zhang, X. (2015). Foundations for Envelope Models and Methods. *Journal of the American Statistical Association*, **110**, 599 – 611.

Khare, K., Pal, S., and Su, Z. (2017). A Bayesian approach for envelope models. *The Annals of Statistics*, **45**, 196 – 222.

Marin, J-M, Mengersen, K., Robert, C. P. (2005). Bayesian Modelling and Inference on Mixtures of Distributions. *Handbook of Statistics*, **25**, 459 – 507.

Statistical Information-Criteria-Based Neural Network Input and Hidden Node Selection

Andrew McInerney[1], Kevin Burke[1]

[1] University of Limerick, Ireland

E-mail for correspondence: andrew.mcinerney@ul.ie

Abstract: Feedforward neural networks (FNNs) have many similarities to the models typically used in statistical modelling. The calculation of an associated likelihood function opens the door to information-criteria-based variable and architecture selection. A novel model selection method is proposed using the Bayesian information criterion for FNNs, wherein the optimal weights for one model are carried over to the next.

Keywords: Neural networks; model selection; variable selection.

1 Introduction

FNNs can be used for a variety of problems, and, in particular, they are useful for non-linear regression. These models consist of an input layer, which allows the covariates to enter the model; one or more hidden layers, which determine the complexity of the model; and an output layer, which provides the model's prediction. Although FNNs have similarities to the models typically used in statistical modelling, the majority of neural network research has been conducted outside of the field of statistics (Hooker and Mentch, 2021). This has resulted in a lack of statistically-based methodology, such as model and variable selection, which focus on developing parsimonious models. Instead, neural networks are viewed as 'black-box' models—with the level of complexity not of great concern (Efron, 2020). Thus, many neural networks are over-parameterised and miscalibrated (Sun et al., 2022). Employing a more statistical approach to model selection can allow for simpler, and, hence, more stable neural networks.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Feedforward Neural Network

We assume a model of the form

$$y_i = \gamma_0 + \sum_{k=1}^{q} \gamma_k \phi \left(\sum_{j=0}^{p} \omega_{jk} x_{ji} \right) + \varepsilon_i$$

where y_i is the response for the ith individual with covariate vector $x_i = (1, x_{1i}, x_{2i}, \ldots, x_{pi})^{\mathrm{T}}$, p is the number of input nodes (covariates), q is the number of hidden nodes, and ε_i is a random error that we assume to have a $N(0, \sigma^2)$ distribution. The parameters are: ω_{jk}, the weight that connects the jth input node to the kth hidden node; γ_k, the weight that connects the kth hidden node to the output node; and γ_0, the bias term associated with the output layer. The function $\phi(\cdot)$ is the activation function for the hidden layer, which is often a logistic function. Given our assumption that $\varepsilon_i \sim N(0, \sigma^2)$, we will use maximum likelihood to estimate the parameters.

3 Model Selection

Model selection in FNNs requires two decisions: the optimal set of input nodes, and the optimal number of hidden nodes. Our proposed method aims to determine both of these using a stepwise-BIC procedure. We take a 'top-down' approach, where a large model is considered initially, where all subsequent smaller models are initialised using warm starts. This ensures that all candidate models lie in a similar region of the objective function, which allows for fairer model comparison. A schematic of the model selection method is shown in Figure 1, and is described at a high level in the following paragraphs.

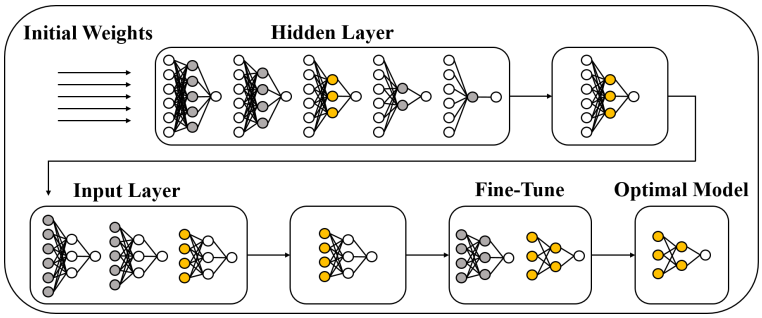


FIGURE 1. Model selection schematic. Nodes coloured grey are being considered in current step. Nodes coloured gold represent optimal nodes.

As with all model selection methods, a set of candidate models must be considered. For the input layer, we can have up to p_{max} inputs, where p_{max}

is the maximum number of covariates being considered, and, is defined by the data available. For the hidden layer, we must specify a q_{max} value, which is the maximum number of hidden nodes to be considered. This controls the complexity of the candidate models. We can then have between one and q_{max} nodes in the hidden layer.

The model selection procedure begins by determining the optimal number of hidden nodes. We start by fitting the largest possible FNN with p_{max} covariates and q_{max} hidden nodes. This model is supplied with random vectors of initial parameters, the log-likelihood function is maximised (over the different random initial vectors), and the optimal parameters are found. Given this optimal model with q_{max} hidden nodes, the importance of each hidden node is assessed by dropping each node in turn, reoptimising the log-likelihood function (using the optimal parameters from the larger model as initial values), and comparing the BIC of each model. The hidden node that is least important (i.e., the one that resulted in the lowest BIC when removed) is then dropped from the model. This process is repeated with $q_{max} - 1$ hidden nodes (where random initial vectors are again generated), and continued until all $q = 1, \ldots, q_{max}$ models have been fitted. Then, the optimal number of hidden nodes, $\hat{q}$, is determined using the BIC.

Once $\hat{q}$ has been found, model selection switches focus to the input layer. Similar to the hidden node selection, each input node is dropped in turn, the log-likelihood function is reoptimised, and the optimal parameters and associated BIC value are obtained. If the removal of a given input covariate results in a lower BIC value than the full model, that input node is dropped. This is repeated until no covariate, when removed from the model, results in a lower BIC. This yields the optimal set of covariates.

Both the hidden layer and covariate selection stages are backward elimination procedures, but with random initialisations and warm starts for stability in the neural network context. We use the phrase "optimal" loosely above, since we then follow up the previous two stages with a fine-tuning stage that looks for improved solutions in a neighbourhood of the previous "optimal". This is done by considering the addition or removal of one hidden node, then the addition or removal of one input node, and these two steps are repeated alternately until no further adjustment decreases the BIC. This fine-tuning stage is analogous to stepwise model selection with backward and forward steps.

4 Simulation

The performance of the proposed model selection method was evaluated using a simulation study. We simulated data from an FNN with three input nodes and three hidden nodes. The weights were generated so that there are three important inputs, x_1, x_2, x_3, with non-zero weights, and ten

unimportant inputs, $x_4, \ldots, x_{13}$, with zero weights. Each simulation was repeated 1,000 times, for different sample sizes ($n = 250, 500, 1000$).

The metrics calculated to evaluate the performance of the proposed method are the average number of true zero weights *correctly* dropped from the model (C), the average number of true non-zero weights *incorrectly* dropped from the model (IC), the probability of choosing the correct set of inputs (PI), the probability of choosing the correct number of hidden nodes (PH), and the probability of choosing the true model (PT). The results of the simulation study are in Table 1.

TABLE 1. Model selection metrics.

n	C(10)	IC(0)	PI	PH	PT
250	8.200	0.002	0.726	0.429	0.269
500	8.424	0.000	0.765	0.780	0.591
1000	9.679	0.000	0.958	0.948	0.936

5 Discussion

The proposed approach takes a statistical approach to neural network model selection through a likelihood function, and, hence, BIC. It also implements warms starts, which aims to keep the neural network learning algorithm in the same weight space region across all candidate models. Our results above demonstrate the favourable performance of the procedure with increasing sample size. Some other interesting findings, to be discussed in our presentation, are that the fine-tuning stage provides a non-negligible improvement and that input node (covariate) selection does not perform well when the hidden layer structure is misspecified (even if the correct hidden structure is a sub-model of the one assumed).

Acknowledgments: This work has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 18/CRT/6049.

References

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, **88**, S28 – S59.

Hooker, G. and Mentch, L. (2021). Bridging Breiman's Brook: From Algorithmic Modeling to Statistical Learning. *Observational Studies*, **7**, 107 – 125.

Sun, Y., Song, Q., and Liang, F. (2022). Learning sparse deep neural networks with a spike-and-slab prior. *Statistics & Probability Letters*, **180**, 109246.

A varying coefficient state-space model for investigating betting behaviour within in-play markets

Rouven Michels[1], Marius Ötting[1], Roland Langrock[1]

[1] Bielefeld University, Germany

E-mail for correspondence: marius.oetting@uni-bielefeld.de

Abstract: We investigate the potential effects of in-game dynamics on betting behaviour. Considering two comprehensive data sets from the 2017/18 Bundesliga season comprising in-play betting volumes and match events, we use state-space models to analyse the dynamics and drivers of betting volumes. Within this state-space framework, we use (penalised) B-splines to model the potentially time-varying effect of in-game dynamics as implied by measurable events such as shots and passes. Preliminary results suggest that volumes in the in-play market are driven by such in-game dynamics and that this effect varies over the course of a match.

Keywords: betting markets, B-splines, state-space model, time series analysis

1 Introduction

The revenue of gambling markets has considerably increased in recent years, with about 40 billion euro generated from sports betting in Europe in 2021. Sports betting takes place both in the pre-game (bets placed before kick-off) and in the in-play (bets placed during matches) market, where the latter market accounts for about 55% of the overall volume.
Despite the relatively high demand for in-play betting, empirical research to date has largely focused on the pre-game market. To investigate the corresponding drivers of bet placements in the in-game market, we consider two comprehensive data sets from the 2017/18 Bundesliga season, comprising 1) minute-by-minute betting odds and volumes and 2) information on events such as shots on goal, passes, and tackles.
The first data cover bets on the match outcome, i.e. home win or away win (we exclude bets on a draw from our analysis). These data are aggregated into intervals of one minute. The response variable in our analysis is the

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

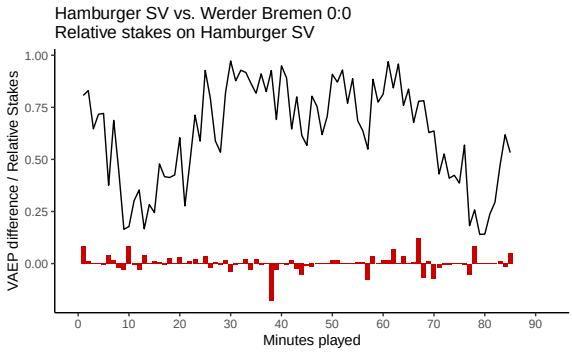


FIGURE 1. Time series of the relative stakes placed on a win of Hamburger SV for one example match from the data set (Hamburger SV vs. Werder Bremen). The vertical bars denote the differences between both teams' VAEP values from the perspective of Hamburger SV.

relative stake placed on each team, where for each interval we divide the amount of stakes placed on each team by the total amount placed on either team.

Our final data set comprises $N = 306$ time series, $\{y_{n,t}\}$, $n = 1, \ldots, 306$, with t, $t = 1, \ldots, T$, indicating the minute of the match. Towards the end of a match, much less stakes are placed, such that we truncate all time series at minute $t = 85$, resulting in a total of 26 010 observations of relative stakes.

To investigate the drivers of betting behaviour, we consider two covariates. First, we use the Elo rating (taken from `http://clubelo.com/`) as a proxy for the pre-game strength of a team. Second, to additionally investigate the effect of in-game actions on the stakes placed, we consider data on in-game events provided by the company WyScout. Using such in-game events, we consider the so-called Valuing Actions by Estimating Probabilities (VAEP) approach by Decroos et al. (2019). The VAEP measures the value of any action, e.g. a pass or a tackle, with respect to both the probability of scoring and the probability of conceding a goal. For both covariates, we consider the difference between the two teams' values in each 1-minute interval, i.e. between the Elo ratings (*elodiff*) and between the VAEP values (*vaepdiff*). Figure 1 shows an example time series of the relative stakes along with the associated *vaepdiff* values.

2 Model formulation

For the example time series shown in Figure 1, we observe a fairly high serial correlation. The correlation is induced by the market progressing through different phases, corresponding for example to extended periods of time with bets placed predominantly on the home team (cf. minutes 25–65 in Figure 1). In our modelling framework, such different phases are

captured by a latent variable within a state-space model (SSM).
For the relative stakes y_t with support $[0, 1]$, we use the beta-inflated distribution (BEINF) as an extension of the regular beta distribution to account for the fact that in some intervals stakes are placed on one team only (in which case $y_t = 0$ or $y_t = 1$). We follow the parametrisation proposed by Rigby et al. (2019), such that

$$y_t \sim \text{BEINF}(\mu_t, \sigma, p, q), \quad \text{with} \quad f(y_t) = \begin{cases} p, \text{ if } y_t = 0; \\ (1-p-q)h(y_t), \text{ if } 0 < y_t < 1; \\ q, \text{ if } y_t = 1, \end{cases}$$

for $0 \leq y_t \leq 1$ and $h(y_t)$ being the density function of the regular beta distribution.
To account for the dynamic nature of the relative stakes within matches as indicated by Figure 1, the mean μ_t is assumed to be time-varying and is modelled as follows:

$$\mu_t = \text{logit}^{-1}(\alpha_0 + g_t + \alpha \textit{elodiff}). \tag{1}$$

The unobserved variable corresponding to the market phase, g_t, is modelled as an AR(1) process, with additional covariate dependence:

$$g_t = \phi g_{t-1} + \beta \textit{vaepdiff}_{t-1} + \omega \eta_t, \tag{2}$$

with $\eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, $\omega > 0$.
Due to the dynamic nature of football matches, it may very well be the case that the effects of *elodiff* and *vaepdiff* vary over time. To account for this in our modelling framework, we replace α and β in Eqs. (1) and (2) by time-varying parameters α_t and β_t. To avoid a priori assumptions on the functional forms of α_t and β_t, we model these functions nonparametrically using B-splines. Specifically, α_t and β_t are modelled as linear combinations of a finite number of section-wise defined basis functions,

$$\alpha_t = \sum_{k=1}^{K} \nu_k^{\alpha} B_k(t), \quad \beta_t = \sum_{k=1}^{K} \nu_k^{\beta} B_k(t), \tag{3}$$

for $t = 1, \ldots, 85$, where $B_1, \ldots, B_k$, $k = 1, \ldots, K$, are fixed, equidistant B-spline basis functions of order three.
To prevent overfitting, we add a roughness penalty term, thus considering P-splines (Eilers and Marx, 1996). The resulting penalised log-likelihood function is then given as follows (cf. Langrock et al., 2017):

$$\ell_p = \log\left(\mathcal{L}_{\text{approx}}\right) - \frac{\lambda_\alpha}{2} \sum_{k=3}^{K} (\Delta^2 \nu_k^{\alpha})^2 - \frac{\lambda_\beta}{2} \sum_{k=3}^{K} (\Delta^2 \nu_k^{\beta})^2, \tag{4}$$

with the unpenalised likelihood function $\mathcal{L}_{\text{approx}}$ (see Zucchini et al., 2016), the second-order differences $\Delta^2 \nu_k = \nu_k - 2\nu_{k-1} + \nu_{k-2}$, and smoothing parameters λ_α and λ_β.

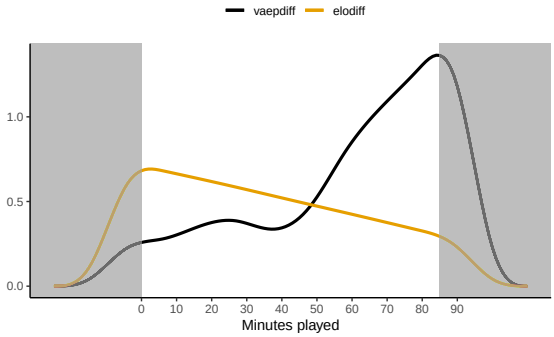


FIGURE 2. Time varying P-spline effects of the Elo variable ($\lambda_\alpha = 5000$) as well as the VAEP variable ($\lambda_\beta = 10$). The white area indicates the intervals data was considered for. The grey area indicates where outer knots are set.

3 Results

The varying coefficient SSM was estimated using $K = 10$ basis functions. The results confirm a fairly high serial correlation in the state process ($\hat{\phi} = 0.971$). For the time-varying effects, we used the AIC to select the tuning parameters λ_α and λ_β from a specified grid, with the optimal choice being $(\lambda_\alpha, \lambda_\beta) = (5000, 10)$. The estimated time-varying effects of *elodiff* and *vaepdiff* are shown in Figure 2. The effect of *elodiff* on bet placement is estimated to be positive throughout the match, but with the effect size decreasing as the match progresses. In contrast, for *vaepdiff* we find a highly non-linear functional form of the effect size over time, with the strongest effect found towards the end of the match.

References

Decroos, T., Bransen, L., Van Haaren, J., and Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1851 — 1861.

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

Langrock, R., Kneib, T., Glennie, R., and Michelot, T. (2017). Markov-switching generalized additive models. *Statistics and Computing*, **27**, 259 – 270.

Rigby, R., Stasinopoulos, M., Heller, G., and De Bastiani, F. (2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. CRC Press.

Zucchini, W., MacDonald, I.L., and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall/CRC.

A boosting model for survival analysis with dependent censoring

Alise Danielle Midtfjord[1], Riccardo De Bin[1], Arne Bang Huseby[1]

[1] Department of Mathematics, University of Oslo, Norway

E-mail for correspondence: debin@math.uio.no

Abstract: We propose a boosting model for the analysis of censored data with a dependent censoring scheme, based on the accelerated failure time model and the Clayton copula. Both in the motivating example, related to aeroplane landing, and in a classic biomedical dataset, our proposed approach provides excellent results.

Keywords: Clayton copula; Gradient boosting; Survival analysis.

1 Motivating Problem and Data Description

Motivating Problem. Assessing the airports' runway conditions is central for the aviation industry. To activate appropriate safety procedures and avoid accidents, pilots need to be informed before landing about the available friction on the runways. Rapid changes in weather conditions and the impossibility of constantly stopping the air traffic to mechanically assess the friction, drive the need for data-based approaches.

FRICTION data. To perform this task, Avinor, the largest airport operator in Norway, provided us with a large data set about 16 Norwegian airports, out of which we select the part related to Oslo Airport Gardemoen. The dataset includes weather data (wind speed, temperature, humidity, precipitation, etc.), runway reports (type of runaway contamination, use of sanding, chemicals, etc.), and flight data (acceleration, brake pressure, flap position, etc.). While weather data and runaway reports are used as covariates, flight data are used to compute the response. The runway friction is indeed computed based on the deacceleration of the airplane during the landing. Unfortunately, this information is censored, as often the pilots do

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

not need to use all the available friction. When instead they fully apply the brakes, the maximum attainable friction from the runway is recorded. In statistical words, this is the event. For the other observations, we only know that the true runway friction is larger than that actually used (right censoring). In total, we have 15154 observations (690 events) and 106 covariates (see Table 1 for the complete list).

TABLE 1. Overview of the available information.

Weather information	Runway information
Sand	Precipitation Intensity (lag 0, 1, 3, 6, 12, 24 h.)
Warm Sand	Air Temperature (lag 0, 1, 3, 6, 12, 24 h.)
Deice	Runway Temperature (lag 0, 1, 3, 6, 12, 24 h.)
Aice	Runway Temp. Trend (wrt 1, 3, 6, 12, 24 h.)
Contamination Depth	Relative Humidity (lag 0, 1, 3, 6, 12, 24 h.)
Contamination Cover	Relative Humid. Trend (wrt 1, 3, 6, 12, 24 h.)
Contamination Dry	Air Pressure (lag 0, 1, 3, 6, 12, 24 h.)
Contamination Wet	Air Pressure Trend (wrt 1, 3, 6, 12, 24 h.)
Contamination Solid	Dew Point (lag 0, 1, 3, 6, 12, 24 h.)
Contamination Loose	Horizontal Visibility (lag 0, 1, 3, 6, 12, 24 h.)
Contamination Base	Precipitation Type (lag 0, 1, 3, 6, 12, 24 h.)
	Dry Snow (+ cumulated by 1, 3, 6, 12, 24 h.)
	Wet Snow (+ cumulated by 1, 3, 6, 12, 24 h.)
	Sleet (+ cumulated by 1, 3, 6, 12, 24 h.)
	Rain Rain
	Wind Direction
	Maximum Wind Speed
	Mean Wind Speed
	Along Wind Speed
	Across Wind Speed
	Absolute Air Temperature
	Absolute Runway Temperature
	Airport Runway

Aim. Previous works only used the uncensored data to estimate the runway conditions (see, e.g., Midtfjord et al, 2021). Since 95% of the landing are censored observations, it seems reasonable to develop an approach that also includes the latter in the analysis. While currently available boosting algorithms based on classical survival models can help in handling censored observations, they cannot be directly used in this situation as they are based on the assumption of independent censoring. This is not realistic in our problem. The lower the available friction, the higher the chances of fully applying the brakes. For this reason, we develop a boosting algorithm to model censored data with dependent censoring.

2 Model

Boosting. To construct our prediction model, we start from the work of Midtfjord et al (2021), who show the excellent performance of a gradient boosting approach for this kind of problem. Gradient boosting (Friedman et al., 2000) is a forward stagewise additive procedure that builds the model by iteratively including small improvements to it. Practically, it minimises the empirical risk function (the empirical counterpart of the loss function) by fitting a function, called base learner, to the negative gradient of the loss function. The base learner can be any function that relates the covariates to the response, in our case, regression trees. See Table 2 for a schematic view of the gradient boosting algorithm.

TABLE 2. Schematic view of the gradient boosting algorithm.

Gradient boosting
1: Initialize $\hat{h}(\mathbf{x})$;
2: Update the model by, for $k = 1, \ldots, K$,
2.1: computing the negative gradient of the loss function at the current model;
2.2: obtaining the improvement $\hat{h}_k(\mathbf{x})$ by fitting the base learner on the negative gradient;
3: Aggregate the results, $\hat{h}(\mathbf{x}) = \sum_{k=0}^{K} \hat{h}_k(\mathbf{x})$.

Accelerated Failure Time model. We base our loss function on the negative log-likelihood of an Accelerated Failure Time (AFT) model (Pike, 1966). Although less popular than the Cox model, the AFT model suits well our case because it directly relates the response T to the covariates $\mathbf{X}$,

$$\log(T) = h(\mathbf{X}) + \epsilon, \tag{1}$$

where $h : R^p \to R$ captures the effect of $\mathbf{X}$ on the response, and ϵ is the error, that follows a baseline distribution Z, with mean 0 and variance σ_Z^2. Typical choices for the probability distributions of T are the Weibull, the log-normal and the log-logistic distributions, which lead the baseline function Z to be Gumbel, normal and logistic distributed, respectively.

Clayton copula. In order to take into account the dependent censoring mechanism, we take advantage of the Clayton copula (Clayton, 1978). As any copula, it allows expressing the joint distribution, in this case of the friction and censoring scheme, as a function of the marginals. The Clayton copula is among the simplest copulas, as it has only one parameter and does not require any logarithmic or exponential operation (Wang et al, 2010). Given the general definition of (an Archimedean) copula,

$$C_\theta(u, v) = \phi_\theta^{-1}\big(\phi_\theta(u) + \phi_\theta(v)\big),$$

indeed, for the Clayton copula

$$\phi_\theta(t) = \frac{t^{-\theta} - 1}{\theta} \quad \text{and} \quad \phi_\theta^{-1}(t) = (t\theta + 1)^{-\frac{1}{\theta}},$$

where the parameter θ is strictly larger than 0. The Clayton copula is particularly suitable for our problem because it exhibits greater dependence in the lower tail, but our approach works in principle with any type of copula (Gumbel, Frank, ...).

The copula-based boosting model. To implement our boosting algorithm, in addition to the regression trees that we chose as a base learner, we need to specify the loss function. Starting from (1) and including the copula, we obtain

$$loss_i = \left(1+\tfrac{1}{\theta}\right) \log\left((1-F_Z(s(t_i)|\mathbf{x}_i))^{-\theta} + (1-F_V(r(t_i)|\mathbf{x}_i))^{-\theta} - 1\right) + g(t_i, \delta_i)$$

where

$$g(t_i, \delta_i) = \begin{cases} (1+\theta)\log\left(1 - F_Z(s(t_i)|\mathbf{x}_i)\right) - \log\left(f_Z(s(t_i)|\mathbf{x}_i)\frac{1}{\sigma_Z t_i}\right) & \text{if } \delta_i = 1 \\ (1+\theta)\log\left(1 - F_V(r(t_i)|\mathbf{x}_i)\right) - \log\left(f_V(r(t_i)|\mathbf{x}_i)\frac{1}{\sigma_V t_i}\right) & \text{if } \delta_i = 0, \end{cases}$$

$(t_i, \delta_i, \mathbf{x}_i)$ is the i-th observation, made of the measurement of the friction t_i, the indication of the use of the anti-skid system δ_i, and the values of the covariates, $\mathbf{x}_i$; $\theta > 0$ is the parameter of the Clayton copula; $s(t_i) = (\log t_i - \hat{h}(\mathbf{x}_i))/\sigma_Z$; $r(t_i) = (\log t_i - \hat{h}(\mathbf{x}_i))/\sigma_V$; F and f are the CDF and the PDF of the random variable shown in the subscript; Z and V come from (1), for the friction and the censoring, respectively; and, finally, σ_Z^2 and σ_V^2 are the related variances.

3 Results

The model is evaluated in terms of calibration (calibration plot, see Figure 1 and 2) and discrimination ability (C-index, see Table 3). We also provide the mean absolute error for the uncensored responses (MAE, see Table 3).

TABLE 3. Performance of the proposed model (Clayton-boost) on the two datasets, contrasted to: boosting without copula (Std-boost), standard AFT model (Std-AFT), and Cox model (Cox).

	Clayton-boost		Std-boost		Std-AFT		Cox	
Dataset	C-ind	MAE	C-ind	MAE	C-ind	MAE	C-ind	MAE
FRICTION	0.799	**0.047**	**0.838**	0.356	0.835	1.148	-	-
GBSG	**0.708**	**749**	0.685	1300	0.678	1323	0.679	875

FRICTION data. As can be seen in Table 3 (first row) and Figure 1, the proposed method works very well, clearly outperforming the competitors in terms of calibration. While it struggles to correctly rank the response (worse C-index), it performs very well in minimizing the absolute error for the uncensored observations. Since for our specific problem this is the most important point (we need to provide the pilot with the best approximation of the friction in the dangerous cases in which (s)he needs to fully use it), these results are encouraging.

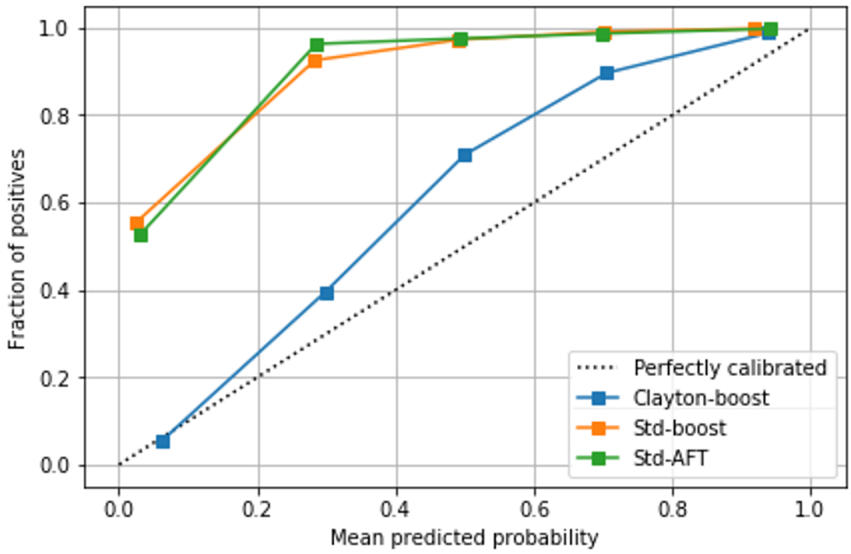


FIGURE 1. Calibration plots for the FRICTION data.

GSBG cancer data. We also applied our method to a classic biomedical example, namely the German Breast Cancer Study Group dataset analysed by Sauerdrei et al. (1999). Here the response is the recurrence-free survival (in days), while the covariates are clinical information on the patient. The sample size is 686, with 56% of censored observations (299 events). The results are shown in Table 3 (second row) and Figure 1, and confirm the good performance of our model, in this case also in terms of C-index, compared to the competitors.

References

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141 – 151.

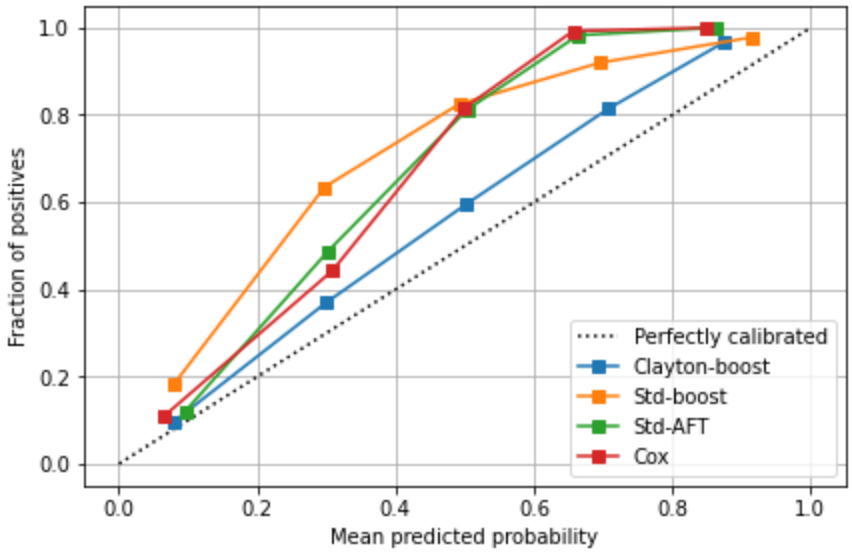


FIGURE 2. Calibration plots for the GSBG cancer data.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, **28**, 337 – 374.

Midtfjord A.D., De Bin R., and Huseby A.B. (2021). A machine learning approach to safer airplane landings: Predicting runway conditions using weather and flight data. *arXiv preprint*, arXiv:2107.04010.

Pike, M.C. (1966). A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, **22**, 142 – 161

Sauerbrei, W., Royston, P., Bojar, H., Schmoor, C., and Schumacher, M. (1999). Modelling the effects of standard prognostic factors in node-positive breast cancer. *British Journal of Cancer*, **79**, 1752 – 1760.

Wang, L., Zeng, J., Hong, Y., and Guo, X. (2010). Copula estimation of distribution algorithm sampling from Clayton copula. *Journal of Computational Information Systems*, **6**, 2431 – 2440.

Detecting statistical interactions in immune receptor datasets

T. Minotto[1], I. Hobæk Haff[1], G. K. Sandve[2]

[1] Department of Mathematics, University of Oslo, Norway
[2] Department of Informatics, University of Oslo, Norway

E-mail for correspondence: thomamin@math.uio.no

Abstract: Recent progress in the understanding of immune receptors suggests that complex interactions between amino acids are important in determining binding to antigens. Yet, current methods focus mostly on constructing good predictors for the data with complex models, and less on the understanding of the underlying interaction effects. Here, we attempt to retrieve high-order statistical interactions in immune receptor data with different methods. We study performance at this task in a large simulation study, and how it depends on the order, amount and strength of the interactions, witness rate and sample size. The results show that pairwise interactions are easily retrieved, but model complexity harms detection. Interactions are better detected in larger samples, but the process is then slower.

Keywords: Main effects; Interactions; Modelling; Detection; Immune receptors.

1 Introduction

Immune receptors are a key defense of the body from pathogens, but modeling them or predicting to which entity they will bind to is difficult, due to their complex 3D structures. Indeed, an immune receptor is made of chains of amino acids, small molecules whose combination can take multiple forms in the 3D space, allowing high variability in the recognition of antigens. Recent progress in the prediction of binding has been made with complex machine learning techniques such as convolutional neural networks or recurrent neural networks , and these have performed better than more simple models. This supports the idea that high-order interactions between amino acids are at stake.

In a dataset with p covariates, the number of possible interactions of order

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

k is $\binom{p}{k}$, which increases as $O(p^k)$ with the interaction order. Thus, techniques explore the interaction space in a non-exhaustive way. These include hierNet, glinternet, Monte-Carlo logic regression, logic feature selection, Bayesian logic regression, iterative random forests and neural interaction detection. We propose to apply these methods to the immune receptor problem of explaining binding based on the sequence of amino acids, which has not been done before, to study how they work in this context.

2 Material and methods

2.1 Settings

We study the binding problem, which consists in predicting $p_{\boldsymbol{x}} = \mathbb{P}(Y = 1|\boldsymbol{x})$, where $\boldsymbol{x}$ is an amino acid sequence, typically of length 15, whose components are categorical with 20 categories, and Y is a binary response variable indicating whether the sequence binds to an antigen or not ($Y =$ 1 or 0).

Statistical interactions can be defined as non-additivity of the effect of covariates on the response. In our context, we assume that the probability of binding $p_{\boldsymbol{x}}$ can be expressed as $p_{\boldsymbol{x}} = g(F(\boldsymbol{x}))$, with g the logit function, and $F(\boldsymbol{x})$ an additive function that describes the dependence on the covariates. Then, the function $F(\boldsymbol{x})$ shows no interaction of order k between covariates $(x_{i_j})_{j\in[1,k]}$ if it can be expressed as the sum of k functions, $(f_{\setminus i_j})_{j\in[1,k]}$, where each $f_{\setminus i_j}$ does not depend on x_{i_j}:

$$F(\boldsymbol{x}) = \sum_{j=1}^{k} f_{\setminus i_j}(x_1, ..., x_{i_j-1}, x_{i_j+1}, ..., x_p).$$

2.2 Methods for detecting interactions

At least two lasso-based methods have been proposed in the literature to retrieve pairwise interactions. The first one, hierNet (Bien et al., 2013), builds on the all-pairs lasso, that includes both main effects and interactions in the model. Interactions are present as a product of covariates, and constraints are added to respect a hierarchy: interactions are included in the model only if one or both of its variables are marginally important.
The second method, glinternet (Lim and Hastie, 2015), is based on a group lasso. The goal is also to fit a model with main effects and interactions, but coefficients are penalised by groups. This allows to respect the hierarchy constraint, and at the same time speed up the search for relevant covariates.

Other methods have been proposed to retrieve higher-order interactions. Some build on the logic regression model, that is a generalised linear model

where covariates are Boolean combinations of the original covariates, called logic trees. Simulated annealing is used to explore the space of trees and find best solutions. Based on this, Kooperberg and Ruczinski (2005) use reversible jump Markov chain Monte Carlo to generate several logic regression models from a first fit and develop Monte-Carlo logic regression (MCLR). Another method, logic feature selection (logicFS) (Schwender and Ickstadt, 2008), creates models for different bootstrap samples of the data. A last one, Bayesian logic regression (BLR) (Hubin et al., 2020), uses a genetic algorithm to generate different trees, and a mode jumping MCMC to switch from one model to the other. By studying which covariates appear together in the trees, it is possible to retrieve information about interactions.

Finally, two methods fit a complex prediction model to the data, then retrieve interactions by different metrics. Iterative random forests fits a random forest and studies how frequently covariates appear together in the trees (Basu et al., 2018). Neural interaction detection fits a neural network and computes an aggregation measure of the weights telling which interactions have the most influence on the outcome (Tsang et al., 2018).

3 Simulation study

Immune sequences are generated with the OLGA software (Sethna et al., 2019), and only sequences of length 15 are kept. We implant interactions at a given rate in the sequences as motifs: a group of amino acids at different positions takes a given value, for instance amino acids at positions 1, 2, 3, are taken to be A, A, A.
Then, we use 2 different types of model in the study. In a discriminative model, a logistic regression is created where each binary covariates is attributed a coefficient (main effects), and some products of covariates are also included in the model (interactions), i.e.

$$\log\left(\frac{p_{\boldsymbol{x}}}{1-p_{\boldsymbol{x}}}\right) = \alpha\left(\tilde{\beta}_0 + \sum_{j\in S_m}\tilde{\beta}_j x_j + \sum_{i\in S_I}\tilde{\beta}_i \prod_{l\in S_{I_i}} x_l + \sum_{n\in S_n}\tilde{\beta}_n x_n\right) + \epsilon\,,$$

$\alpha, \tilde{\beta}$s coefficients, S_m and S_n main effects, S_I interactions, $\epsilon \sim \mathcal{N}(0, 0.01)$.

In a generative model, we either implant interactions in the binder sequences, or generate many sequences and keep the ones that arrive with the interactions in the binder class until we have enough of them. The latter procedure takes much more time so we mainly use the former one. In both models we tune model parameters to have a few mislabelled sequences (around 2.5% in each class). Selected methods are then applied to the final dataset $(\boldsymbol{X}, \boldsymbol{Y})$ to attempt to retrieve interactions.

We vary the interaction order (from 2 to 4), implantation (witness) rate (in 50%, 20%, 10% and 5% of sequences), number of sequences in a dataset (10^3, 10^4, 10^5). In the discriminative model we also vary the strength of the interaction with respect to the main effects (1, 2, 4 or 8), and the hierarchy between interactions and main effects (strong, weak or no hierarchy).

4 Results

As expected, pairwise interactions are much easier to retrieve than higher-order ones, and for higher order interactions, the methods often find sub-interactions of order 2 instead of the correct ones. Detection is better when more sequences are present in the dataset, but this increases detection time and the gain is not that much. For low implantation rates the performance decreases, but the highest implantation rate (motif in half of the sequences) does not always lead to the best detection. For the discriminative model, ability to separate main effects from interactions depends on coefficient values, and is best for a high interaction coefficient. When the interaction is not selected first, it often appears in the top 10. Interacting covariates are also frequently retrieved as separate main effects. The hierNet and neural interaction detection methods have shown robust performance in a wide range of settings, for reasonable computation times, while Monte-Carlo logic regression and BLR have performed more poorly, and despite BLR having the largest running time. The other methods had good performance when the interaction was easy to retrieve.

References

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, **115**, 1943 – 1948.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, **41**, 1111 – 1141.

Hubin, A., Storvik, G., and Frommlet, F. (2020). A Novel Algorithmic Approach to Bayesian Logic Regression (with Discussion). *Bayesian Analysis*, **15**, 263 – 333.

Kooperberg, C. and Ruczinski, I. (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genetic Epidemiology*, **28**, 157 – 170.

Lim, M. and Hastie, T. (2015). Learning Interactions via Hierarchical Group-Lasso Regularization. *Journal of Computational and Graphical Statistics*, **24**, 627 – 654.

Schwender, H. and Ickstadt, K. (2008). Identification of SNP interactions using logic regression. *Biostatistics*, **9**, 187 – 198.

Sethna, Z., Elhanati, Y., Callan Jr., C. G., Walczak, A. M., and Mora, T. (2019). OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, **35**, 2974 – 2981.

Tsang, M., Cheng, D., and Liu, Y. (2018). Detecting Statistical Interactions from Neural Network Weights. In: *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada

Estimated Covid-19 burden in Spain: ARCH underreported non-stationary time series

David Moriña[1], Amanda Fernández-Fontelo[2], Alejandra Cabaña[2], Argimiro Arratia[3], Pedro Puig[2]

[1] Universitat de Barcelona, Barcelona, Spain
[2] Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain
[3] Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail for correspondence: dmorina@ub.edu

Abstract: The problem of dealing with misreported data is very common in a wide range of contexts. The current situation caused by the Covid-19 worldwide pandemic is a clear example, where the data provided by official sources were not always reliable due to data collection issues and to the large proportion of asymptomatic cases. In this work, we explore the performance of Bayesian Synthetic Likelihood to estimate the parameters of a model capable of dealing with misreported information and to reconstruct the most likely evolution of the phenomenon.

Keywords: under-reported data; ARCH models; infectious diseases; Covid-19; Bayesian synthetic likelihood.

1 Introduction

The Covid-19 pandemic that is hitting the world since late 2019 has made evident that having quality data is essential in the decision making chain, especially in epidemiology but also in many other fields. Many methodological efforts have been made to deal with misreported Covid-19 data, following ideas introduced in the literature since the late nineties. As a large proportion of the cases run asymptomatically (Oran and Topol (2020)) and mild symptoms could have been easily confused with those of similar diseases at the beginning of the pandemic, its reasonable to expect that Covid-19 incidence has been notably underreported. Very recently several approaches based on discrete time series have been proposed (see Fernández-Fontelo et al. (2020)) although there is a lack of continuous time

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

series models capable of dealing with misreporting, a characteristic of the Covid-19 data and typically present in infectious diseases modeling. In this sense, a new model capable of dealing with temporal structures using a different approach is presented by Moriña et al. (2020). A typical limitation of these kinds of models is the computational effort needed in order to properly estimate the parameters. Synthetic likelihood is a recent and very powerful alternative for parameter estimation in a simulation based schema when the likelihood is intractable and, conversely, the generation of new observations given the values of the parameters is feasible. The method was introduced in Wood (2010) and placed into a Bayesian framework in Price et al. (2018), showing that it could be scaled to high dimensional problems and can be adapted in an easier way than other alternatives like approximate Bayesian computation (ABC).

2 Methods

AutoRegressive Conditional Heteroskedasticity (ARCH) models are a well-known approach to fitting time series data where the variance error is believed to be serially correlated. Consider an unobservable process X_t following an AutoRegressive ($AR(1)$) model with ARCH(1) errors structure, defined by

$$X_t = \phi_0 + \phi_1 \cdot X_{t-1} + Z_t,$$

where $Z_t^2 = \alpha_0 + \alpha_1 \cdot Z_{t-1}^2 + \epsilon_t$, being $\epsilon_t \sim N(\mu_\epsilon(t), \sigma^2)$. The process X_t represents the actual Covid-19 incidence. In our setting, this process X_t cannot be directly observed, and all we can see is a part of it, expressed as

$$Y_t = \begin{cases} X_t \text{ with probability } 1-\omega \\ q \cdot X_t \text{ with probability } \omega, \end{cases} \tag{1}$$

where q is the overall intensity of misreporting (if $0 < q < 1$ the observed process Y_t would be underreported while if $q > 1$ the observed process Y_t would be overreported) and ω can be interpreted as the overall frequency of misreporting (proportion of misreported observations). To model consistently the spread of the disease, the expectation of the innovations ϵ_t is linked to a simplified version of the well-known compartimental Susceptible-Infected-Recovered (SIR) model. At any time $t \in \mathbb{R}$ there are three kinds of individuals: Healthy individuals susceptible to be infected ($S(t)$), infected individuals who are transmitting the disease at a certain speed ($I(t)$) and individuals who have suffered the disease, recovered and cannot be infected again ($R(t)$). As shown by Fernández-Fontelo et al. (2020), the number of affected individuals at time t, $A(t) = I(t) + R(t)$ can be approximated by

$$A(t) = \frac{M^*(\beta_0, \beta_1, \beta_2, t) A_0 e^{kt}}{M^*(\beta_0, \beta_1, \beta_2, t) + A_0(e^{kt} - 1)}, \tag{2}$$

where $M^*(\beta_0, \beta_1, \beta_2, t) = \beta_0 + \beta_1 \cdot C_1(t) + \beta_2 \cdot C_2(t)$, being $C_1(t)$ and $C_2(t)$ dummy variables indicating if time t corresponds to a period where a mandatory confinment was implemented by the government and if the number of people with at least one dose of a Covid-19 vaccine in Spain was over 50% respectively. At any time t the condition $S(t) + I(t) + R(t) = N$ is fulfilled. The expression (2) allow us to incorporate the behaviour of the epidemics in a realistic way, defining $\mu_\epsilon(t) = A(t) - A(t-1)$, the new affected cases produced at time t.
The Bayesian Synthetic Likelihood (BSL) simulations are based on the described and the chosen summary statistics are the mean, standard deviation and the three first coefficients of autocorrelation of the observed process. Parameter estimation was carried out by means of the *BSL* (An et al. (2019)) package for R. Taking into account the posterior distribution of the estimated parameters, the most likely unobserved process is reconstructed, resulting in a probability distribution at each time point. The prior of each parameter is set to a uniform on the corresponding feasible region of the parameter space and zero elsewhere.

3 Results

This work focuses on the weekly Covid-19 incidence registered in Spain in the period (2020/02/23-2022/02/27). It can be seen in Figure 1 that the registered data (turquoise) reflect only a fraction of the actual incidence (red). The grey area corresponds to 95% probability of the posterior distribution of the weekly number of new cases (the lower and upper limits of this area represent the percentile 2.5% and 97.5% respectively), and the dotted red line corresponds to its median.
In the considered period, the official sources reported 11,056,797 Covid-19 cases in Spain, while the model estimates a total of 25,283,406 cases (only 43.73% of actual cases were reported). This work also revealed that while the frequency of underreporting is extremely high for all regions (values of $\hat{\omega}$ over 0.90 in all cases), the intensity of this underreporting is not uniform across the considered regions: Aragón is the CCAA with highest underreporting intensity ($\hat{q} = 0.05$) while Extremadura is the region where the estimated values are closest to the number of reported cases ($\hat{q} = 0.50$). Detailed underreported parameter estimates for each region can be found in Table 1. Although the main impact of the vaccination programmes can be seen in mortality data, the results of this work also showed a significant decrease in the weekly number of cases as well in all CCAA except Aragón. Figure 2 represents the estimated and registered processes globally for Spain.

Acknowledgments: Investigation funded by Fundación MAPFRE.

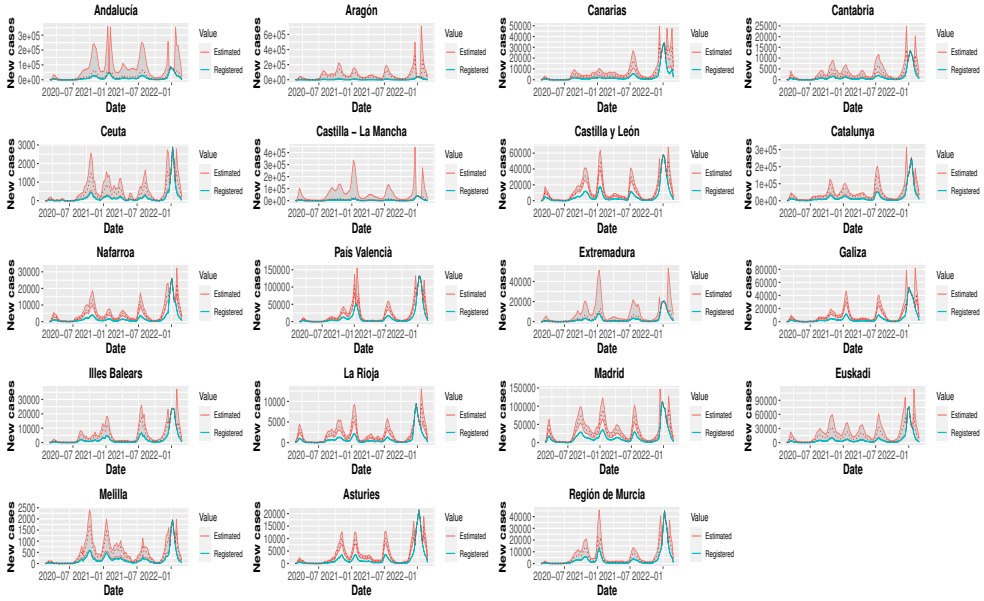


FIGURE 1. Registered and estimated weekly new Covid-19 cases in each Spanish region.

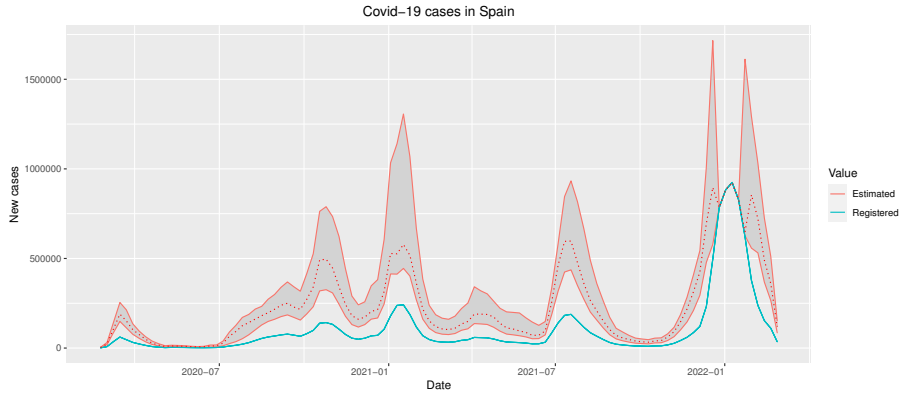


FIGURE 2. Registered and estimated weekly new Covid-19 cases globally in Spain.

References

An, Z., South, L.F. and Drovandi, C. (2019). BSL: An R Package for Efficient Parameter Estimation for Simulation-Based Models via Bayesian Synthetic Likelihood. *arXiv preprint*.

Fernández-Fontelo, A., Moriña, D., Cabaña, A., Arratia, A. and Puig P. (2020). Estimating the real burden of disease under a pandemic situation: The SARS-CoV2 case. *PLoS ONE*, **15**, e0242956.

Moriña, D., Fernández-Fontelo, A., Cabaña, A. and Puig P. (2021) New statistical model for misreported data with application to current public health challenges. *Scientific Reports*, **11**, 23321.

Oran, D.P. and Topol, E.J. (2020). Prevalence of asymptomatic SARS-CoV-2 infection. *Annals of Internal Medicine*, **173(5)**, 362 – 367.

Price, L.F., Drovandi, C.C., Lee, A. and Nott D.J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, **27(1)**, 1 – 11.

Wood S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, **466(7310)**, 1102 – 1104.

TABLE 1. Estimated underreported frequency and intensity for each Spanish region. CI stands for Credible Interval.

Region	Parameter	Estimate (95% CI)
Andalucía	$\hat{\omega}$	0.96 (0.89 - 0.98)
	$\hat{q}$	0.45 (0.41 - 0.51)
Aragón	$\hat{\omega}$	0.97 (0.97 - 0.98)
	$\hat{q}$	0.05 (0.05 - 0.29)
Principado de Asturias	$\hat{\omega}$	0.98 (0.97 - 0.99)
	$\hat{q}$	0.35 (0.33 - 0.37)
Cantabria	$\hat{\omega}$	0.97 (0.95 - 0.99)
	$\hat{q}$	0.31 (0.28 - 0.35)
Castilla y León	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.38 (0.34 - 0.40)
Castilla - La Mancha	$\hat{\omega}$	0.96 (0.93 - 0.99)
	$\hat{q}$	0.36 (0.30 - 0.39)
Canarias	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.32 (0.29 - 0.36)
Catalunya	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.35 (0.33 - 0.39)
Ceuta	$\hat{\omega}$	0.97 (0.94 - 0.99)
	$\hat{q}$	0.30 (0.27 - 0.35)
Extremadura	$\hat{\omega}$	0.90 (0.49 - 0.97)
	$\hat{q}$	0.50 (0.39 - 0.70)
Galiza	$\hat{\omega}$	0.98 (0.97 - 0.99)
	$\hat{q}$	0.33 (0.30 - 0.35)
Illes Balears	$\hat{\omega}$	0.96 (0.88 - 0.98)
	$\hat{q}$	0.39 (0.35 - 0.61)
Región de Murcia	$\hat{\omega}$	0.97 (0.92 - 0.99)
	$\hat{q}$	0.43 (0.38 - 0.48)
Madrid	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.40 (0.36 - 0.42)
Melilla	$\hat{\omega}$	0.97 (0.95 - 0.99)
	$\hat{q}$	0.35 (0.32 - 0.38)
Comunidad Foral de Navarra	$\hat{\omega}$	0.98 (0.97 - 0.99)
	$\hat{q}$	0.31 (0.29 - 0.34)
Euskadi	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.30 (0.28 - 0.35)
La Rioja	$\hat{\omega}$	0.98 (0.96 - 0.99)
	$\hat{q}$	0.32 (0.29 - 0.35)
País Valencià	$\hat{\omega}$	0.99 (0.97 - 0.99)
	$\hat{q}$	0.38 (0.37 - 0.41)

Movement Pattern Discovery With Applications In Elite Soccer

Pouyan Nejadi[1], Dr. Davood Roshan[12], Prof. John Newell[12]

[1] School of Mathematical and Statistical Sciences, National University of Ireland, Galway
[2] CURAM, SFI Research Centre for Medical Devices, National University of Ireland, Galway

E-mail for correspondence: p.nejadi1@nuigalway.ie

Abstract: In group sports, motion tracking spatiotemporal position data is typically used to learn about the tactical strategies (i.e. formation, players' role) in a game, where all players' movement data can be recorded via installed motion capture sensors around a playing field. In this paper, we propose using such motion tracking data to find each player's most typical movement patterns for designing personalised drills. Such personalised drills can be used as rehabilitation programs for injured players.

Keywords: Movement Pattern; angular distribution; pattern finding, Von mises, Football, Soccer.

1 Introduction

In recent years, the advances in sensor technologies allowed enormous data to be collected within the sports industry. This is particularly true for group sports (e.g. football, basketball, American football), where all players' movement data can be recorded via installed motion capture sensors around a playing field. To date, the primary use of such data is for tactical decision making. Player heatmaps are a popular approach. Such plots summarise a player's movement by highlighting regions on a pitch that the player occupied. The aim of this paper, though, is to use motion tracking data in a new and novel way where the emphasis could be more on players' welfare rather than tactical decision makings. To this end, a new statistical method was developed to identify player-specific movement patterns that could be further used in arranging more effective and personalised training

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

sessions. Such personalised sessions will not only prepare a player for the physiological demands in the position they play, but also make them ready for games after an injury.
The approach taken is to fit a Bivariate Generalised Linear Model (GLM) model based on the joint angular-linear distribution proposed by Johnson R. A. and Wehrly T. E. (1978) to the sets of trajectories. So, in this probabilistic framework, each trajectory can be summarised as parameters of the Bivariate GLM.

2 Methodology

Suppose a trajectory T is recorded as a sequence of n (X, Y) coordinates:

$$T = \{(X_1, Y_1), ..., (X_n, Y_n)\} \tag{1}$$

where $(X_i, Y_i) \sim MVN(\mu, \Sigma); i = 1, ..., n, \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$. Also it should be noted that (X_i, Y_i) is the player's displacement at ith coordinate with respect to the previous $(i-1)$th coordinate. For the proposed method, we transformed (X_i, Y_i) into angular form:

$$(X_i, Y_i) \to (R_i, \Theta_i) \tag{2}$$

where R_i is distance to the previous coordinate and Θ_i is angles between each three consecuitve coordinates. In such a condition, we know that if:

$$R_i = \sqrt{X_i^2 + Y_i^2} \tag{3}$$

then

$$R_i \sim Rayleigh(\sigma); \qquad i = 1, ..., n. \tag{4}$$

Also, Θ can follow any arbitrary angular distribution. In this paper, we assume it follows Von Mises (VM) distribution as follows:

$$\Theta_i \sim VM(\theta_0, K). \tag{5}$$

The Von Mises distribution is a circular Normal distribution that can be employed for statistical inference of angular random variables.
Now, to derive their joint density function, Johnson R. A. and Wehrly T. E. (1978) proposed an angular-linear distribution as follows:

$$f_{R,\Theta}(r, \theta) = 2\pi g(\zeta) f_R(r) f_\Theta(\theta); \quad -\pi \leq \theta \leq \pi \;\; ; \;\; -\infty \leq R \leq \infty \tag{6}$$

where $g(.)$ is an angular disribution function of variable ζ, given by:

$$\zeta = 2\pi(F_R(r) - F_\Theta(\theta)) - \pi. \tag{7}$$

In the proposed approach in this paper, we used the joint density function as a foundation for constructing the bivariate GLM. In this framework, the two $tan(.)$ and $log(.)$ functions have been employed for θ and R, respectively. So, for each trajectory, the location parameter of the VM distribution and scale parameter of Rayleigh disribution can be modeled as follow:

$$\theta_0 = 2atan(\beta_0 + \beta_1 \text{time}) \tag{8}$$

$$\sigma = exp(\gamma_0 + \gamma_1 \text{time}) \tag{9}$$

After plugging the above linear predictor in the likelihood function, we can estimate the parameters by Maximum Likelihood approach through the joint density function as follows:

$$(\hat{\beta_0}, \hat{\beta_1}, \hat{\gamma_0}, \hat{\gamma_1}) = argmax(L(\beta_0, \beta_1, \gamma_{0j}, \gamma_1)) \tag{10}$$

where

$$L(\beta_0, \beta_1, \gamma_0, \gamma_1) = \prod_{i=1}^{n} f_{\Theta,R}(\theta_i, r_i | \sigma, \theta_0). \tag{11}$$

In modern football (soccer) games, data are collected on a player's movement 25 times every second over 90 minutes match. Such a long trajectory, for analysing, is required to be broken down into segments representing movement transitions. After estimating the model parameters for each segment, we employed a hierarchical k-means clustering to cluster those segments with similar parameters into the same movement pattern.

3 Results

The segmentation process has been done with respect to the velocity, and those segments with the highest value in velocity (sprinting) are considered to be applied to the proposed model. To prevent any bias in the parameters estimation, all trajectories are rotated to have the same origin points, and the ending points are aligned with the origin points. The Figure 1 represents all the rotated spriniting segments for a given football player. Then, the proposed model in (6) were applied to estimate the model parameter. Then, the hierarchical k-means clustering was applied to the estimated parameters. Figure 2 shows the result of the hierarchical k-means clustering approach. Each parameter represents different interpretation for each segment. For instance, as can be seen in Figure 2 (4) negative $\hat{\gamma_1}$ can be considered as a movement which is decreasing in speed.

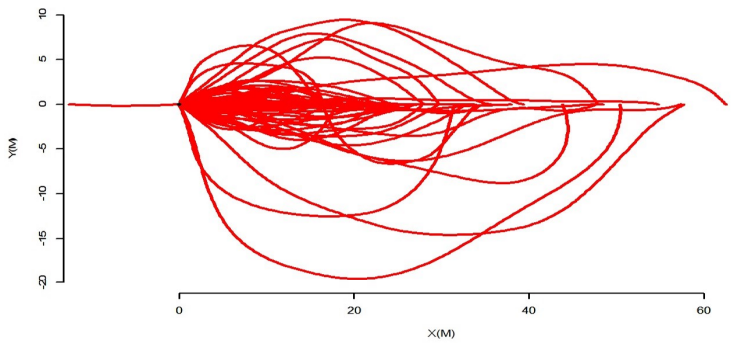


FIGURE 1. All sprinting trajectories with the same origin point

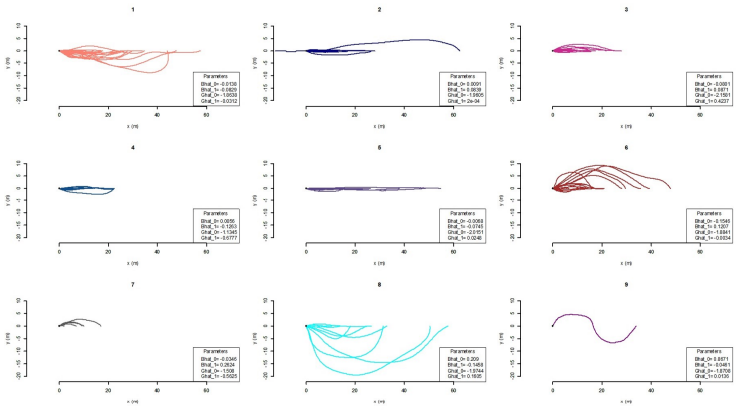


FIGURE 2. Clustered trajectory

4 Conclusion

In this paper, a bivariate GLM was proposed to discover similar movement patterns of a particular player in a football game. The proposed model has

shown it has the capability to differentiate between movements not only in terms of their shape but also in terms of the intensity of the movement. This would be a great advantage to sport scientists and physician to design more effective training sessions and rehab programmes for players who suffer from injury. We are also interested in utilising a sequence pattern-finding approach to uncover the most repeated pattern.

References

Benlakhdar, S., Rziza, M., & Thami, R. O. H. (1994). *Statistical modeling of directional data using a robust hierarchical von mises distribution model: perspectives for wind energy.* , 1 – 21.

Carta, J. A., Ramirez, P., & Bueno, C. (2008). *A joint probability density function of wind speed and direction for wind energy analysis. Energy Conversion and Management.* **49**, 1309 – 1320.

Johnson, R. A., & Wehrly, T. E. (1978). Some angular-linear distributions and related regression models. In: *Journal of the American Statistical Association*,**73(363)**, 602 – 606.

Prati, A., Calderara, S., & Cucchiara, R. (2008). Using circular statistics for trajectory shape analysis. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 1 – 8.

Automatic Variable Selection in Distributional Regression Models using a Smooth Information Criterion

Meadhbh O'Neill[1], Kevin Burke[1]

[1] University of Limerick, Ireland

E-mail for correspondence: `meadhbh.oneill@ul.ie`

Abstract: We propose a variable selection method that is based on a smooth information criterion. Our differentiable problem can be optimized directly with automatic tuning parameter selection and is implemented in a flexible distributional regression setting, where covariates can enter the model through multiple distributional parameters, such as the mean and variance, simultaneously. We apply the method to prostate cancer data using the `smoothic` package in `R`.

Keywords: Variable selection; Information criteria; Penalized maximum likelihood; Heteroscedasticity; Distributional regression.

1 Introduction

Variable selection involves identifying and selecting a subset of relevant and important variables to utilize in model construction. Popular methods, including the LASSO (least absolute shrinkage and selection operator) (Tibshirani, 1996), include a penalty term consisting of the non-differentiable L_1 norm. These procedures have been studied mostly in the areas of normal linear regression and generalized linear models. Covariates enter these classical models through a single location parameter. We expand these methods to include multiple distributional parameters for implementation in more flexible scenarios, such as, when the process under study displays heteroscedastic behaviour.

The level of penalization is controlled by a tuning parameter, which is typically obtained in a computationally expensive manner where an array of different models are fitted to the data (one for each value of the tuning parameter). A "best model" is then selected based on, for example, the

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

cross-validation error or an information criterion (IC), such as the Bayesian IC (BIC). Our proposed approach is in line with the latter, except that we directly optimize the IC in a way that the tuning parameter is fixed from the start, so that an array of models do not need to be fitted. Furthermore, our approach is much more straightforward to implement than existing procedures due to its smooth (differentiable) penalty that permits the use of standard Newton Raphson optimization.
We refer to this as the "smooth IC" (SIC) procedure and extend its use to the emerging area of "distributional regression" (Stasinopoulos et al, 2018), which is also referred to as "multiparameter regression" (MPR) (Burke et al, 2020). This is a more flexible approach where multiple distributional parameters, such as the location and dispersion parameters, are regressed on covariates at the same time. In this MPR setting, the standard penalized regression approaches would become even more computationally intensive since there will be a separate tuning parameter for each distributional parameter, e.g., one for the mean and one for the variance. However, our proposed SIC variable selection procedure is ideally suited to this setting since the tuning parameter is known and fixed from the start.

2 Smooth Information Criterion

We focus on the normal model, although the methodology can easily be used with other models. The log-likelihood function for the MPR normal model is

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n} x_i^T\alpha - \frac{1}{2}\sum_{i=1}^{n} e^{-{x_i}^T\alpha}(y_i - {x_i}^T\beta)^2, \qquad (1)$$

where y_i is the response value and $x_i = (1, x_{1i}, \ldots, x_{pi})^T$ is a vector of covariates for the ith individual over the predictor variables $j = 0, 1, \ldots, p$. The vectors of regression coefficients for the location and dispersion parameters are $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ and, $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)^T$ respectively. Arranging the model selection information criteria as a penalized likelihood (Su et al, 2018) and introducing a smooth approximation to the L_0 norm yields our proposed approach of MPR with smooth IC (MPR-SIC):

$$\ell_\lambda^{\text{SIC}}(\theta) = \ell(\theta) - \frac{\lambda}{2}\big[||\tilde{\beta}||_{0,\epsilon} + ||\tilde{\alpha}||_{0,\epsilon}\big], \qquad (2)$$

where $\theta = (\beta, \alpha)^T$, $\lambda = 2$ or $\lambda = \log(n)$ in the AIC and BIC respectively, $\tilde{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\tilde{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$, i.e., the intercepts are not penalized. The "smooth L_0 norm" is defined as $||\theta||_{0,\epsilon} = \sum_{j=1}^{p} \phi_\epsilon(\theta_j)$, where $\phi_\epsilon(x) = x^2/(x^2 + \epsilon^2)$. This is differentiable for $\epsilon > 0$ and $\lim_{\epsilon\to 0} \phi_\epsilon(x) = ||x||_0$. Smaller values of the smoothing parameter, ϵ, approximate the L_0 norm closely and encourage sparsity, but this also makes the method become less numerically stable.

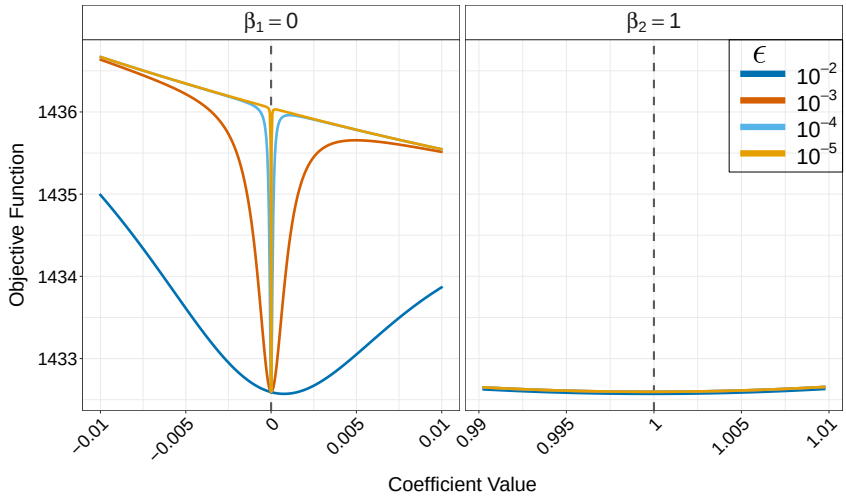


FIGURE 1. Shape of objective function for different values of ϵ.

We recommend "telescoping" through a decreasing sequence of ϵ values to achieve sparsity by squeezing the coefficient values to zero. The estimates from the previous problem are used as initial values for the current nearby optimization problem, therefore making use of "warm starts". This procedure can attain final estimates that are arbitrarily close to zero, and can therefore be treated as zero for practical purposes. The effect of telescoping in relation to the objective function is shown in Figure 1. Different ϵ values in the telescope sequence are shown by different curves. For the true zero coefficient, β_1, the width of the curves become narrower as ϵ decreases, and it is clear that the minimum is concentrated at zero (true values indicated by dashed vertical lines). This shows that there is less uncertainty around the estimate. Moreover, the shape of the objective function for the true non-zero coefficient, β_2, is not impacted by the telescoping method.

3 Analysis of Prostate Cancer Data

We apply our method using the `smoothic` package in `R` (O'Neill and Burke, 2021) to the prostate cancer data. The data come from a study by Stamey et al (1989) and appear in Tibshirani (1996). The estimates and associated standard errors (in brackets) are shown in Table 1. A measure of the effect of the variable is indicated by the ΔBIC value, which is the change in BIC observed after removing the variable from the location or dispersion component of the model. The variable `svi` is selected in both components. Interestingly, the removal of `svi` from the dispersion parameter results in an increase in the BIC of 4.09 units, which is more than the increase in the BIC of dropping the variable from the location parameter (1.64 units).

TABLE 1. Estimation metrics for the prostate cancer data.

	$\hat{\beta}_j$	ΔBIC	$\hat{\alpha}_j$	ΔBIC
`inter`	-1.26 (0.53)		3.15 (1.36)	
`lcavol`	0.47 (0.06)	40.39		
`lweight`	0.82 (0.14)	19.79	-1.17 (0.38)	4.78
`svi`	0.58 (0.22)	1.64	1.07 (0.38)	4.09

Variables not selected: `age`, `lbph`, `lcp`, `gleason`, `pgg45`.

4 Discussion

Our proposed smooth IC distributional regression procedure performs variable selection and parameter estimation simultaneously. Standard gradient based optimization techniques can be used in combination with the telescoping method to produce sparse estimates. Fixing the tuning parameter at $\log(n)$ for the BIC is computationally advantageous, as it avoids the need to fit an array of different models with different tuning parameters. The effectiveness of our method and the advantage of modelling the dispersion are evident from the results of the prostate cancer data, where some dispersion effects have more of an impact on the BIC than location effects.

Acknowledgments: This work was supported by the Confirm Research Centre and Science Foundation Ireland (grant number: 16/RC/3918).

References

Burke, K., Jones, MC., and Noufaily, A. (2020). A flexible parametric modelling framework for survival analysis. *Journal of the Royal Statistical Society, Series C*, **69(2)**, 429 – 457.

O'Neill, M., and Burke, K. (2021). smoothic: Variable Selection Using a Smooth Information Criterion. CRAN R package version 0.1.0.

Stamey, TA., Kabalin, JN., McNeal, JE., et al (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of urology*, **141(5)**, 1076 – 1083.

Stasinopoulos, MD., Rigby, RA., Bastiani, FD. (2018). Gamlss: a distributional regression approach. *Statistical Modelling*, **18(3-4)**, 248 – 273.

Su, X., Fan, J., Levine, RA., et al (2018). Sparse estimation of generalized linear models (glm) via approximated information criteria. *Statistica Sinica*, **28(3)**, 1561 – 1581.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58(1)**, 267 – 288.

Aggregating from Point to Areal Prevalences: A Complete Population Model

John Paige[1], Geir-Arne Fuglstad[1], Andrea Riebler[1], Jon Wakefield[2]

[1] Department of Mathematical Sciences, NTNU, Trondheim, Norway
[2] Department of Statistics and Biostatistics, University of Washington, Seattle, USA

E-mail for correspondence: john.paige@ntnu.no

Abstract: Spatial aggregation of prevalence based on point-referenced observations is used to provide areal predictions at a desired resolution. When observations arise from a population, we argue that spatial aggregation requires a *sampling frame model* that incorporates uncertainty about the population in order to account for three major sources of *aggregation error*: aggregation weights, fine scale variation, and finite population variation. We show via simulation study that, by addressing aggregation error in areal estimates, our proposed sampling frame model is more robust to aggregation grid resolution than common methods. We demonstrate the practical importance of the proposed model with an application to neonatal mortality using the 2014 Kenya demographic and health survey with binary individual outcomes.

Keywords: Aggregation Models; Bayesian Inference; Demographic Health Surveys; Neonatal Mortality; Small Area Estimation.

1 Introduction

Spatial aggregation based on point-referenced observations is an important problem in spatial statistics. If the quantities of interest can be written as integrals of a spatial field, the desired posterior distributions can be computed by block kriging or may be directly available in basis decomposition methods. However, in some cases, point-referenced measurements may represent responses from a 'target' population, a population of interest. In these cases, one might desire aggregate estimates for the target population from which the responses were gathered. We term this problem *spatial aggregation with respect to a population distribution.*

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

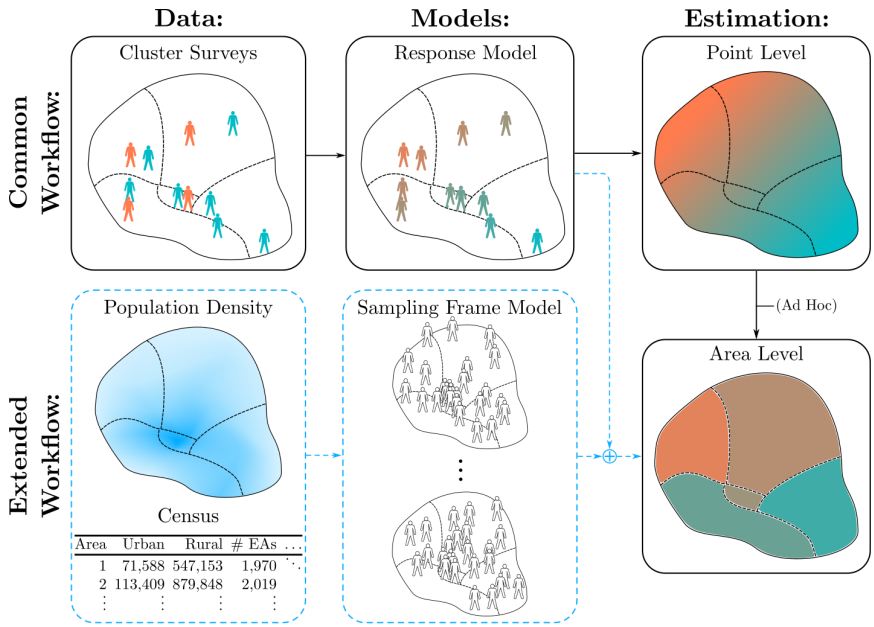


FIGURE 1. Common geostatistical approach to spatial aggregation with respect to a population distribution (in black), with proposed additions (in dashed blue). While population data is sometimes used aggregating from point to areal level in common approaches, this procedure is ad hoc and does not account for several sources of aggregation error.

Our focus is small area estimation of *prevalence*, i.e. the proportion of individuals with outcome 1, based on binary responses (0 or 1), although common approaches instead approximate the prevalence by the *risk*, which is the expected number of individuals with outcome 1. It is worth emphasizing that even if we knew the risk in an area exactly, say $r = 0.7$, the prevalence p could vary widely around this number for a small population size just as a binomial proportion might vary around its probability.

In this context we identify three major sources of *aggregation error*: 1) aggregation weights, 2) fine scale variation, and 3) finite population variation. By aggregation weights we mean the weights used to take a weighted integral or average of point level estimates to produce areal estimates. These weights may involve population density, for example, or the proportion of population in the urban or rural part of an area. Fine scale variation is variability occurring at the finest modeled spatial scale, such as the scale of the response. Fine scale variability could be induced by unmodeled non-spatial or discrete spatial covariates, for example, or other local conditions. Finite population variation is variability caused by the finite size of the target population, and is the cause of variation in prevalence about the

underlying risk.

To address the three major sources of aggregation error we identify, we propose two additions to the common geostatistical workflow depicted in Figure 1. First, we propose accounting for information about the population including population density and census information. Second, and most importantly, we propose using a *sampling frame model* that expresses uncertainty about the population distribution, and depends on the added population density and sampling frame information. The "sampling frame" is the full list of the individuals and auxiliary information such as spatial locations and covariate values in the target population.

2 Data and methods

We estimate the neonatal mortality rate (NMR), prevalence of mortality among children within 28 days of birth using the 2014 Kenya Demographic and Health Surveys (KDHS) survey (Kenya National Bureau of Statistics, 2015). The survey selects 1,582 of the 96,215 enumeration areas (EAs), which are villages or city neighborhoods that compose the sampling frame, and whose exact locations are unknown. 25 households sampled within each selected EA form the survey 'clusters'. We focus on NMR in 2010–2014 in the 301 constituencies in Kenya that sub-partition Kenya's 47 counties. Population density estimates from WorldPop (Tatem, 2017) are normalized to match the urban/rural population census totals as in Paige et al. (2020).

If the locations of each EA with index i were known along with the number of members of the target population that were born, N_i, and died Z_i in the time period, the prevalence in region R could be calculated as:

$$p(R) = \sum_{i=1}^{M} \frac{N_i}{N} \frac{Z_i}{N_i}, \tag{1}$$

where M is the number of EAs in R, and $N = \sum_{i=1}^{M} N_i$.

One geostatistical model for prevalence at the n clusters is:

$$\begin{aligned} y_c | r_c, n_c &\sim \text{Binomial}(n_c, r_c) \\ \text{logit}(r_c) | \boldsymbol{\beta}, \boldsymbol{s}_c, \epsilon_c &= d(\boldsymbol{s}_c)^T \boldsymbol{\beta} + u(\boldsymbol{s}_c) + \epsilon_c, \quad c = 1, \ldots, n. \end{aligned} \tag{2}$$

Here, y_c is and n_c is the number of neonatals that died and were born in the time period in cluster c respectively. The cluster level risk is r_c, and the cluster location is $\boldsymbol{s}_c$. The vector $d(\boldsymbol{s}_c)$ contains spatial covariates, and $\epsilon_c \sim N(0, \sigma_\epsilon^2)$ is the spatial nugget. The spatial effect $u = \{u(\boldsymbol{s}) : \boldsymbol{s} \in \mathbb{R}^2\}$ is a mean zero stationary Gaussian process.

We will assume $d(\boldsymbol{s}) = \beta_0 + \beta^{\text{URB}} I(\boldsymbol{s} \in \text{urban})$, where $I(\boldsymbol{s} \in \text{urban})$ is 1 and 0 if $\boldsymbol{s}$ is urban or rural respectively. We model u via the stochastic partial differential equation (SPDE) approach (Lindgren et al., 2011) using

TABLE 1. 95% credible interval (CI) widths in neonatals per thousand, and empirical coverages in percent of gridded and empirical aggregation models as a function of numerical aggregation grid resolution.

95% CI	**Model**	**Units**	**200m**	**1km**	**5km**	**25km**
Width	Empirical	(per 1,000)	9.4	9.4	9.5	9.5
	Gridded		8.1	18.8	56.3	59.7
Coverage	Empirical	(Percent)	94	94	94	94
	Gridded		90	100	100	100

integrated nested Laplace approximations (Rue et al., 2009) to circumvent the computational expense of Markov chain Monte Carlo.

Estimates for $p(R)$ are commonly obtained by numerically integrating the risk in (2) over a spatial grid indexed by $g \in G$, weighting by population density q normalized to have unit integral on R as, $r_{\text{grid}}(R) = \sum_{g \in G} q(\boldsymbol{s}_g) r_g$, where $r_g = d(\boldsymbol{s}_g)^T \boldsymbol{\beta} + u(\boldsymbol{s}_g) + \epsilon_g$ is the risk of an EA at grid cell g. We call this the 'gridded' sampling frame model. Instead, we propose an 'empirical' sampling frame model using Kenya's 2009 census data and population density information to estimate a distribution for N_i and the EA locations, and to simulate a distribution of possible finite populations. The risk at EA i is then $r(\boldsymbol{s}_i)|\boldsymbol{\beta}, \boldsymbol{s}_i, \epsilon_i = d(\boldsymbol{s}_i)^T \boldsymbol{\beta} + u(\boldsymbol{s}_i) + \epsilon_i$ where $\boldsymbol{s}_i$ is the location of the EA. Conditional on the finite population and risk, Z_i is binomial with risk $r(\boldsymbol{s}_i)$ and with N_i trials.

By assuming there is exactly one nugget effect per grid point, the gridded model does not correctly account for fine scale variation that could be caused by unmodeled nonspatial or discrete spatial covariates and EA-specific conditions. It also assumes prevalence can be exchanged with risk, and so does not account for finite population variability. Although it accounts for aggregation weights using population density, some other common models do not. The empirical model better accounts for fine scale and finite population variability by eliminating these assumptions.

3 Results and discussion

We apply both aggregation models to 100 KDHS-like surveys from 100 populations in Nairobi County simulated using (2) linked with the empirical aggregation model. Populations are simulated on a 5km resolution aggregation grid, while models are fit at various grid resolutions. Mean 95% credible interval (CI) widths and coverages for the 17 constituencies in Nairobi are shown in Table 1. While the gridded model is highly dependent on grid resolution, never achieving nominal coverage with CI widths varying from 8.1 to 59.7 neonatals per thousand, the empirical aggregation

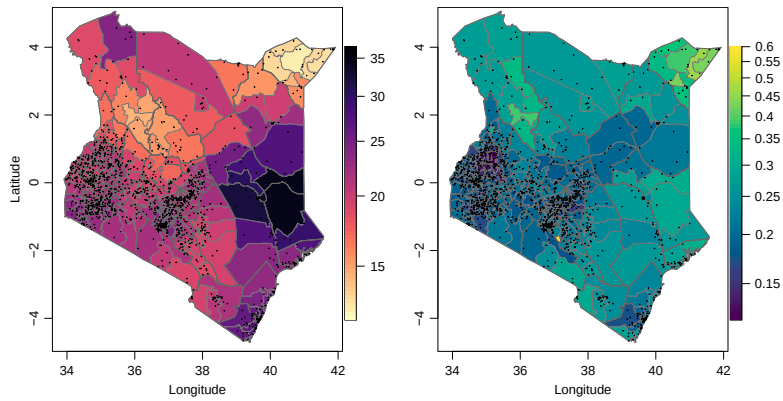


FIGURE 2. NMR estimates in neonatals per thousand (left) and CVs (right) for Kenya constituencies in 2010–2014. Black dots are 2014 KDHS cluster locations.

model is robust to resolution, achieving essentially nominal coverage and similar CI widths in all cases.

We give NMR predictions in neonatals per thousand and coefficients of variation (CV) for the empirical aggregation model applied to the 2014 KDHS in Figure 2. We find central predictions are primarily influenced by the spatial effect and aggregation weights (the urban effect is not statistically significant), while CVs are mainly influenced by fine scale and finite population variation, and EAs and clusters per constituency.

4 Conclusions

We highlight three sources of aggregation error when estimating population prevalence: aggregation weights, fine scale variation, and finite population variation. We propose adding additional steps to the common workflow for estimating population prevalence with geostatistical models. In particular, we propose addressing aggregation error by adding to the workflow a sampling frame model based on population density and census data. We show via simulation study that the resulting predictions and associated uncertainties are more robust to the aggregation grid.

References

KDHS (2014). Kenya demographic and health survey 2014. Kenya National Bureau of Statistics.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). *J. Roy. Stat. Soc. B Met.*, **73**, 423–498.

Paige, J., Fuglstad, G.A., Riebler, A., and Wakefield, J. (2020). Design- and Model-Based Approaches to Small-Area Estimation in a Low and Middle Income Country Context: Comparisons and Recommendations. *J. Surv. Stat. Methodol.*, **10**, 50–80.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Stat. Soc. B Met.*, **71**, 319–392.

Tatem, A.J. (2017). WorldPop, open data for spatial demography. *Sci. Data*, **4**, 1–4.

Bayesian mixture models for time series based on context trees

Ioannis Papageorgiou[1], Ioannis Kontoyiannis[1]

[1] University of Cambridge, United Kingdom

E-mail for correspondence: ip307@cam.ac.uk

Abstract: A rich class of general Bayesian mixture models is introduced for real-valued time series. The mixture models are defined in terms of partitions of the state space, with a different time series model associated to each region of the partition. The state space partitions are defined in terms of a discretized version of the most recent samples and are represented by context-tree models. Together with the general Bayesian modelling framework, a collection of methodological and algorithmic tools are also developed, allowing for exact and computationally efficient Bayesian inference. In particular, it is shown that the maximum *a posteriori* probability (MAP) mixture model can be identified exactly, including the MAP context-tree partition and the MAP time series model fitted to each region. The proposed framework can be used with an arbitrary class of time series models associated to the different state-space regions. Special attention is given to the case of context-tree mixtures of autoregressive (AR) models. The performance of the proposed methods in model selection and forecasting is illustrated through a real-world applications from economics, where they are found to outperform several commonly used approaches.

Keywords: Time series; Bayesian mixture models, State space partitions, Autoregressive models; Context trees.

Extended abstract

Time series modelling. The statistical analysis of time series is an important task with applications across the entire spectrum of applied science and engineering. A wide variety of modelling approaches have been proposed, including autoregressive (AR) models, hidden Markov models, state-space models, deep neural networks and Gaussian processes. However, there still remains the need for rich classes of flexible models that are easily interpretable and suitable for applications with limited training

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

data. In this work, we introduce such a general class of Bayesian mixture models for real-valued time series.

The key element of the proposed approach is that the state space is partitioned based on discretized values of the most recent samples, with a different time series model associated to each region of the partition. We refer to this discretized version of the most recent samples as the *discrete context*, which is extracted from the real-valued observations using a finite-valued quantizer. These state space partitions are represented by discrete *context-tree* models, which are typically simple and easily interpretable, while at the same time capturing important aspects of the underlying structure present in the data.

Context-tree models were initially introduced in the information-theoretic literature in the 1980's, and have been used widely in data compression since then. Recently, they were studied from a Bayesian statistics point of view by Kontoyiannis et al. (2020), who introduced the Bayesian Context Trees (BCT) framework. Extending the ideas and algorithms of the BCT framework, we show that exact Bayesian inference is possible in a computationally very efficient manner for the class of mixture models proposed in this work. In particular, the maximum *a posteriori* probability (MAP) mixture model can be identified exactly, including the MAP context-tree partition and the MAP time series model in each region.

Although the general modelling framework can be used with an arbitrary class of time series models associated to each state-space region, the focus of this paper is when AR models are used. This results in a class of flexible AR mixtures that generalizes popular AR mixtures, including the Threshold Autoregressive (TAR) models (Tong and Lim, 1980) and the Mixture Autoregressive (MAR) models (Wond and Li, 2001). The performance of the proposed methods in model selection and forecasting is illustrated through an applications on real-world data from economics, where they are found to outperform several state-of-the-art approaches.

Bayesian mixture models. The discrete context that will be used to define the state space partitions is extracted from the real-valued observations via a piecewise constant quantizer $Q : \mathbb{R} \to A$. Here $A = \{0, \ldots, m-1\}$, and we assume that the m quantization levels are defined via the thresholds $\{c_1, \ldots, c_{m-1}\}$.

State space partitions. Given a quantizer Q a maximum context length $D \geq 0$, and a context tree T, we define a partition of the state space $\mathbb{R}^D$ via T as follows. For a time series $x = \{x_n\}$, let $t = (Q(x_{n-1}), \ldots, Q(x_{n-D}))$ be the discrete context of length D corresponding to the sample x_n at time n, and let s be the unique leaf of T that is a suffix of t. For example, for the context tree of Figure 1, if $Q(x_{n-1}) = 0$ and $Q(x_{n-2}) = 1$ then $s = 01$, whereas if $Q(x_{n-1}) = Q(x_{n-2}) = 1$ then $s = 1$.

This defines a partition of $\mathbb{R}^2$ into three regions indexed by the contexts $\{1, 01, 00\}$ corresponding to the leaves of T.

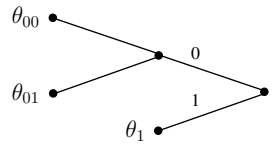


FIGURE 1. Example of a binary context tree T representing the partition.

To complete the specification of the hierarchical model, we associate a different model to each leaf s of the context tree T, giving a different conditional density for x_n: At time n, given the context s determined by the past D samples $(x_{n-1}, \ldots, x_{n-D})$, the distribution of x_n is given by the model assigned to leaf s. Although arbitrary time series models could be used in general, here we associate AR models with parameters θ_s to each leaf s, and refer to the resulting model class as the *Bayesian context tree autoregressive model* (BCT-AR).

Prior structure. At the top level, we consider state space partitions represented by context trees T in the collection $\mathcal{T}(D)$ of all proper m-ary trees with depth no greater than D, where a tree T is called proper if any node in T that is not a leaf has exactly m children. Following Kontoyiannis et al. (2020), for the trees $T \in \mathcal{T}(D)$ we use the BCT prior,

$$\pi(T) = \pi_D(T; \beta) = \alpha^{|T|-1} \beta^{|T|-L_D(T)} \ ,$$

where $\beta \in (0, 1)$ is a hyperparameter, $\alpha = (1-\beta)^{1/(m-1)}$, $|T|$ is the number of leaves of T, and $L_D(T)$ is the number of leaves of T at depth D. This prior penalizes larger trees by an exponential amount.
Given a tree $T \in \mathcal{T}(D)$, we place an independent prior on each θ_s, so that $\pi(\theta|T) = \prod_{s \in T} \pi(\theta_s)$. For the parameters θ_s, we use a Gaussian prior for the AR coefficients and an inverse-gamma prior for the noise variance.

Exact Bayesian inference. The proposed prior structure allows for exact Bayesian inference, in a computationally very efficient manner, by using modified versions of the algorithms of Kontoyiannis et al. (2020). In particular, for a time series x, the prior predictive likelihood $p(x)$ can be computed exactly, with all tree models T and parameter vectors θ integrated out. Also, the MAP BCT-AR model can be identified exactly, including the MAP context-tree partition and the MAP AR parameters within each region.
Next, we give an example application of these methods to a real-world data set.

US unemployment rate. The first example we consider is the quarterly US unemployment rate, in the time period 1948-2019 (288 observations).

The fitted MAP BCT-AR model is the tree of Figure 1, identifying three meaningful states: Jumps higher than a threshold ($c = 0.15$) correspond to economic contractions (context 1), context 00 corresponds to a stable economy, and context 01 corresponds to stabilizing just after a contraction. In Table 1, the performance of our methods in forecasting is compared with the most successful earlier approaches (Montgomery et al., 1998).

TABLE 1. Mean squared error of forecasts (with a 50-50 training/test set split)

	Prediction step				
Model	1	2	3	4	5
Seasonal ARIMA	5.40	7.71	10.1	11.6	11.0
SETAR	5.42	8.34	8.82	9.48	9.95
MAR	5.33	7.61	8.92	9.56	9.71
BCT-AR	**4.90**	**7.33**	**8.44**	**9.08**	**9.48**

References

Kontoyiannis, I., Mertzanis, L., Panotopoulou, A., Papageorgiou, I., and Skoularidou, M. (2020). Bayesian Context Trees: Modelling and exact inference for discrete time series. *Journal of the Royal Statistical Society: Series B*, to appear, 2022. Available at *arXiv:2007.14900*.

Montgomery, A., Zarnowitz, V., Tsay, R., and Tiao, G. (1998). Forecasting the US unemployment rate. *Journal of the American Statistical Association*, **93**, 478 – 493.

Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B*, **42**, 245 – 268.

Wong, C.S. and Li, W.K. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society: Series B*, **62**, 95 – 115.

Markov-switching step selection analysis

Jennifer Pohle[1], Johannes Signer[2], Ulrike Schlägel[1]

[1] University of Potsdam, Germany
[2] University of Goettingen, Germany

E-mail for correspondence: jennifer.pohle@uni-potsdam.de

Abstract: Integrated step selection analysis is a popular statistical tool to study animal's movement and habitat selection using conditional logistic regression. In this paper, we extend this framework by introducing an underlying latent Markov chain which allows preferences and movement patterns to vary over time. A simulation study is used to investigate the performance of the resulting Markov-switching integrated step selection analysis and to compare it to alternative candidate models. Besides habitat selection, the inherent regression model is also applicable for longitudinal discrete choice and case-control studies.

Keywords: time series; latent variables; conditional logistic regression; animal movement; hidden Markov model

1 Introduction

A central interest in ecology lies on animals' habitat and resource use. Combining animal movement and environmental data, integrated step selection analysis (iSSA) is a popular framework for studying fine-scale habitat selection of the animal moving through the landscape, while also considering movement capacities (Avgar et al., 2016). It relies on a conditional logistic regression (CLR) to compare the characteristics of *used* locations where the animal was observed, against alternative locations *available* at a given time point (see Figure 1). However, preferences and movement patterns might depend on the animals' usually unobserved behavioural modes such as resting or foraging. Ignoring such states in the analysis might lead to biased results and misleading conclusions. Therefore, it has recently been suggested to first apply hidden Markov models (HMMs) to the movement data for latent state decoding, before fitting state-specific iSSAs (HMM-iSSAs) in a second step (Karelus et al., 2020). This two-step approach accounts for the latent state structure, but it ignores uncertainties in the state classification

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

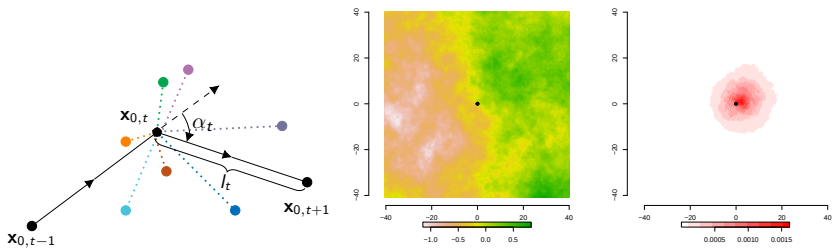


FIGURE 1. Left panel: Illustration of a movement trajectory with used locations in black and alternative available locations for time point $t+1$ in colour. Middle and right panel: Example landscape and example spatial density from the simulation study (state 2.)

and might confound movement and selection patterns. Similar to Nicosia et al. (2017), in this paper, we therefore propose a Markov-switching iSSA (MS-iSSA) which allows for a simultaneous estimation of the state, selection and movement parameters. A simulation study is used to compare the performance of the MS-iSSA to alternative models.

2 Methodology

Let $\{\mathbf{x}_{0,t}\}_{t=1}^{T}$ denote the time series of observed locations of length T, with $\mathbf{x}_{0,t} \in \mathbb{R}^2$. We assume the movement to be driven by an underlying N-state discrete-time Markov chain $\{S_t\}_{t=1}^{T}$, defined by the transition probability matrix $\Gamma = (\gamma_{ij})$ with $\gamma_{ij} = \Pr(S_t = j \mid S_{t-1} = i)$. Conditional on the current state $S_t = i$, and locations $\mathbf{x}_{0,t-1}$ and $\mathbf{x}_{0,t}$, the spatial density for the next location $\mathbf{x}_{0,t+1}$ can be modelled as:

$$f_i(\mathbf{x}; \boldsymbol{\theta}^{(i)}, \boldsymbol{\beta}^{(i)}) = \frac{\overbrace{\phi(\mathbf{x} \mid \mathbf{x}_{0,t}, \mathbf{x}_{0,t-1}; \boldsymbol{\theta}^{(i)})}^{\text{selection-free movement kernel}} \cdot \overbrace{\exp(\mathbf{Z}(\mathbf{x})^\top \boldsymbol{\beta}^{(i)})}^{\text{movement-free selection function}}}{\underbrace{\int_{\tilde{\mathbf{x}} \in D_x} \phi(\tilde{\mathbf{x}} \mid \mathbf{x}_{0,t}, \mathbf{x}_{0,t-1}; \boldsymbol{\theta}^{(i)}) \cdot \exp(\mathbf{Z}(\tilde{\mathbf{x}})^\top \boldsymbol{\beta}^{(i)}) d\tilde{\mathbf{x}}}_{\text{normalising constant}}},$$

where $\mathbf{Z}(\mathbf{x})$ denotes the location-specific covariate vector (including, e.g., land-cover type, snow depth), and $D_{\mathbf{x}}$ the availability domain. $\Phi(\cdot)$ describes the space use density in absence of any habitat selection. We consider a gamma distribution for step length (i.e. moved distance) $l_t = \|\mathbf{x} - \mathbf{x}_{0,t}\|$ with state-dependent parameters $\theta_1^{(i)}$ and $\theta_2^{(i)}$ for shape and rate, respectively, and a uniform distribution for turning angle (i.e. directional change) $\alpha_t \sim \mathcal{U}_{[-\pi,\pi)}$, but other choices are possible. The movement kernel is weighted by a log-linear function involving the state-dependent

selection coefficient vector $\boldsymbol{\beta}^{(i)}$ which indicates preference for or avoidance of habitat characteristics.
The integral in the denominator is usually intractable. However, justified by the law of large numbers, parameter estimation can rely on CLR using J control locations $\mathbf{x}_{j,t+1}$ randomly drawn from $D_{\mathbf{x}}$ (Nicosia et al., 2017). This leads to the state-specific choice probability:

$$p_{0ti} = \frac{\exp\left[\log(l_{0,t})(\theta_1^{(i)} - 1) - l_{0,t}\theta_2^{(i)} + \mathbf{Z}(\mathbf{x}_{0,t+1})^\top \boldsymbol{\beta}^{(i)}\right]}{\sum_{j=0}^{J} \exp\left[\log(l_{j,t})(\theta_1^{(i)} - 1) - l_{0,t}\theta_2^{(i)} + \mathbf{Z}(\mathbf{x}_{j,t+1})^\top \boldsymbol{\beta}^{(i)}\right]}$$

Plugging p_{0ti}, $i = 1, \ldots, N$, into the HMM likelihood (Zucchini et al., 2016), parameters can be estimated using numerical maximisation of the likelihood, which is evaluated using the forward algorithm.

3 Simulation Study

We use a simulation study to evaluate the performance of the proposed method and compare it to the two-step approach (HMM-iSSA) and basic iSSAs. In each of the 100 simulation runs, we generated a covariate as a synthetic landscape $\mathbf{Z}$ based on Gaussian random fields. Then $T = 1000$ locations were simulated based on an MS-iSSA model with $N = 2$ states, transition probabilities $\gamma_{11} = \gamma_{22} = 0.9$, gamma distribution parameters $\boldsymbol{\theta}^{(1)} = (1.2, 1.25)$ in state 1, and $\boldsymbol{\theta}^{(2)} = (2.25, 0.29)$ in state 2. State 2 was associated to selection for the landscape feature using $\beta^{(2)} = 2$ (see Figure 1). For state 1, we considered two scenarios: (1) $\beta^{(1)} = 0$, i.e. no selection, (2) $\beta^{(1)} = -2$, i.e. avoidance of the landscape feature.
Figure 2 shows the estimates of the MS-iSSA, HMM-iSSA and iSSA fitted to the simulated data in each simulation run using $J = 20$, 100 and 500 randomly drawn control locations. The results indicate that for the HMM-iSSA and iSSA, the selection coefficients are biased. The MS-iSSA estimates are reasonable, especially for $J = 500$, although the variance of $\hat{\beta}^{(2)}$ is rather large in both scenarios. In additional simulations with $T = 5000$, the estimation performance further improves.

4 Discussion

The simulation study suggests that the MS-iSSA is a promising tool for analysing habitat selection and movement in $\mathbb{R}^2$ while considering state-switching dynamics over time. Current work focuses on refining the sampling procedure for the control locations, model selection, and a case study using telemetry data from hares. While the presented MS-iSSA method is designed for habitat and movement analyses, the inherent Markov-switching CLR is applicable also in other areas, e.g. for longitudinal discrete choice studies in transportation or case-control-studies in medicine.

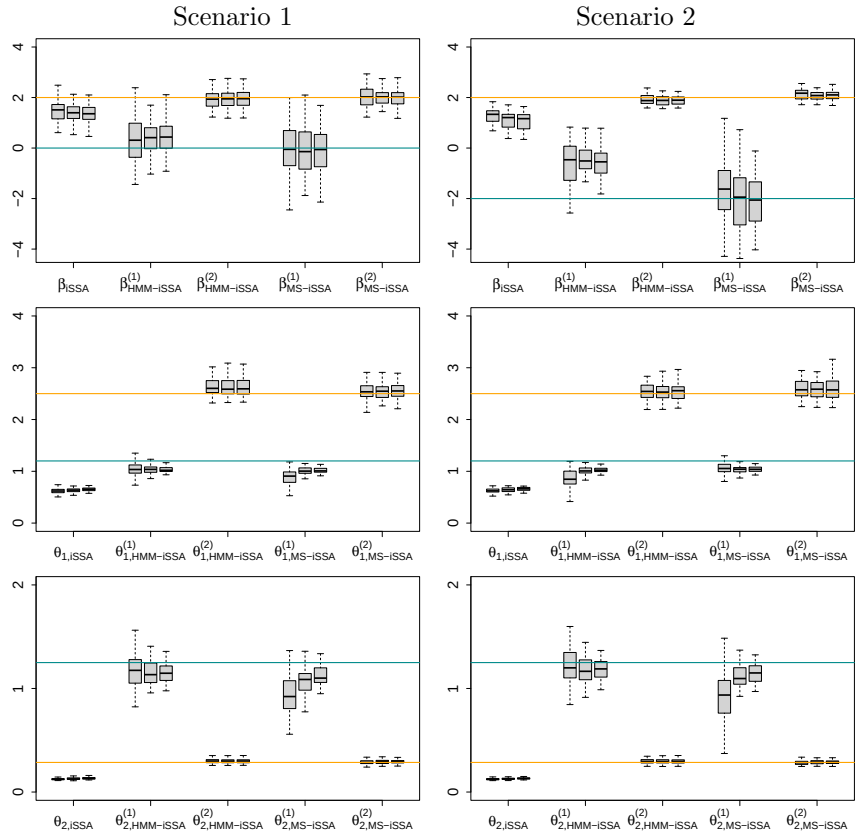


FIGURE 2. Boxplots of the parameter estimates across the 100 simulation runs for each fitted model, respectively. The upper row displays the selection coefficients, the middle row the shape parameters and the bottom row the rate parameters. The three boxplots per parameter and model indicate the use of $J = 20$ (left), 100 (middle), and 500 (right), respectively. True parameter values are indicated by the green (state 1) and yellow (state 2) lines.

References

Avgar, T., Potts, J.R., Lewis, M.A., and Boyce, M.S. (2016). Integrated step selection analysis: bridging the gap between resource selection and animal movement. *Methods in Ecology and Evolution*, **7**, 619–630.

Karelus, D.L., McCown, J.W., Scheick, B.K., van de Kerk, M., Bolker, B.M., and Oli, M.K. (2020). Corrigendum to: Incorporating movement patterns to discern habitat selection: black bears as a case study. *Wildlife Research*, **47**, 359 – 360

Nicosia, A., Duchesne, T., Rivest, and L.-P., and Fortin, D. (2017). A multi-state conditional logistic regression model for the analysis of animal movement. *The Annals of Applied Statistics*, **11**, 1537 – 1560.

Zucchini, W., MacDonald, I.L., and Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R*, 2nd Edition. Boca Raton: Chapman & Hall/CRC.

Spatial Joint Models through Bayesian Structured Piecewise Additive Joint Modelling for Longitudinal and Time-to-Event Data

Anja Rappl[1], Thomas Kneib[3], Stefan Lang[2], Elisabeth Bergherr[3]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] Universität Innsbruck, Austria
[3] Georg-August-Universität Göttingen, Germany

E-mail for correspondence: anja.rappl@fau.de

Abstract: Joint models for longitudinal and time-to-event data have seen a lot of extension over the recent years, spatial joint models, however, are still rare. By substituting the commonly used proportional hazards model by a piecewise additive mixed model we allow for more flexibility with respect to the baseline hazard and the form of the predictors included in the model. The results of a simulation study support this approach.

Keywords: joint model for longitudinal and time-to-event outcomes, piecewise additive mixed model, Bayesian statistics

1 Introduction

Biometrical studies often capture time-to-event and longitudinal data simultaneously. Since separate analysis of these outcomes leads to biased estimates, both should be modelled jointly. These joint models consist of two submodels: A longitudinal submodel and a survival submodel with both being linked through an association parameter. The former is traditionally a linear mixed model and the latter a proportional hazards model. This model type has seen a lot of expansions in recent years covering different estimation approaches (for a comparison see Rappl et al., 2020), model variations and types of submodels. Alsefri et al. (2020) give a concise summary of advances in Bayesian joint models in particular.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

Still joint models with a spatial component are rare, e.g. Martins et al. (2016) have described their Bayesian estimation with a Weibull baseline hazard and Köhler et al. (2017) a Bayesian flexible tensor-product approach using Newton-Raphson procedures and derivative-based Metropolis-Hastings sampling. However, assuming a parametric baseline hazard can be quite restrictive and derivative-based Metropolis-Hastings algorithms are computationally expensive. Therefore, we propose a similarly flexible and faster Bayesian approach to joint modelling by choosing a piecewise additive mixed model (PAMM, Bender et al., 2018) in the time-to-event submodel instead. This allows for spatial, (non-)linear and random effects to be included as well as estimation of the baseline hazard without any assumptions about its distributional form.
More theoretical background of this approach can be found in Section 2, while Section 3 highlights the results of a simulation study followed by a short discussion of our findings in Section 4.

2 Methodological background

Let $\boldsymbol{y}$ denote the vector of longitudinal outcomes across all individuals $i = \{1, \ldots, n\}$ and observations times points $t = \{1, \ldots, n_i\}$ and $\boldsymbol{\lambda}(t)$ the vector of individual specific risks to experience an event at time t proportional to the baseline hazard $\lambda_0(t)$ and based on the observed event times $\boldsymbol{T}$ and censoring/event indicator $\boldsymbol{\delta}$. Then in its most generic form a joint model looks like

$$\boldsymbol{y}(t) = \boldsymbol{\eta}_{\text{l}}(t) + \boldsymbol{\eta}_{\text{ls}}(t) + \varepsilon, \quad \varepsilon \sim \text{N}(0, \sigma^2 \boldsymbol{I}) \tag{1}$$

$$\boldsymbol{\lambda}(t) = \lambda_0(t) \exp\{\boldsymbol{\eta}_{\text{s}} + \gamma_\alpha \cdot \boldsymbol{\eta}_{\text{ls}}(t)\}, \tag{2}$$

where $\boldsymbol{\eta}_{\text{s}}$ and $\boldsymbol{\eta}_{\text{l}}$ are survival and respectively longitudinal submodel specific predictors and $\boldsymbol{\eta}_{\text{ls}}$ is the predictor that connects both model parts through the association parameter γ_α.
Now (2) can be re-written as a log-linear Poisson-model of counting events $\boldsymbol{\delta}_j$ in interval j by dividing the observation time $(0, t_{max}]$ into J intervals with boundaries $0 = \kappa_0 < \cdots < \kappa_J = t_{max}$ and assuming constant baseline hazards λ_j within each interval (piece-wise exponential model). If the interval-specific log-baseline hazard $\log \lambda_j$ is further represented as a smooth function of time $f_0(t_j)$ instead of a step-function the model generalises to a PAMM. Together with the exposure times $\boldsymbol{o}_j = (o_{1j}, \ldots, o_{nj})'$ of each individual i in each interval j as offsets we can equivalently write (2) as

$$\boldsymbol{\lambda}(t) = \exp\left\{f_0(t_j) + \boldsymbol{o}_j + \boldsymbol{\eta}_{\text{s}} + \alpha \boldsymbol{\eta}_{\text{ls}}\right\}, \quad \forall\, t \in (\kappa_{j-1}, \kappa_j]. \tag{3}$$

The predictors $\boldsymbol{\eta}_{\cdot}$ are generically specified such that they may include any function $f^{\cdot} = f_k(\boldsymbol{z}_k)$ of any covariate vector $\boldsymbol{z}_k \forall\ k = 1, \ldots, K$, where $f^{\cdot}$

may denote linear ($\boldsymbol{f}^{lin}$), smooth ($\boldsymbol{f}^{sm}$), spatial ($\boldsymbol{f}^{geo}$) and random ($\boldsymbol{f}^{rdm}$) effects of their respective covariates. The only restrictions here are that the random effects $\boldsymbol{f}^{rdm}$ are part of the shared predictor $\boldsymbol{\eta}_{\text{ls}}$ and there must be only one spatial effect f^{geo} in the model for identifiability reasons. Hence a specific predictor can be expressed as

$$\boldsymbol{\eta}_{\cdot} = \boldsymbol{f}^{lin} + \boldsymbol{f}^{sm} + f^{geo} + \boldsymbol{f}^{rdm} = \boldsymbol{\eta}_{\cdot} = \boldsymbol{Z}_{1,\cdot}\boldsymbol{\gamma}_{1,\cdot} + \ldots \boldsymbol{Z}_{K,\cdot}\boldsymbol{\gamma}_{K,\cdot}, \quad (4)$$

where the second part corresponds to the matrix notation of an effect as effect specific design matrix $\boldsymbol{Z}_{k,\cdot}$ and corresponding coefficients $\boldsymbol{\gamma}_{k,\cdot}$.
The prior structure for all parameters in the model then is straightforward - note that the baseline hazard $f_0(t_j)$ and the association parameter γ_α are no different from any other effect in the predictors - and given by

$$p\left(\boldsymbol{\gamma}_k \mid \sigma^2_{\gamma_k}\right) \propto \sigma^{-\text{rk}(\boldsymbol{K}_k)}_{\gamma_k} \exp\left\{-\frac{1}{2\sigma^2_{\gamma_k}}\boldsymbol{\gamma}'_k\boldsymbol{K}_k\boldsymbol{\gamma}_k\right\},$$
$$\sigma^2_{\gamma_k} \sim \text{IG}(a,b), \sigma^2_\varepsilon \sim \text{IG}(a_0,b_0)$$

with the data likelihoods as follow

$$\boldsymbol{y} \sim \text{N}(\boldsymbol{\eta}_{\text{l}}(t) + \boldsymbol{\eta}_{\text{ls}}(t), \sigma^2_\varepsilon \boldsymbol{I}),$$
$$\boldsymbol{\delta}_j \sim \text{Poi}(\log \boldsymbol{\lambda}(t)) \quad \forall\, t \in (\kappa_{j-1}, \kappa_j].$$

3 Simulation study

In order to assess the performance of our approach we examined $R = 100$ replications of estimates from simulated data of $n = 200$ individuals across originally $n_i = 6$ observation times points from model (1) with $\gamma_\alpha = -0.3$ and the predictors

$$\eta_{\text{l}} = -0.5\, x_l(t),\ \eta_{\text{s}} = 0.1\, x_s \text{ and}$$
$$\eta_{\text{ls}} = 0.9\, x_{ls,1} - 0.5 \sin(x_{ls,2}) - 0.5\, x_{ls,3} + f_{geo}(s) + 0.4\, t + b_0 + b_1\, t.$$

The model variance was set to $\sigma^2 = 0.5$ and the variance-covariance of the random effects to $\boldsymbol{B} = \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$.
Coverage of the true effects within the 95%-high density region and mean squared error (MSE) were used to evaluate the performance of the above model and the results can be found in Table 1. The results in the table indicate low MSE for the estimation of all effects and that coverage of 95% is met. The spatial effect is covered in 96 out of 100 replications with an MSE of 0.057. Estimated with the largest error are the random effects, yet the estimations still cover the true values in 95.3% and 94.3% of the cases respectively. Illustrations of the effect estimates of the geographical covariate as well as the random effects against the true values can be found in Figure 1. Given the amount and type of the effects in this model as well as the rather small sample size of $n = 200$ individuals the results already indicate the solid performance of this approach.

TABLE 1. Mean squared error and coverage of 100 replications by effect

	MSE	HDI95 (in %)
longitudinal		
intercept	0.017	91.0
x_l	0.001	93.0
shared		
time	0.026	96.0
$x_{ls,1}$	0.014	91.0
$x_{ls,3}(t)$	0.004	88.0
$f(x_{ls,2})$	0.039	95.4
f_{geo}	0.057	96.0
$\boldsymbol{b}_0$	0.315	95.3
$\boldsymbol{b}_1$	1.223	94.3
survival		
γ_α	0.004	96.0
x_s	0.011	95.0
$f_0(t)$	0.039	96.6

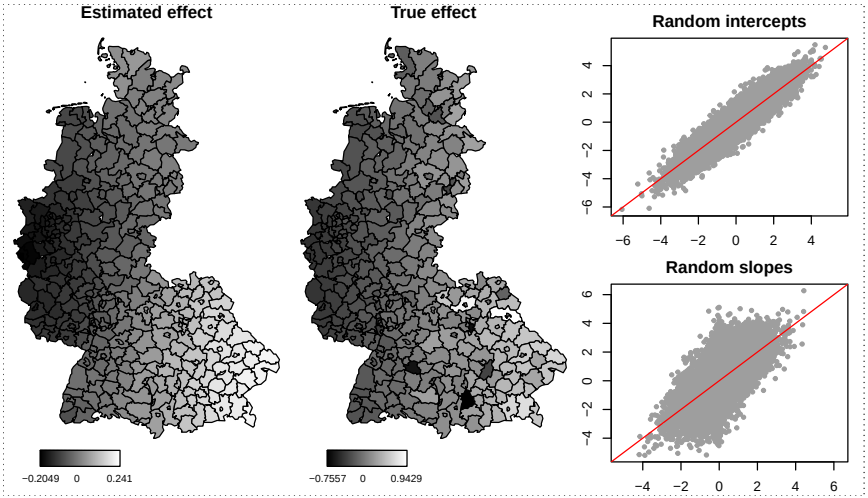


FIGURE 1. Illustration of estimates of spatial effects and random intercepts and slopes against their true values.

4 Summary and Discussion

Despite the advances joint models have seen over the recent years spatial joint models are still rare, which could be due to the restrictions based on the baseline hazard for estimation. By using a piecewise additive mixed model for the survival submodel we suggest a more flexible approach to not just estimate the baseline hazard but theoretically include any type of effect in the predictors of the model.
The results of a first simulation study are very promising and we are already in the process of substantiating these findings with more elaborate simulated models and real data examples. Those more elaborate models could compare the predictor specific performance of spatial covariates, time-varying effects and different shapes of baseline hazards.

References

Alsefri, M., Sudell, M., García-Fiñana, M. and Kolamunnage-Dona, R. (2020). Bayesian joint modelling of longitudinal and time to event data: a methodological review. *BMC Med Res Methodol*, **20(94)**.

Bender, A., Groll, A. and Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, **18(3-4)**, 299 – 321.

Köhler, M., Umlauf, N., Beyerlein, A., Winkler, C., Ziegler, A., Greven, S. (2017). Flexible Bayesian additive joint models with an application to type 1 diabetes research.*Biometrical Journal*, **00**, 1 – 18.

Martins, R., Silva, G. L., and Andreozzi, V. (2016). Bayesian joint modeling of longitudinal and spatial survival AIDS data. *Statist. Med.*, **35**, 3368 – 3384.

Rappl, A., Mayr, A. and Waldmann, E. (2021).More than one way: Exploring the capabilities of different estimation approaches to joint models for longitudinal and time-to-event outcomes. *International Journal of Biostatistics*, **20200067**.

Non-linear modelling of systolic and diastolic blood pressures via environmental factors

Dayasri Ravi[1], Andreas Groll[1], Tamara Schikowski[2]

[1] Department of Statistics, TU Dortmund University, Germany
[2] IUF-Leibniz Research Institute for Environmental Medicine, Germany

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

Abstract: Systolic and diastolic blood pressures have always been closely associated with environmental factors such as temperature and relative humidity. However, the interaction effect between these environmental factors in modelling blood pressure is often not considered. We aim to use generalized additive models to model blood pressures as the environmental data often display a non-linear pattern. The explanatory variables may often have different measuring units. The tensor product spline approach is practical to model the interaction effect among the environmental explanatory variables instead of the isotropic smoothing.

Keywords: Generalized additive models; Tensor product splines; P-splines.

1 Introduction

Several studies have shown a significant relationship between blood pressure, temperature and relative humidity (e.g., Barnett et al., 2007). Although blood pressures are predicted reasonably well using environmental variables such as temperature, it is interesting to consider the interaction among these variables. We see that Generalized additive models (GAMs) are a better choice over Generalized linear models (GLMs) for modelling the effects of climatic and environmental variables (see Ravindra et al., 2019). Additive models are the sum of smooth, typically non-linear functions of the explanatory variables. These models allow for more flexible relationship between the variables compared to linear modelling. As the nature of the relationship between the response and explanatory variable(s) dictates the model, it can be analyzed non-parametrically. GAMs can be seen as an extension of GLMs, allowing the linear predictor to be expressed as smooth additive functions. GAMs can thus handle both non-linear and non-

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

monotonous relationships between the response and explanatory variables. The variety of choices in smoothing functions can substantially improve the performance of the GAM model.
We analyzed the systolic and diastolic blood pressures of 635 women examined during two follow-ups. The aim was to model the blood pressures via socio-demographic covariates such as age, household, smoking pattern, schooling years, etc., and focus on the interaction effect between the environmental covariates.

2 Materials, Methods and Data

In the following section, we introduce the idea of GAMs and bivariate tensor product splines to model interaction effects. We briefly explain the data and the statistical analyses used. Finally, we state the critical results found in our study.

2.1 Methodology

Consider response variable y_i associated to covariates $\{x_{1i}, x_{2i}, \ldots, x_{pi}\}$. The GAM for modelling the data $\{y_i, x_{1i}, x_{2i}, \ldots, x_{pi}; i = 1, 2, .., n\}$ can be extended (see Wood, S. N., 2006) from GLMs with $g(\cdot)$ as a known monotonic differentiable link function to have the following structure

$$g(E[y_i]) = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \ldots + f_p(x_{pi}) \ . \tag{1}$$

Here, the functions $f_1(\cdot), f_2(\cdot), \ldots, f_p(\cdot)$ are unknown smooth functions to be estimated. To simplify the estimation of these functions, we represent them in such a way that Equ. (1) becomes a linear model. For simplicity, in the following we consider a single function $f(x)$ of one covariate x. We can represent the function $f(\cdot)$ through a linear combination of a set of basis functions $\{b_j; j = 1, 2, \ldots, d\}$ as

$$f(x) = \sum_{j=1}^{d} \alpha_j b_j(x) \ , \tag{2}$$

where the α_j's are unknown spline coefficient parameters and d is the corresponding number of basis functions.
In the following, we use the B-spline basis function approach for the basis functions $b_j(x)$ (see Eilers, P. H., and Marx, B. D., 2021). The smoothness and the number of B-splines depend on the number of knots. So, we consider a roughness penalty to overcome this strong dependence on the number of knots. Penalization can be applied to k-th order differences between the coefficients α_j from Equ. (2). Thus, instead of regular least squares, i.e. $\sum_{i=1}^{n}(y_i - f(x_i))^2$, we minimize the penalized least squares (PLS) criterion.

This is called the penalized spline (P-splines; see Eilers, P. and Marx, B., 2021) approach.
The k-th order P-spline based on $d = l + m - 1$ B-splines can be estimated by the following penalized residual sum of squares

$$PLS(\lambda) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{d}\alpha_j b_j(x_i))^2 + \lambda \sum_{j=k+1}^{d}(\Delta^k \alpha_j)^2, \tag{3}$$

where Δ^k is the k-th order difference among the coefficients α_j. $\lambda > 0$ is known as the smoothing parameter that controls the trade-off between the model fitting and smoothness.
In order to model the interaction between two covariates, x_1 and x_2, we can use the tensor product bases. We construct the univariate bases for x_1 and x_2 via $a_j(x_1), j = 1, 2, \ldots, d_1$, and $b_k(x_2), k = 1, 2, \ldots, d_2$, respectively. The bivariate smooth function $f_{12}(x_1, x_2)$ of x_1, x_2 has the following form

$$f_{12}(x_1, x_2) = \sum_{j=1}^{d_1}\sum_{k=1}^{d_2}\gamma_{jk} a_j(x_1) b_k(x_2) \ . \tag{4}$$

Here, we apply double penalization to both rows and columns of B-splines with respective smoothing parameters λ_1 and λ_2.
In case that random effects are included in the model, we can extend the predictor from Equ. (1) by $\mathbf{Zb}$, where $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_b^2)$ represents the vector of random effects and $\mathbf{I}$ an identity matrix of suitable dimension with respect to the random effects components.

2.2 Study population

Data were collected from the IUF-Leibniz Research Institute for Environmental Medicine, as a part of the Study on Influence of Air Pollution on Lung, Inflammation and Aging (SALIA) cohort. The data comprises 635 women examined for systolic and diastolic blood pressures on two follow-ups recorded between 2007 to 2008 and 2012 to 2013. The climate data, temperature (in °C) and relative humidity (in %), were collected up to 30 days before the examination. Additional socio-demographic covariates such as age, Body Mass Index (BMI), years in school, smoking behaviour, diabetes, living conditions, location, etc., were updated at each follow-up.

2.3 Statistical analyses

We fit the blood pressure to a GAM with an identity link, and we select P-splines as the smoothers for the covariates. We model the temperature and relative humidity interaction via a bivariate tensor product P-spline. Additionally, we consider random intercepts for each study participant to

account for subject-specific variability. Taking the seasonal variations into consideration, we subset the data into warmer and colder months based on the date of examination and temperature. We examine the delayed effects of environmental variables by taking moving average lags up to 30 days before the examination. The lag structure includes lag 0, lag 0-1, lag 0-3, lag 0-5, lag 0-10, lag 0-20 and lag 0-30. Here, lag 0 represents the daily mean temperature and daily relative humidity taken on the day of examination (`temp_0`,`rh_0`), i.e. no lag at all, and lag 0-30 represents the moving average of daily mean temperature and daily relative humidity taken from the day of examination to 30 days before the examination (`temp_mean30`,`rh_mean30`). We choose the best moving average lag structure based on the Akaike information criterion (AIC).

2.4 Results

We find a huge subject-specific random effect for both systolic and diastolic blood pressures. The estimated standard deviations for the random effects are 15.289 (12.807-18.252, 95 % CI) and 7.579 (6.298-9.121, 95 % CI) for systolic and diastolic blood pressures, respectively. Based on the AIC score, we select the model with lag 0 for systolic and diastolic blood pressures for the warmer months (April - September). During colder months (October - March), we select the model with lag 0-10 for systolic blood pressure and lag 0-20 for diastolic blood pressure. Fig. 1 displays a robust and highly non-linear interaction effect between daily mean temperatures and daily relative humidity. From Fig. 2, we observe that the effect of age is linear for both systolic and diastolic blood pressures. We also notice a negative effect of age on diastolic blood pressure, which seems to be consistent with epidemiological results (e.g., Pinto E., 2007).
Our findings suggest that the bivariate interaction between temperature and relative humidity plays a critical role in modelling the blood pressures during the warmer months (p-values < 0.01) more than in colder months, where the effect was not significant. During colder months, the covariate location (rural or urban) shows significance (p-value $= 0.00674$ for diastolic and p-value $= 0.0099$ for systolic). Age also plays a significant role (p-value < 0.01) for all models except for systolic blood pressure in colder months. Some covariates such as the number of smoking packets per day and BMI show a non-linear effect on blood pressure. However, other covariates, such as years in school, diabetes, and heating conditions do not seem relevant for modelling systolic and diastolic blood pressures.

3 Conclusion

We have proposed a generalized additive model to understand the effect of environmental factors on systolic and diastolic blood pressures. We also

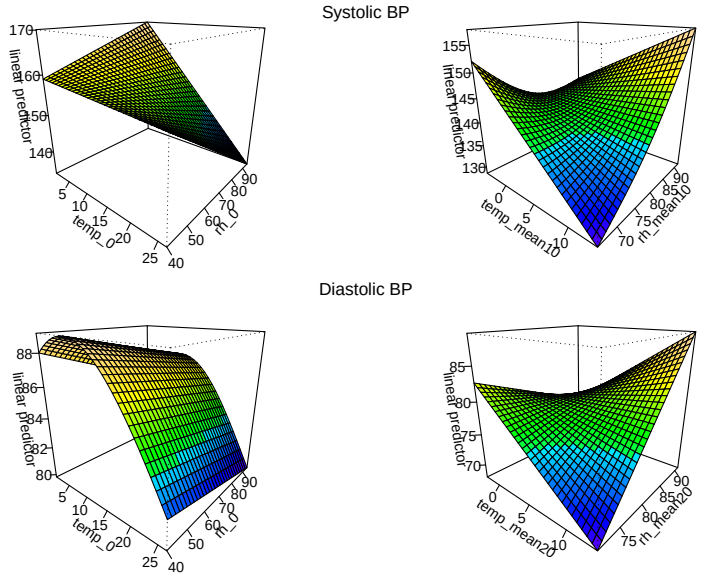


FIGURE 1. Bivariate tensor product of daily temperature and humidity. *Left panels*: Warmer months, *Right panels*: Colder months.

quantified the interaction effect between temperature and relative humidity using a bivariate tensor spline. To account for the repeated measurement structure, we incorporated the random effects. As various studies indicate the importance of seasonal variations in blood pressures, we subset our study population into warmer and colder months (see Rosenthal, T., 2004). Our data suggest that individual-specific covariates such as age and location influence blood pressures. We also see a significant effect of the interaction between climatic factors on blood pressure. Our finding of systolic and diastolic blood pressures modelled best with climatic data taken on the day of examination for warmer days and longer lags for colder days confirms earlier research studies (Brook et al., 2011). However, note that our study has a few limitations. First, the blood pressures have different scales of measurement for each follow-up. Although this seems to be a small-scale impact on the results, it may still induce some bias. Second, this study does not include the popular methods used in literature to detect interaction effects. For example, stratification of the climatic factors can give a more comprehensive and quantitative comparison of temperature and relative humidity effects on blood pressure. While GAMs are widely used in studying the effect of environmental factors on blood pressure, this study considers the interaction among these environmental factors through a bivariate tensor product spline.

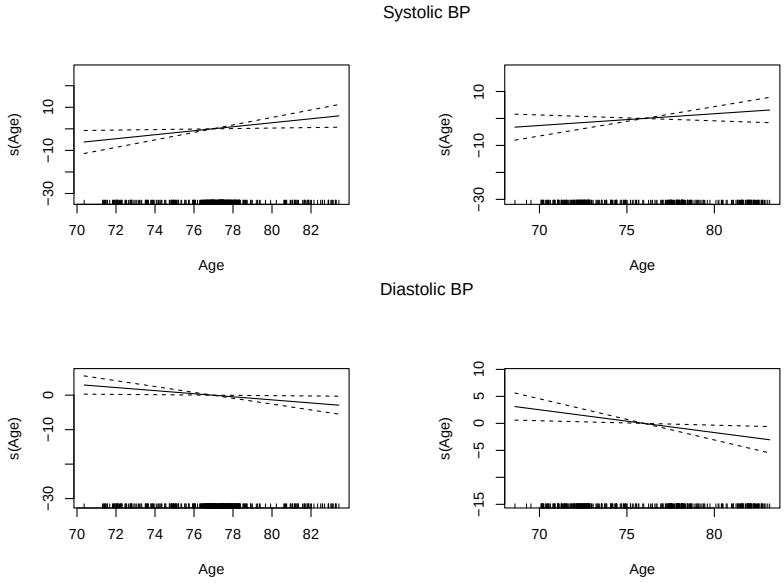


FIGURE 2. Relationship between age and blood pressures. *Left panels*: Warmer months, *Right panels*: Colder months. Estimated effects in solid lines with 95 % confidence intervals in dashed lines.

References

Barnett et al., (2007). The effect of temperature on systolic blood pressure. *Blood Pressure Monitoring*, **12**, 195 – 203.

Brook, R. D. (2017). The environment and blood pressure. *Cardiology clinics*, **35(2)**, 213-221.

Eilers, P.H and Marx, B.D (2021). *Practical Smoothing: The Joys of P-splines*. Cambridge: Cambridge University Press.

Pinto E. (2007). Blood pressure and ageing. *Postgraduate medical journal*, **83(976)**, 109–114.

Ravindra et al., (2019). Generalized additive models: Building evidence of air pollution, climate change and human health). *Environment International*, **132**, 104987.

Rosenthal, T. (2004). Seasonal variations in blood pressure. *The American journal of geriatric cardiology*, **13(5)**, 267-272.

Wood, S. N. (2006). *Generalized additive models: an introduction with R.* chapman and hall/CRC.

Univariate and Multivariate Adaptive Reference Ranges for Longitudinal Monitoring

Davood Roshan[1,2], John Ferguson[3], Charles R. Pedlar[4], John Newell[1,2]

[1] School of Mathematical and Statistical Sciences, National University of Ireland, Galway
[2] CURAM, SFI Research Centre for Medical Devices, National University of Ireland, Galway
[3] HRB Clinical Research Facility, National University of Ireland, Galway
[4] Faculty of Sport, Health and Applied Science, St Mary's University, Twickenham, United Kingdom

E-mail for correspondence: Davood.roshansangachin@nuigalway.ie

Abstract: In clinical settings, population-based reference ranges (also known as normal or static reference ranges) are typically used when interpreting a biomarker result for an individual by classifying their value as typical or atypical. In this paper, we propose the use of personalised adaptive reference ranges when biomarkers are collected longitudinally for an individual. The method is illustrated by an analysis of data collected longitudinally from elite athletes.

Keywords: Biomarker; Longitudinal monitoring; Reference ranges, Adaptive multivariate reference region, Mixed effect models.

1 Introduction

In a clinical setting, biomarkers are typically measured and evaluated as biological indicators of a physiological state. A reference range, generated from a cross-sectional analysis of healthy individuals free of the condition of interest, is typically used when interpreting a set of biomarker test results for a particular individual. An arbitrary percentile cut-point (typically the 95th or 97.5th percentile) is chosen to define abnormality. In practice, it is quite important to estimate such reference ranges appropriately. For instance, the clinical and biological assessment of individuals is usually based

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

on longitudinal monitoring of their biomarkers. In such cases, the individual has started to generate their own reference values and the interest is on identifying meaningful changes in these values. Therefore, reference ranges which adapt to account for both between and within individual variability are needed for an effective monitoring. To rectify the issue, we propose the use of personalised adaptive reference ranges in which reference ranges adapt successively whenever a new measurement is recorded for the individual. The approach undertaken includes a random intercept model and will be illustrated by applying to real biomarker data collected longitudinally from athletes in elite sport. We will also discuss the development of multivariate adaptive reference regions, as an extension of our proposed adaptive reference ranges when the interest is to assess an individual physiology using multiple biomarkers over time.

2 Development of Adaptive Reference Ranges

Consider I independent individuals, with individual i consisting of n_i independent normally distributed measurements, $y_{ij};\ j = 1, ..., n_i$, each with an unknown mean μ_i and an unknown variance σ_i^2. A random intercept model then will be defined for the biomarkers' values as:

$$y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = 1, ..., I,\ j = i, ..., n_i \tag{1}$$

where μ_is are the subject level random intercepts following a normal distribution as $\mu_i \sim N(\mu, \tau^2)$ in which μ represents an overall mean and τ^2 indicates the between individual variability. The error term indicated by ϵ_{ij} represents the within individual variability and is assumed to be independent of μ_i and also normally distributed, i.e. $\epsilon_{ij} \sim N(0, \sigma_i^2)$. For convenience, we set $\boldsymbol{\gamma}_i = (\mu, \tau^2, \sigma_i^2)$.
Measurement j from subject i, y_{ij}, will be considered as 'atypical' if it falls beyond the $\frac{\alpha}{2} * 100\%$ and $(1 - \frac{\alpha}{2}) * 100\%$ quantiles of the distribution: $P(y_{ij}|y, \boldsymbol{\gamma}_i)$; where y refers to the both subject i measurements prior to time j (i.e. $y_{i1}, ..., y_{ij-1}$) and other historical information from other individuals (i.e. $y_{i'j};\ i' \neq i\ \&\ j = 1, ..., n_{i'}$). Due to the (assumed) independence of individuals, $P(y_{ij}|y, \boldsymbol{\gamma}_i)$ can be written in the form $P(y_{ij}|y_{i1}, ..., y_{ij-1}, \gamma_i) \equiv P(y_{ij}|\bar{y}_i^{j-1}, \gamma_i)$; where $\bar{y}_i^{j-1}$ is the average value of the biomarker measurements for subject i before time j. It can be shown that the distribution: $P(y_{ij}|\bar{y}_i^{j-1}, \boldsymbol{\gamma}_i)$ can be found through (2) as:

$$y_{ij}|\bar{y}_i^{j-1}, \boldsymbol{\gamma}_i \sim N\left(\frac{\frac{\mu}{\tau^2} + \frac{(j-1)\bar{y}_i^{j-1}}{\sigma_i^2}}{\frac{1}{\tau^2} + \frac{j-1}{\sigma_i^2}},\ \frac{1}{\frac{1}{\tau^2} + \frac{j-1}{\sigma_i^2}} + \sigma_i^2\right), \tag{2}$$

An approximate Expectation-Maximisation (EM) algorithm (Ippel L, 2016), that relies on a few summary statistics that contain all the required

information from previous observations, will be used to find the estimates of the model parameters at time j (see Roshan et. al, 2021). Once the parameters are estimated, the $100(1-\alpha)\%$ adaptive reference ranges for subject i at time j will be generated as:

$$\frac{\frac{\hat{\mu}}{\hat{\tau}^2}+\frac{(j-1)\bar{y}_i^{j-1}}{\hat{\sigma}_i^2}}{\frac{1}{\hat{\tau}^2}+\frac{j-1}{\hat{\sigma}_i^2}} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{\frac{1}{\hat{\tau}^2}+\frac{j-1}{\hat{\sigma}_i^2}}+\hat{\sigma}_i^2} \tag{3}$$

3 Longitudinal monitoring of elite runners

The longitudinal monitoring of oxidative stress (OS) are now widely used in elite sports to inform the identification of fatigued states and underperformance in athletes. The Free Oxygen Radical Test (FORT) and the Free Oxygen Radical Defence test (FORD) are two Point of Care (POC) tests that indirectly provide an index of OS (Lewis, Nathan A., et al. 2015). Such test results were collected from 11 elite distance runners to identify meaningful changes in their test results over their training period. Interpretation of changes in FORT and FORD test results in practice is typically based on a comparison against a population-based static reference range. However, for a more accurate assessment, we initially developed personalised adaptive reference ranges for the two longitudinally recorded FORT and FORD test results for a particular athlete (see Figure 1). Bivariate Adaptive Reference Regions were also generated to account for the joint assessment of the two test results (see Figure 2). As can be seen from Figure 1, the developed adaptive reference ranges are considerably narrower than the population-based static reference ranges suggesting that they are more sensitive to any changes. For instance, although the static reference ranges identified all of the test results as typical for that individual, the adaptive methods identified the 7th FORT test result as atypical which may need further consideration. Figure 2 on the other hand shows the developed bivariate reference regions for two FORT and FORD test results. As it can be seen the suspected measurement at time point 7 is still outside the joint reference region which provides further evidence of a potential abnormality for the athlete in question at this time point which should be investigated further.

To construct the above mentioned bivariate adaptive reference regions we extended the mixed effect model proposed in 1 where:

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix} \sim MN\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, D\right); \qquad D = \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{pmatrix}$$

$$\epsilon_{ij} = \begin{pmatrix} \epsilon_{ij1} \\ \epsilon_{ij2} \end{pmatrix} \sim MN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right); \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

$$y_{ij} = \begin{pmatrix} y_{ij1} \\ y_{ij2} \end{pmatrix} \sim MN\left(\begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \Sigma\right)$$

In the same fashion, the pair of measurements j from subject i, $y_{ij} = \begin{pmatrix} y_{ij1} \\ y_{ij2} \end{pmatrix}$ will be considered as 'atypical' if it falls beyond the $\frac{\alpha}{2} * 100\%$ and $(1 - \frac{\alpha}{2}) * 100\%$ quantiles of the distribution: $P(y_{ij}|y, \boldsymbol{\gamma}_i)$; where y refers to the both subject i measurements prior to time j and other available measurements from other individuals and $\boldsymbol{\gamma}_i$ is also a set of all model parameters (i.e. $m_1, m_2, \tau_{11}, \tau_{12}, \tau_{21}, \tau_{22}, \sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22}$).

4 Conclusion

In this study, methods for developing personalized adaptive reference ranges for longitudinally recorded clinical biomarkers were presented with an intention to help researchers and physicians make more reliable decisions in terms of what can be considered as normal physiology for an individual. The results have shown that the proposed adaptive methods are capable of triggering 'alerts' and can be used as an early warning system that warrant further attention and review. We further extended our model to circumstances where more than one biomarker is of interest by developing multivariate adaptive reference regions.

Acknowledgments: This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) and the European Regional Development Fund (ERDF) under grant number 13/RC/2073_P2.

References

Ippel L, Kaptein MC, Vermunt JK (2016). *Estimating random-intercept models on data streams*. Computational Statistics & Data Analysis 104: 169-182.

Lewis, Nathan A., et al. (2015). *Alterations in redox homeostasis in the elite endurance athlete*. Sports medicine, 45(3), 379-409.

Roshan, Davood, et al. (2021). *A comparison of methods to generate adaptive reference ranges in longitudinal monitoring*. Plos one 16.2, e0247338

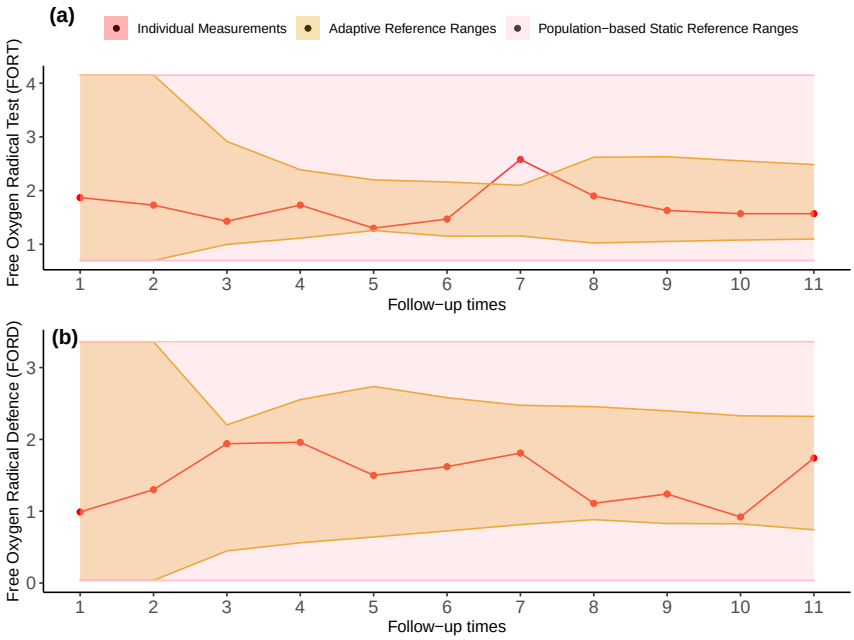


FIGURE 1. Static and adaptive reference ranges for (a) FORT and (b) FORD test results for a female athlete.

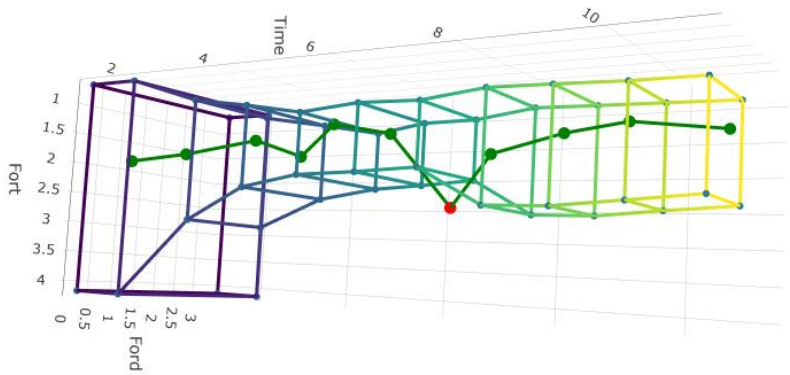


FIGURE 2. Bivariate adaptive reference regions for the two FORT and FORD test results for a female athlete.

Learning the structure of the mTOR protein signalling pathway

Abdul Salam[12], Marco Grzegorczyk[1]

[1] Bernoulli Institute, Groningen University, Groningen, Netherlands
[2] Department of Statistics, University of Malakand, Chakdara, Dir Lower, KP, Pakistan

E-mail for correspondence: m.a.grzegorczyk@rug.nl

Abstract: We propose a new globally coupled hierarchical dynamic Bayesian network (DBN) model for learning the structure of the mTOR protein signalling pathway from immunoblotting protein phosphorylation data. We follow an hierarchical Bayesian modelling approach, as data were collected under two different experimental conditions. Moreover, since protein phosphorylation measurements were taken at non-equidistant time points, we propose to use smoothing splines and Gaussian processes for predicting the missing equidistant values.

Keywords: Network learning, mTOR pathway, non-equidistant measurements.

1 Introduction

Dynamic Bayesian networks (DBNs) and their non-homogeneous hierarchical extensions are popular statistical models for learning the structures of cellular networks, such as gene regulatory networks and protein signalling pathways from time series data. In our study we aim to learn the structure of the mammalian target of rapamycin complex 1 pathway (mTOR) protein signalling pathway from immunoblotting protein phosphorylation data. The mTOR signalling pathway plays a fundamental rule in regulating and controlling many important cellular processes, such as cell growth and proliferation.
After two experimental treatments $k = 1$ (without insulin) and $k = 2$ (with insulin) $N = 11$ phosphorylation sites of eight key proteins were measured after $t = 0, 1, 3, 5, 10, 15, 30, 45, 60, 120$ minutes. On the one hand, since insulin is known to have an effect on the intensities of some of the protein interactions, the two time series cannot be merged and be analysed in

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

one batch. On the other hand, both time series are potentially too short to be analyzed separately. As a consensus, we propose a new hierarchical Bayesian model with interaction-specific coupling (strength) parameters. For each individual protein-protein interaction our new model features a specific coupling (strength) parameter that regulates the similarity of this particular interaction across the two conditions. A low coupling parameter enforces the interaction to stay (rather) constant, while a large coupling parameter allows the interaction intensity to change among conditions. Our new model improves upon the globally coupled hierarchical DBN model from Grzegorczyk and Husmeier (2013). It allows for more flexibility by introducing interaction-specific coupling parameters rather than letting all regulators of a regulatee share the same coupling parameter. To derive the model we borrow modelling ideas from Shafiee Kamalabad and Grzegorczyk (2021), which introduced a sequentially coupled hierarchical DBN with segment-specific coupling parameters. We adapt the conceptual idea to define a new globally coupled hierarchical DBN with covariate-specific coupling parameters.
We also show that interpolation methods (smoothing splines and Gaussian processes) can predict the required missing equidistant protein values and that this improves upon the widely applied but rather naive approach to treat observed non-equidistant data points as if they were equidistant.

2 Methods

Consider two linear regression models that share the same response, the same m covariates and the same noise variance parameter, $\sigma^2 > 0$, but have different regression coefficient vectors $\boldsymbol{\beta}_k \in \mathbb{R}^m$ $(k = 1, 2)$:

$$\mathbf{y}_k|(\boldsymbol{\beta}_k, \sigma^2) \sim \mathcal{N}(\mathbf{X}_k\boldsymbol{\beta}_k, \sigma^2\mathbf{I}) \qquad (k = 1, 2)$$

where $\mathbf{y}_k \in \mathbb{R}^{n_k}$ and $\mathbf{X}_k \in \mathbb{R}^{n_k,m}$ are the k-th response vector and design matrix, respectively, and $\sigma^{-2} \sim GAM(\alpha_\sigma, \beta_\sigma)$. We introduce the following new hierarchical prior:

$$\begin{aligned} \boldsymbol{\beta}_k|\left(\sigma^2, \boldsymbol{\lambda}^2, \boldsymbol{\mu}\right) &\sim \mathcal{N}_m\left(\boldsymbol{\mu}, \sigma^2\text{diag}(\boldsymbol{\lambda}^2)\right) \qquad (k = 1, 2) \\ \boldsymbol{\mu} &\sim \mathcal{N}_m(\mathbf{0}, \mathbf{I}) \end{aligned}$$

where $\boldsymbol{\lambda}^2 := (\lambda_1^2, \ldots, \lambda_m^2)$ is a vector of m covariate-specific coupling parameters, and $\text{diag}(\boldsymbol{\lambda}^2)$ is a diagonal matrix with the elements of $\boldsymbol{\lambda}^2$ on the diagonal. On the coupling parameters we impose another hierarchical prior:

$$\begin{aligned} \lambda_l^{-2}|\tilde{b} &\sim GAM(\tilde{a}, \tilde{b}) \qquad (l = 1, \ldots, m) \\ \tilde{b} &\sim GAM(\alpha, \beta) \end{aligned}$$

For the density of the posterior distribution we then have:

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\lambda}^2, \tilde{b}, \sigma^2, \boldsymbol{\mu} | \mathbf{y}_1, \mathbf{y}_2) \propto \left(\prod_{k=1}^{2} p(\mathbf{y}_k | \boldsymbol{\beta}_k, \sigma^2) p(\boldsymbol{\beta}_k | \sigma^2, \boldsymbol{\lambda}^2, \boldsymbol{\mu}) \right) \cdot p(\boldsymbol{\mu}) \cdot p(\sigma^2) \cdot \left(\prod_{l=1}^{m} p(\lambda_l^2 | \tilde{b}) \right) \cdot p(\tilde{b})$$

For a known covariate set, Gibbs sampling can be used to generate posterior samples of the model parameters. All full conditional distributions can be computed analytically. When the true covariate set is unknown, additional Reversible Jump Markov Chain Monte Carlo (RJMCMC) moves on the covariate set can be implemented to also average across all possible covariate sets. We implement covariate addition, deletion and exchange moves to allow the covariate set (and so the dimensionality of the parameter space) to vary during the simulation. For lack of space we here cannot provide the technical details.

A conventional assumption for dynamic Bayesian networks (DBNs) is that all interactions are subject to a time lag of one time unit. The task of learning a dynamic network among variables $Z_1, \ldots, Z_N$ can then be subdivided into N separate regression tasks. In the n-th regression model Z_n takes the role of the response and the t-th observation of Z_n is explained by the values of the other variables (covariates) at the preceding time point $t-1$. Sampling covariate sets for Z_n then refers to sampling the regulators of Z_n. The score of each possible network interaction $Z_i \rightarrow Z_j$ is the relative frequency with which the sampled covariate sets for Z_j contained Z_i.

3 Results

DBN models assume that data points have been measured at equidistant time points ($t = 1, 2, 3, \ldots$), but the mTOR measurements are non-equidistant ($t_k \in \{0, 1, 3, \ldots, 60, 120\}$). Often this critical mismatch between model and data is ignored, treating the temporally ordered data points as if they were equidistant ('simple-shift'). We instead propose to predict the missing equidistant data using (smoothing) splines and Gaussian processes (GPs), and we cross-compare the methods in terms of predictive probabilities (PPs) obtained by leave-one-out cross validation (LOOCV-PPs). Figure 1 shows the LOOCV-PPs results for the widely applied naive simple shift approach and the two proposed interpolation methods. It can seen that the simple shift yields much worse results than the two interpolation methods. Moreover, (smoothing) splines perform slightly (but significantly) superior to Gaussian processes (GPs).

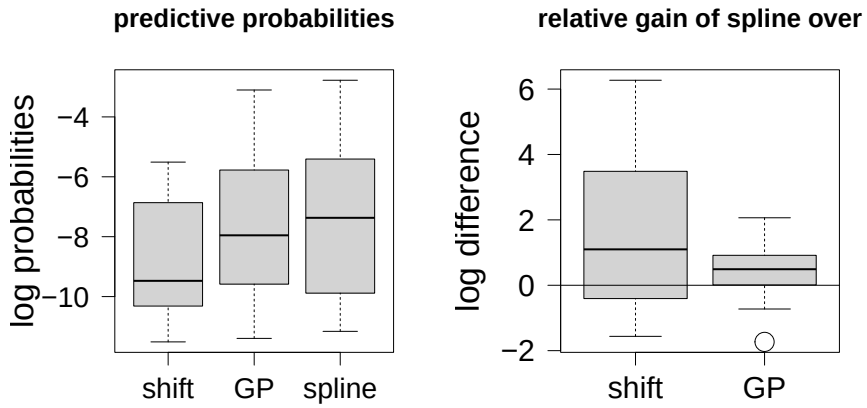


FIGURE 1. **Comparing the three interpolation methods.** All results were obtained with the newly proposed hierarchical DBN; cf. Section 2. Left: Boxplots of LOOCV-PPs for the naive shift and the two proposed interpolation methods (GP and splines). Right: Boxplots of the relative differences in favour of splines.

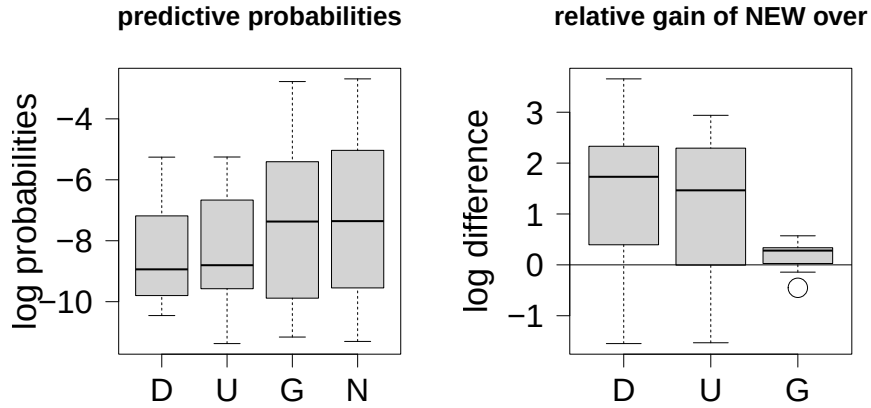


FIGURE 2. **Comparing the four (hierarchical) DBN models.** All results were obtained on the smoothing splines interpolated data set. Left: Boxplots of the logarithmic LOOCV-PPs for **D**BN ,**U**NCLOUPLED, **G**LOBAL and **N**EW. Right: Boxplots of the relative differences in favour of **N**EW.

On the smoothing splines interpolated data set we compare the newly proposed model (**N**) with a standard DBN (**D**) that ignores the conditions and merges both time series, an uncoupled model (**U**) which learns independent regression vectors for both conditions, and the less flexible globally coupled hierarchical DBN (**G**) from Grzegorczyk and Husmeier (2013), which couples all covariates with the same coupling strength.

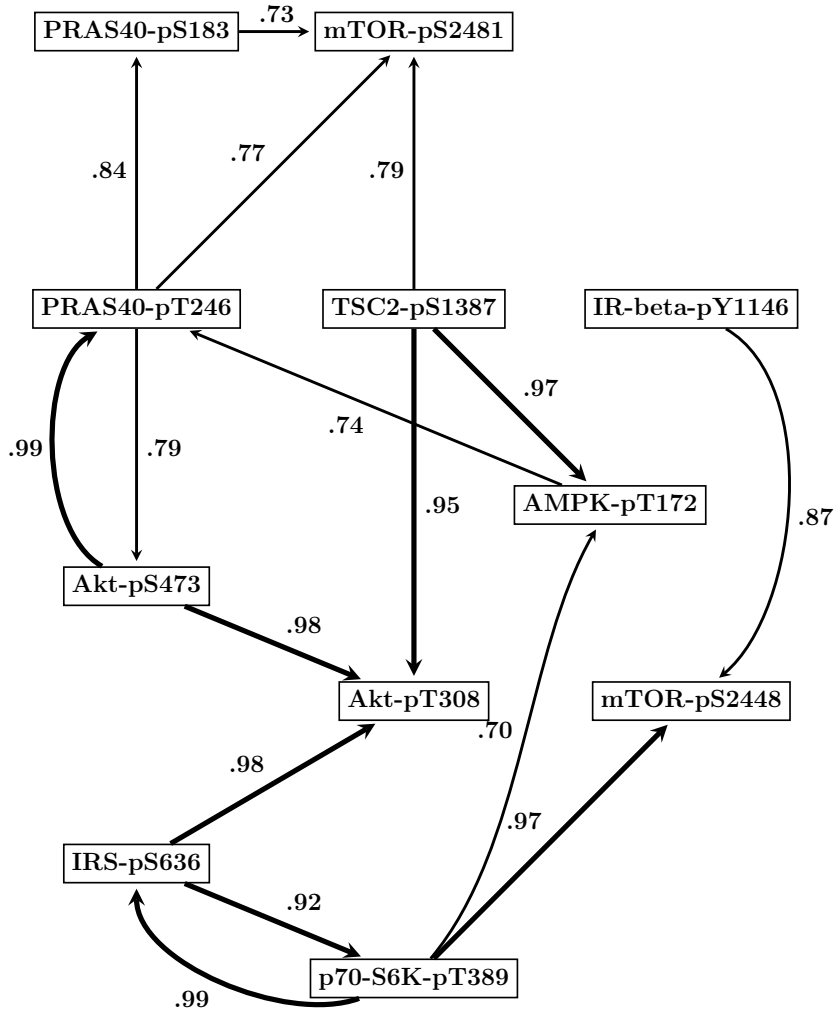


FIGURE 3. **Predicted structure of the mTOR protein signalling pathway.** Results were obtained with the newly proposed hierarchical DBN model (cf. Section 2) on the smoothing splines interpolated data set. There are $N = 11$ nodes (phosphorylation sites) and the 16 edges (protein interactions) whose scores exceeded 0.7 are shown. Edges whose scores exceeded 0.9 are in bold. The edge scores refer to the relative frequencies with which the corresponding interactions (i.e covariate-response relations) were posterior sampled within our RJMCMC based model averaging approach.

Figure 2 shows the LOOCV-PP results for this DBN model comparison. It can be seen that the two extreme approaches, merging all data (**D**) or allowing for independent interaction parameters (**U**), yield relatively bad results. The two globally coupled models yield better results with the newly proposed refined model (**N**) improving slightly (but significantly) upon the original globally coupled DBN (**G**).

Based on the empirical results of our comparative evaluation study (cf. Figures 1-2), we apply the new hierarchical DBN model (**N**) to the smoothing splines interpolated data set to learn the structure of the mTOR protein signalling pathway. Figure 3 shows the predicted mTOR pathway.

4 Conclusions

To predict the structure of the mTOR protein signalling pathway from immunoblotting data we have proposed a new refined globally coupled hierarchical DBN model and we have used two established interpolation methods to estimate missing equidistant data. The results of a first comparative evaluation study (see Figure 1) suggest that smoothing splines (and Gaussian processes) can be used to interpolate missing equidistant data values and that these interpolation methods lead to better results than the naive simple shift approach that treats non-equidistant data as if they were equidistant. A second comparative evaluation study (see Figure 2) revealed that the new model performs better than three earlier proposed (hierarchical) DBN models. In particular the new model improves upon an earlier proposed globally coupled hierarchical model. Finally we have applied the new DBN model on the smoothing splines interpolated data set to learn the structure of the mTOR signalling pathway. The predicted structure of the mTOR protein signalling pathway is shown in Figure 3.

References

Grzegorczyk, M. and Husmeier, D. (2013). Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models. *Machine Learning*, **91**, 105 – 154.

Shafiee Kamalabad, M. and Grzegorczyk, M. (2021). A new Bayesian piecewise linear regression model for dynamic network reconstruction. *BMC Bioinformatics*, **22**, Article number: 196.

Tree-based Modeling of Confounding Effects in Matched Case-Control Studies

Gunther Schauberger[1], Luana Fiengo Tanaka[1], Moritz Berger[2]

[1] Chair of Epidemiology, Department of Sport and Health Sciences, Technical University of Munich, Germany
[2] Institute of Medical Biometry, Informatics and Epidemiology, University of Bonn, Germany

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

Abstract: Conditional logistic regression is the standard approach for the analysis of matched case-control studies. This parametric model consists of fitting an additive linear predictor function, which focuses on main effects. An alternative tree-based method is proposed that allows for more flexibility, in particular when non-linear effects and interactions between confounding variables are present. The method is illustrated by a case-control study on cervical cancer.

Keywords: CART; Conditional logistic regression; Matched pairs.

1 Parametric Conditional Logistic Regression

Consider a case-control study conducted to determine the effect of an exposure on a binary outcome, for example, the presence of a specific disease. To guarantee balance of potential confounders between cases and controls one frequently applies matching, where for each case a certain number of controls is selected that coincide with regard to important factors (Mansournia et al., 2018). The classical method to analyse matched case-control data is the conditional logistic regression model of the form

$$\log\left(\frac{\mathrm{P}(y_{ij}=1)}{\mathrm{P}(y_{ij}=0)}\right) = \alpha_i + \mathbf{z}_{ij}^T\boldsymbol{\gamma}, \quad i = 1,\ldots,n,\ j = 1,\ldots,m_i\,, \qquad (1)$$

where $y_{ij} \in \{0,1\}$ denotes the binary outcome and $\mathbf{z}_{ij}$ is a set of covariates with coefficient vector $\boldsymbol{\gamma}$. The observations come in n clusters $s_i,\ i = 1,\ldots,n$, of size m_i, which is accounted for in the model via cluster-specific intercepts α_i. In matched case-control studies, it holds that $y_{ij} = 1$

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

for cases and $y_{ij} = 0$ for controls and that the clusters are defined by the matching strata. The vector $\mathbf{z}_{ij}$ collects the exposure variable and all further confounding variables of observation j in cluster i.
We assume each of the n strata to contain only one case, i.e. $\sum_{j=1}^{m_i} y_{ij} = 1$. For simplicity, in each stratum we assume the first observation to be the case, i.e. $y_{i1} = 1$ for $i = 1, \ldots, n$. Estimates of the parameters in (1) are then derived by maximizing the conditional log-likelihood

$$l_c(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left(\mathbf{z}_{i1}^T \boldsymbol{\gamma} - \log \left(\sum_{j=1}^{m_i} \exp(\mathbf{z}_{ij}^T \boldsymbol{\gamma}) \right) \right) ,$$

where the stratum-specific intercepts α_i are eliminated from the likelihood by conditioning on the number of cases per stratum, see Breslow and Day (1980) for further details.

2 Tree-based Conditional Logistic Regression

In the parametric model (1) it is assumed that the effect of the confounding variables on the outcome can be described by a linear predictor function. This, however, may be not appropriate when non-linear effects or interactions between the covariates occur in the data. To address this issue, we propose to implement the concept of recursive partitioning in the covariate space (Breiman et al., 1984) within the framework of conditional logistic regression. Thus, we allow for a very flexible confounding structure and take advantage of the fact that conditional logistic regression accounts for the matching structure.
The basic idea is to replace the linear predictor of model (1) by a decision tree, which yields a conditional logistic regression model of the form

$$log \left(\frac{P(y_{ij} = 1)}{P(y_{ij} = 0)} \right) = \alpha_i + f(\mathbf{z}_{ij}) \tag{2}$$

where $f(\mathbf{z}_{ij})$ describes the effect of the variables collected in $\mathbf{z}_{ij}$ represented by a tree. With $S_1, \ldots, S_t$ representing the terminal nodes of the tree, $f(\mathbf{z}_{ij})$ can be denoted as

$$f(\mathbf{z}_{ij}) = \delta_1 I(\mathbf{z}_{ij} \in S_1) + \ldots + \delta_t I(\mathbf{z}_{ij} \in S_t) \, .$$

All nodes are determined by a product of indicator functions. For example, if the splits were in the metric variables z_1 and z_4 a node may be determined by $I(\mathbf{z}_{ij} \in S) = I(z_{ij1} > 20) \, I(z_{ij4} \leq 10)$.
Starting from an initial model containing the strata-specific intercepts, only, we gradually search for binary splits in the covariates collected in $\mathbf{z}$ that further improve the model fit. A split divides the current node into two child nodes, and is incorporated into the model using a corresponding indicator

variable. For a metric or ordinal variable, the model after the first split at threshold c in variable z_k has the form

$$log\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = \alpha_i + \delta I(z_{ijk} \leq c)\,,$$

where $I(\cdot)$ is the indicator function. Successively, the variable and the split point are chosen that lead to the highest improvement of the conditional log-likelihood of the model.
It should be noted that each split has an effect on all remaining terminal nodes, because all parameter estimates change if an additional split is performed. This is in contrast to the way trees are grown in traditional recursive partitioning where the remaining terminal nodes are not affected by a new split.
We propose to directly control the size of the tree by early stopping. To do so, one examines all the null hypotheses $H_0 : \delta = 0$ and $H_1 : \delta \neq 0$ (of the newly added parameter) and selects the split with the maximal LR test statistic (minus two times the difference in the conditional log-likelihood values). To decide whether the selected split should be performed, we apply a concept based on maximally selected statistics (Hothorn and Lausen, 2003). The basic idea is to investigate the dependence of the binary outcome and the selected variable at a global level that takes the number of split points into account. For one fixed variable z_k, one simultaneously considers all LR test statistics T_{kc_k}, where c_k are from the set of possible split points, and computes the maximal value statistic $T_k = \max_{c_k} T_{kc_k}$. The p-value that can be obtained from the distribution of T_k provides a measure for the relevance of variable z_k. As the distribution of T_k is unknown we use a permutation test (with significance level α/p, where p denotes the number of covariates) to obtain a decision on the null hypothesis. By computing the maximal value statistics for a large number of permutations of variable z_k one obtains an approximation of the distribution under the null hypothesis and the corresponding p-value. If significant, the selected split is performed, otherwise the algorithm is terminated.
The tree algorithm may lead to partitions in the covariate space where perfect discrimination between cases and controls is reached. Therefore, to guarantee the existence of all parameter estimates, we allow for an additional refitting step after the final tree has been determined. This refitting step employs regularization using a small L_2 penalty to stabilize the tree parameter estimates $\delta_1, \ldots, \delta_t$. Optimization is then applied to the penalized log-likelihood

$$l_p(\cdot) = l_c(\cdot) + \lambda \sum_{o=1}^{t} \delta_o^2\,,$$

where λ is a tuning parameter set to a small value. A standard value we used was $\lambda = 10^{-20}$.

3 Separating the Exposure Variable

In most matched case-control studies one is specifically interested in the association (or rather the causal effect) of the exposure variable (in the following denoted by x) on the outcome. A tree-based conditional logistic regression model that separates the exposure effect from the remaining covariates can be denoted as

$$log\left(\frac{P(y_{ij}=1)}{P(y_{ij}=0)}\right) = \alpha_i + x_{ij}\beta + f(\mathbf{z}_{ij})\,, \tag{3}$$

where $f(z_{ij})$ describes the effect of the covariates represented by a tree and β is the regression coefficient (interpretable as the adjusted log odds-ratio) of the exposure. Tree building and estimation of the parameters in (3) can be performed in the same way as described in the previous section.
To derive a confidence interval for the exposure effect, one cannot directly use the confidence interval provided by the underlying conditional logistic model as it will not take the selection process into account when building the tree. Accordingly, these confidence intervals would underestimate the variability of the point estimate. Therefore, we propose to calculate confidence intervals via a nonparametric bootstrap approach. A bootstrap sample is generated by sampling n strata with replacement from the original data. For each bootstrap sample, we fit the model and save the corresponding estimate for the exposure effect. From a total of B estimates we then calculate the corresponding empirical quantiles as the boundaries of the confidence interval.

4 TeQaZ Study

For illustration, the proposed method is applied to a real data example. The data originate from the so-called TeQaZ study (Tanaka et al., 2021), a case-control study on cervical cancer. The main focus of the study was to examine the effect of frequent participation in cervical cancer screening (CCS) on the odds of cervical cancer. The exposure variable *CCS* was defined as *frequent participation*, if women had attended CCS at least every three years within the past ten years, including at least once in the three years preceding diagnosis. Matching was done using age and residence area where controls were matched to cases only if they lived in the same area and if their age was at maximum 2 years younger or older than the age of the corresponding case. In total there were 14 potential confounding variables. Additionally, we calculated the variable *Age.Diff* defined as the difference between each individuals age and the average age in the respective stratum. Exclusion of the observations with missing values in any of the variables resulted in an analysis data set with 170 cases and 425 controls.
Figure 1 shows the resulting tree when fitting model (2) without a separate exposure effect. Interestingly, the first chosen split is in the exposure

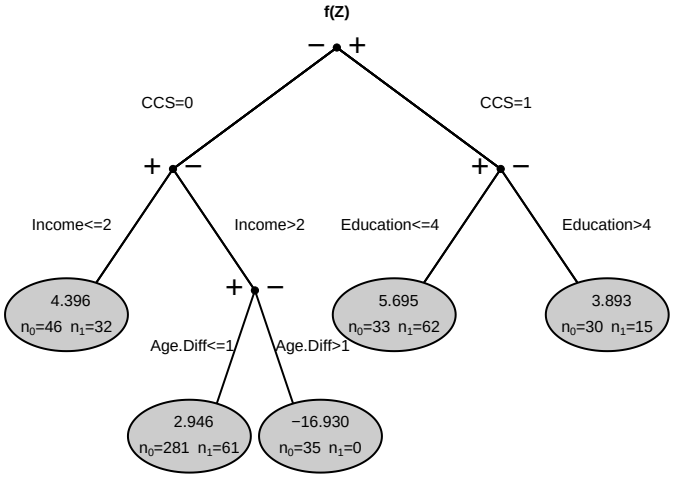


FIGURE 1. Analysis of the TeQaZ study. Tree for the model without a separate CCS exposure effect. For all terminal nodes, the respective parameter estimates and the numbers of cases n_1 and controls n_0 are displayed.

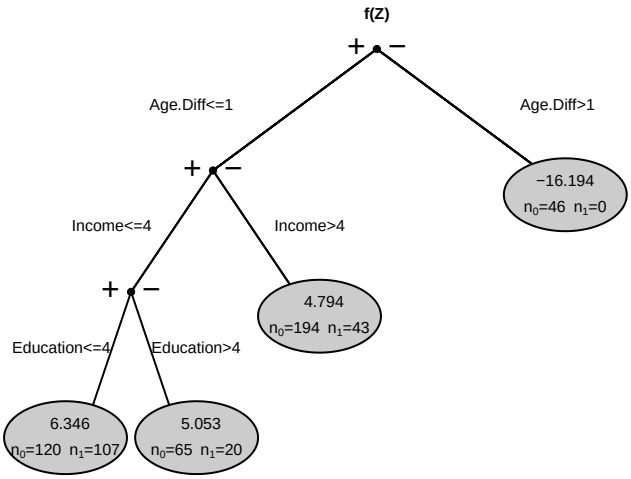


FIGURE 2. Analysis of the TeQaZ study. Tree for the model with a separate CCS exposure effect. For all terminal nodes, the respective parameter estimates and the numbers of cases n_1 and controls n_0 are displayed.

variable *CCS*, where infrequent participation (i.e. $CCS = 1$) is associated to higher odds for cervical cancer than frequent participation. Further splits are done in the variables *Education* (for $CCS = 1$), and *Income* and *Age.Diff* (for $CCS = 0$).

Figure 2 shows the resulting tree of the model where CCS is incorporated in a separate exposure effect. Here, the first split is done for *Age.Diff*, leading to a terminal node with perfect separation ($Age.Diff > 1$) and a very small parameter estimate. This split can be seen as a further age adjustment additional to the effect of age-matching accounting for potential residual age differences. For $Age.Diff \leq 1$ (the node where most of the observations fall into), further splits are performed in *Income* and *Education*, leading to an interaction between these variables. It is seen that higher income and educational level seem to act as protective factors for cervical cancer. The adjusted odds ratio for *CCS* with 95% confidence interval was estimated to be 5.878 [4.119; 28.202], again indicating higher odds for cervical cancer in case of infrequent participation.

References

Breiman, L., Friedman, J.H., Olshen, R.A., et al. (1984). *Classification and Regression Trees*. New York: Routledge.

Breslow, N.E., and Day, N.E. (1980). *Statistical Methods in Cancer Research: The Analysis of Case-Control Studies*. Lyon: IARC.

Hothorn, T., and Lausen, B. (2003) On the exact distribution of maximally selected rank statistics. *Computational Statistics and Data Analysis*, **43**:121-137.

Mansournia, M.A., Jewell, N.P., et al. (2018). Case-control matching: effects, misconceptions, and recommendations. *European Journal of Epidemiology*, **33**(1):5-14.

Tanaka, L.F., Schriefer, D., Radde, K., et al. (2021). Impact of opportunistic screening on squamous cell and adenocarcinoma of the cervix in Germany: A population-based case-control study. *PLOS ONE*, **16**(7):e0253801.

Graphical Assessment of Probabilistic Precipitation Forecasts

Reto Stauffer[1 2], Moritz N. Lang[1], Achim Zeileis[1]

[1] Department of Statistics, Universität Innsbruck, Austria
[2] Digital Science Center, Universität Innsbruck, Austria

E-mail for correspondence: Reto.Stauffer@uibk.ac.at

Abstract: Accurate and reliable probabilistic predictions have been becoming more and more important over the last decades and they are an essential tool for proper risk assessment and strategic planning. In order to provide full probabilistic forecasts, distributional regression models are frequently used. Such models range from basic generalized linear models (GLM) over generalized additive models (GAM) to generalized additive models for locate, scale, and shape (GAMLSS) and other types of refined distributional regression models.
For assessing the goodness of fit of such probabilistic regression models, graphical assessment techniques are an important complement to proper scoring rules and help to identify possible model misspecifications. Based on a case study of probabilistic precipitation forecasts, three different model specifications are evaluated graphically to reveal different sources of misspecification such as censoring at zero, hereoscedasticity, and heavy tails. The graphics either evaluate marginal calibration by comparing observed and fitted frequencies using variations of so-called rootograms. Or alternatively they assess probabilistic calibration by evaluating the distribution of the probability integral transform (PIT) on different scales using histograms or variations of quantile-quantile plots. Relative strengths and weaknesses in revealing the sources of misfit are highlighted.
A unified implementation is provided in the newly developed *R* package *topmodels* (https://topmodels.R-Forge.R-project.org/).

Keywords: Graphical model assessment; Distributional regression

1 Case study

Weather forecasts are typically generated by physically-based numerical weather prediction models. To account for uncertainty, multiple forecasts are created with slightly modified conditions which build an ensemble. This

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

allows to not only retrieve information about the expected amount of precipitation but also the associated uncertainty. To better calibrate these raw ensemble forecasts, statistical post-processing is typically applied.
Revisiting three model specifications considered by Messner, Mayr, and Zeileis (2010), we use different graphical model assessment techniques for identifying possible model misspecifications and thus aiding the selection of a well-calibrated model.
The data used contains observed accumulated precipitation amounts for Innsbruck and the corresponding ensemble mean (ensmean) and ensemble standard deviation (enssd) of total accumulated precipitation amounts between 5 and 8 days in advance. Following previous studies, the square root of precipitation is used which has been shown to improve the calibration.
For comparison, three models are employed. A homoscedastic Gaussian linear regression model which does not properly account for the non-negative nature of precipitation, and two heteroscedastic regression models, left-censored at zero – one assuming a Gaussian and one a logistic underlying response distribution.

Distribution	Location	Scale
$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$	$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{ensmean}_i$	$\log(\hat{\sigma}_i) = \hat{\gamma}_0$
$y_i \sim \mathcal{N}_0(\mu_i, \sigma_i^2)$	$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{ensmean}_i$	$\log(\hat{\sigma}_i) = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot \log(\text{enssd}_i)$
$y_i \sim \mathcal{L}_0(\mu_i, \sigma_i^2)$	$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{ensmean}_i$	$\log(\hat{\sigma}_i) = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot \log(\text{enssd}_i)$

2 Model assessment

According to the seminal work of Gneiting, Balabdaoui, and Raftery (2007), probabilistic forecasts aim to maximize the sharpness of the predictive distributions subject to calibration. Moreover, this can be further distinguished into marginal and probabilistic calibration.

2.1 Marginal calibration

Marginal calibration is generally concerned with whether the observed frequencies of the response variable y_i match the corresponding expected frequencies from the model. For continuous response variables, frequencies for intervals $(b_j, b_{j+1}]$ are considered based on breaks b_j $(j = 1, \ldots, k)$. The expected frequencies are computed based on the predicted cumulative distribution function (CDF) $F(\cdot|\hat{\theta})$ where $\theta = (\mu, \sigma)$ denotes the joint model parameters.

$$\text{observed}_j = \sum_{i=1}^{N} I(y_i \in (b_j, b_{j+1}]),$$

$$\text{expected}_j = \sum_{i=1}^{N} \left[F(b_{j+1}|\hat{\theta}_i) - F(b_j|\hat{\theta}_i)\right].$$

Figure 1 shows hanging rootograms for the two Gaussian models, where the observed frequencies are hanging from the expected ones. The marginal calibration is assessed on the observational scale, which allows a direct interpretation. The heteroscedastic Gaussian model clearly underfits zero precipitation amounts as it is not accounting for the observed point mass at zero. Additionally, a weak wavelike pattern indicates a slight overfitting of precipitation sums between 0 and 5 and an underfitting of precipitation above. In contrast, the heteroscedastic left-censored Gaussian model provides a fairly good marginal fit.
However, due to the aggregation over all individual predictive CDFs $F(y_i|\hat{\theta}_i)$ for $i = 1, \ldots, N$, a statement about a possible violation of the distributional assumption is not easily possible.

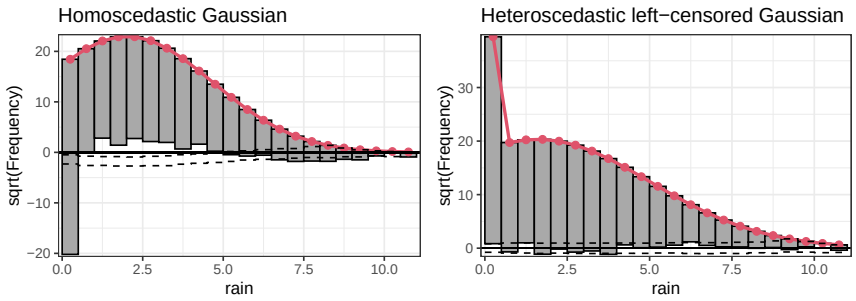


FIGURE 1. Hanging rootograms for the homoscedastic Gaussian model (left) and the heteroscedastic left-censored Gaussian model (right). Expected frequencies shown in red (line), observed frequencies in gray (bars).

2.2 Probabilistic calibration

Compared to the marginal calibration which is obtained on the observation scale, the probabilistic calibration is always performed on the probability scale by considering the probability integral transform (PIT) $u_i = F(y_i|\hat{\theta}_i)$. Additionally, this may need to be randomized for (partially) discrete observations to obtain uniformly distributed PIT values if the model is well calibrated. Alternatively, the PIT values can be mapped to other scales, e.g., by using the inverse of a standard normal CDF $\Phi(\cdot)$, to obtain (randomized) quantile residuals (r_i) as suggested by Dunn and Smyth (1996).

$$r_i = \Phi^{-1}(u_i) \quad \text{with} \quad u_i = \begin{cases} F(y_i|\hat{\theta}_i) & \text{if } F(\cdot) \text{ continuous} \\ U\left[F(y_i - 1|\hat{\theta}_i), F(y_i|\hat{\theta}_i)\right] & \text{if } F(\cdot) \text{ discrete} \end{cases}$$

Whether the distribution of u_i or r_i is uniform or normal, respectively, can then be checked using standard graphics like histograms or quantile-quantile (Q-Q) plots. As small too medium deviations can be quite hard to detect in Q-Q plots, detrending the plot by considering deviations of empirical and theoretical quantiles (also called worm plots) can be helpful.

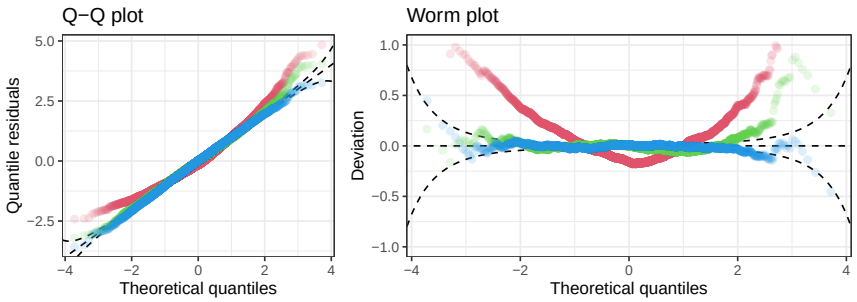


FIGURE 2. Q-Q plot (left) and worm plot (right) for the homoscedastic Gaussian model (red), as well as the heteroscedastic left-censored Gaussian (green) and the heteroscedastic left-censored logistic (blue) model.

Both, the Q-Q plot and the worm plot included in Figure 2 show an obvious misfit of the homoscedastic Gaussian model on both tails of the distribution. While the marginal calibration (Fig. 1) clearly shows that this is mainly due to missing censoring, the probabilistic calibration in this example does not allow to uncover the sources causing this lack of fit.
While the scale itself is not easily interpretable, the probabilistic calibration allows to check if the distributional assumption is correct. Comparing both left-censored heteroscedastic models, a slight advantage can be seen for the one using a left-censored logistic distribution as its heavier tails lead to a better fit, especially for high quantiles.

References

Dunn, P.K., and Smyth G.K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, **5**(3), 236 – 244.

Gneiting, T., Balabdaoui, F., and Raftery, A.E. (2007). Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society: Series B (Methodological)*, **69**(2), 243 – 268.

Messner, J.W., Mayr, G.J., and Zeileis A. (2016). Heteroscedastic Censored and Truncated Regression with crch. *The R Journal*, **8**(1), 173 – 181.

Maximum softly-penalized likelihood for mixed effects logistic regression

Philipp Sterzinger[1], Ioannis Kosmidis[1,2]

[1] Warwick University, Coventry, UK
[2] The Alan Turing Institute, London, UK

E-mail for correspondence: philipp.sterzinger@warwick.ac.uk

Keywords: Logistic Regression, Infinite Estimates, Singular Variance Components, Data Separation, Jeffreys Prior

1 Introduction

Parameter estimation in mixed effects logistic regression by maximum likelihood (ML) is prevalent in statistical practice, because these estimators are expected to achieve optimal maximum likelihood asymptotics under standard regularity conditions. However, there exist data configurations that lead to estimates on the boundary of the parameter space, such as infinite values for fixed effects and singular or infinite variance components. Such estimates can cause havoc to numerical estimation procedures and can, if undetected, substantially impact inferential procedures resulting in spuriously strong or weak conclusions (e.g. Chung et al. 2013, Section 2.1). We introduce a maximum softly penalized likelihood (MSPL) estimator that always lies in the interior of the parameter space and preserves ML asymptotics. The penalty we propose consists of appropriately scaled versions of Jeffreys invariant prior for the model with no random effects, and of compositions of the negative Huber loss functions for the variance components. The resulting MSPL estimates are guaranteed to be in the interior of the parameter space. Scaling the penalty appropriately guarantees that i) penalization is "soft" enough for the MSPL estimator to have the same optimal asymptotic properties expected by the ML estimator, and ii) that the fixed effects estimates are equivariant under linear transformations of the model parameters, such as contrasts, in the sense that the MSPL estimates of linear transformations of the fixed effects parameters are the linear transformations of the MSPL estimates.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Mixed effects logistic regression

Consider a mixed-effects logistic regression with k clusters and n_i observations in the ith cluster ($i = 1, \ldots, k$). The likelihood about the fixed effects $\boldsymbol{\beta} \in \Re^p$ and variance components $\boldsymbol{\Sigma} \in \Re^{q \times q}$ is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = (2\pi)^{-kq/2} \det(\boldsymbol{\Sigma})^{-k/2} \prod_{i=1}^{k} \int_{\Re^q} \prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}} (1-\mu_{ij})^{1-y_{ij}} \exp\left\{ -\frac{\boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{u}_i}{2} \right\} d\boldsymbol{u}_i \,, \tag{1}$$

where $\mu_{ij} = \Pr(y_{ij} = 1 \mid \boldsymbol{u}_i, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$ for fixed and random effects covariates $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$ and random effects $\boldsymbol{u}_i$. The conditional means μ_{ij} are linked to the linear predictor $\eta_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{z}_{ij} \boldsymbol{u}_i$ by $\log(\mu_{ij}/(1 - \mu_{ij})) = \eta_{ij}$. For clarity of presentation, all developments are stated for the exact model likelihood $L(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, which is generally not available in closed form, but they can be extended to approximate log-likelihoods under suitable conditions on the approximation error.

3 Motivating example

TABLE 1. Culcita data of McKeon et al. (2012)

	Block									
Treatment	1	2	3	4	5	6	7	8	9	10
none	0,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,0
crabs	0,0	0,0	0,0	0,0	1,1	1,1	1,1	1,1	1,1	1,1
shrimp	0,0	0,0	0,0	0,0	0,1	1,1	1,1	1,1	1,1	1,1
both	0,0	0,0	0,0	0,0	0,0	0,1	1,1	1,1	1,1	1,1

As a motivating example, we consider the Culcita data set of McKeon et al. (2012) as provided in the worked examples of Bolker (2015), which is shown in Table 1. The complete randomized block design data (four treatments, ten temporal blocks, two replications per block) records coral-eating sea stars (Culcita) attacking coral harbouring different protective symbionts (none, crabs, shrimp, both). Upon removal of the atypical observation in the top right corner of Table 1, we associate predation to treatment effects using a mixed effects logistic regression with a random intercept per block. We estimate $\boldsymbol{\beta}$ and $\log \sigma$ by ML, with a 200-point Gauss-Hermite quadrature approximation to the log-likelihood and three different optimization routines "CG", "BFGS", "nlm" from the R (R Core Team 2022) `optimx` package. Results are shown in Figure 1. The estimates and asymptotic 95% confidence intervals based on the negative Hessian of the approximate log-likelihood are markedly dissimilar and notably extreme on the logistic scale.

The large estimated standard errors are indicative of an almost flat approximate log-likelihood around the estimates. In this case, the ML estimates for the fixed effects $\boldsymbol{\beta}$ are in reality infinite in absolute value. However, due to different stopping criteria of the various optimization routines, the estimates from the various ML implementations appear finite and distinct.

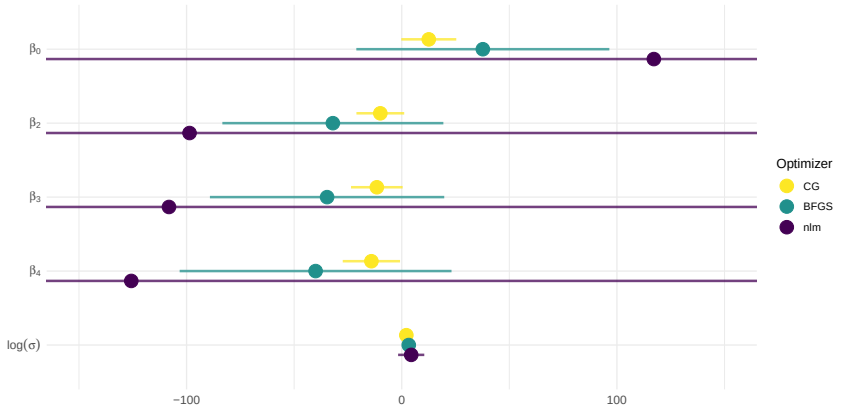


FIGURE 1. ML estimates (points) and asymptotic 95% confidence intervals (lines) from fitting a mixed effects logistic regression model to the Culcita data upon removal of an outlier observation.

4 Non-boundary estimates

Denote by $\partial\Theta$ the boundary of Θ and let $\boldsymbol{\theta}(r)$, $r \in \Re$, be a path in the parameter space such that $\lim_{r\to\infty} \boldsymbol{\theta}(r) \in \partial\Theta$. A common approach to resolving issues with ML estimates being in $\partial\Theta$, like those encountered in the example of Section 3, is to introduce an additive penalty $P(\boldsymbol{\theta})$ to the (approximate) log-likelihood that satisfies $\lim_{r\to\infty} P(\boldsymbol{\theta}(r)) = -\infty$. Hence, if $\ell(\boldsymbol{\theta})$ is bounded from above and there is at least one point $\boldsymbol{\theta} \in \Theta$ such that $\ell(\boldsymbol{\theta}) + P(\boldsymbol{\theta}) > -\infty$, then $\tilde{\boldsymbol{\theta}}$ is in the interior of Θ.

5 Composite penalty

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\psi}^{\mathrm{T}})^{\mathrm{T}}$ and $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\beta}, s(\boldsymbol{\psi}))$ with $s(\boldsymbol{\psi}) = \boldsymbol{\Sigma}$, where $L(\boldsymbol{\beta}, s(\boldsymbol{\psi}))$ is (1). The parameter vector $\boldsymbol{\psi}$ is defined as $\boldsymbol{\psi} = (\log l_{11}, \ldots, \log l_{qq}, l_{21}, \ldots, l_{q1}, l_{32}, \ldots, l_{q2}, \ldots, l_{qq-1})^{\mathrm{T}}$, where l_{ij} $(i > j)$ is the (i, j)th element of the lower-triangular Cholesky factor $\boldsymbol{L}$ of $\boldsymbol{\Sigma}$, i.e. $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$. Consider the estimator

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\ \Theta} \left\{ \ell(\boldsymbol{\theta}) + c_1 P_{(f)}(\boldsymbol{\beta}) + c_2 P_{(v)}(\boldsymbol{\psi}) \right\} , \tag{2}$$

where $c_1 > 0$, $c_2 > 0$, and $P_{(f)}(\boldsymbol{\beta})$ and $P_{(v)}(\boldsymbol{\psi})$ are unscaled penalty functions for the fixed effects and variance components, respectively.
For the unscaled fixed effects penalty, we use the logarithm of Jeffreys invariant prior for the corresponding GLM, that is

$$P_{(f)}(\boldsymbol{\beta}) = \frac{1}{2} \log \det \left(\sum_{i=1}^{k} \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{W}_i \boldsymbol{X}_i \right) , \tag{3}$$

where $\boldsymbol{X}_i$ collects the covariates for the fixed effects in model (1), and $\boldsymbol{W}_i$ is a diagonal matrix with jth diagonal element $\mu_{ij}^{(f)}(1 - \mu_{ij}^{(f)})$ with $\mu_{ij}^{(f)} = \exp(\eta_{ij}^{(f)})/\{1 + \exp(\eta_{ij}^{(f)})\}$ and $\eta_{ij}^{(f)} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$. Note that for any invertible matrix $\boldsymbol{C} \in \Re^{p \times p}$, and for $\boldsymbol{\gamma} = \boldsymbol{C}\boldsymbol{\beta}$, $P_{(f)}(\boldsymbol{\gamma}) = P_{(f)}(\boldsymbol{\beta}) - \log\det(\boldsymbol{C})$. As a result, for any $\boldsymbol{\psi}$, the MSPL fixed effects estimates of $\boldsymbol{\gamma}$ are simply $\tilde{\boldsymbol{\gamma}} = \boldsymbol{C}\tilde{\boldsymbol{\beta}}$. Hence, one can obtain MSPL fixed effects estimates and corresponding estimated standard errors for arbitrary sets of scaled parameter contrasts of the fixed effects, when estimates for one of those sets of contrasts are available and with no need to re-estimate the model.
For the variance components penalty, we use a composition of negative Huber loss functions on the components of $\boldsymbol{\psi}$. In particular,

$$P_{(v)}(\boldsymbol{\psi}) = \sum_{i=1}^{q} D(\log l_{ii}) + \sum_{i>j} D(l_{ij}) , \tag{4}$$

where

$$D(x) = \begin{cases} -\frac{1}{2}x^2, & \text{if } |x| \leq 1 \\ -|x| + \frac{1}{2}, & \text{otherwise} \end{cases} .$$

For the remainder of this work, denote by $P(\boldsymbol{\theta})$ the composite penalty function in (2) with components (3) and (4). It can be shown that $\lim_{r\to\infty} P(\boldsymbol{\theta}(r)) = -\infty$, for any sequence $\boldsymbol{\theta}(r)$ such that $\lim_{r\to\infty} \boldsymbol{\theta}(r) \in \partial\Theta$, so that the resulting estimator is guaranteed to be in the interior of the parameter space as long as there is at least one point in Θ such that the penalized log-likelihood is not minus infinity. Hence, the MPL estimates for $\boldsymbol{\beta}, \boldsymbol{\psi}$ have finite components and the value of $\tilde{\boldsymbol{\Sigma}} = s(\tilde{\boldsymbol{\psi}})$ is guaranteed to be non-degenerate in the sense that it is positive definite with finite entries, implying correlations away from one in absolute value.

6 Consistency and asymptotic normality

To mitigate any distortion of the estimates by the penalization of the log-likelihood, we choose the scaling factors c_1, c_2 to be "soft" enough to control $\|\nabla P(\boldsymbol{\theta})\|$ in terms of the rate of information accumulation r_n, for which $r_n^{-1} J(\boldsymbol{\theta}) \xrightarrow{p} I(\boldsymbol{\theta})$ as $n \to \infty$, where $J(\boldsymbol{\theta}) = -\nabla\nabla^{\mathrm{T}}\ell(\boldsymbol{\theta})$ is the observed information matrix and $I(\boldsymbol{\theta})$ is a $\mathcal{O}(1)$ matrix.

It can be shown that for $c_1 = c_2 = 2\sqrt{p/n}$, which corresponds to the square root of the average approximate variances of $\hat{\eta}_{ij}^{(f)}$ at $\boldsymbol{\beta} = \mathbf{0}_p$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})\| \leq \frac{p^2}{\sqrt{n}} \max_{i,s,t} |[\boldsymbol{X}_i]_{st}| + \sqrt{\frac{2pq(q+1)}{n}},$$

which, under some model regularity conditions, is sufficient to establish consistency and asymptotic normality of $\tilde{\boldsymbol{\theta}}$ as long as $\max_{i,s,t} |[\boldsymbol{X}_i]_{st}| = O_p(n^{1/2})$. The condition on the maximum of the absolute elements of the model matrix is not unreasonable in practice. It certainly holds true for covariates such as dummy variables, as included in the real-data example in this work. It is also true for model matrices with subgaussian random variables with common variance proxy σ^2.

7 Simulation study

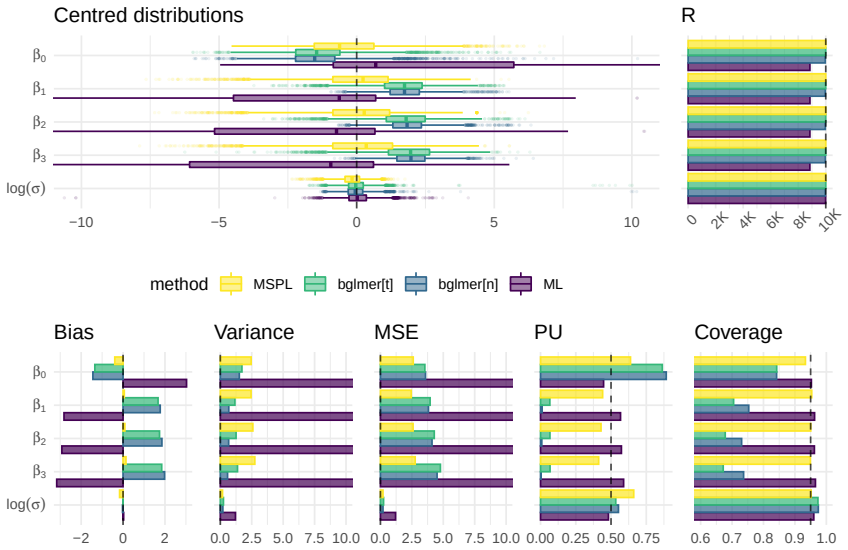


FIGURE 2. Simulation summaries for MSPL, `bglmer` and ML estimators from simulating 10 000 samples from (1) from the Culcita data at the ML estimates.

We simulate 10 000 independent samples of responses for the randomized complete block design data from McKeon et al. (2012) at the ML estimates. For each sample, we compute the ML and MSPL estimates. We compare these estimates with the `bglmer` routine of Chung et al. (2013), which was developed to address degenerate parameter estimates in GLMMs, using a

normal (bglmer[n]) and t (bglmer[t]) prior penalty for the fixed effects and a gamma-prior inspired penalty for the random effects variance. All estimators use a 100-point adaptive Gauss-Hermite quadrature approximation to the model likelihood. Boundary parameter estimates were discarded for the calculation of summary statistics; results are given in Figure 2. The number of used estimates is given in the top right panel (R). The top left panel shows the centred sampling distribution of the parameter estimates for MSPL, `bglmer` and ML as returned by the optimization routines. With the exception of MSPL, these boxplots are not estimates of the actual density of the ML and bglmer estimators, but rather of their conditional density given that they do not take boundary values. The bottom panels give simulation based estimates of the bias, variance, mean squared error (MSE), the probability of underestimation (PU) and the coverage based on 95% Wald-confidence intervals. Clearly, the amount of shrinkage induced by the normal and t priors is excessive. Although the resulting estimators have small finite-sample variance, they have excessive finite-sample bias, which is often at the order of the standard deviation. The combination of small variance and large bias readily impacts first-order inferences. Wald-type confidence intervals about the fixed effects systematically undercover the true parameter value. Finally, both bglmer[n] and bglmer[t] do not appear to prevent extreme positive variance estimates. The MSPL gives estimates in the interior of the parameter space, has very small bias, accurate coverage from the asymptotic 95% confidence intervals, the smallest mean squared error (MSE), and a well calibrated probability of underestimation (PU).

References

Bolker, B. M. (2015). Linear and generalized linear mixed models. In: *Ecological Statistics*, Oxford, 309 – 333.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A. and Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models, *Psychometrika*, **78**, 685 – 709.

Kosmidis, I. and Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models, *Biometrika*, **108**, 71 – 82.

McKeon, C.S., Stier, A.C., McIlroy, S.E. and Bolker, B.M. (2012). Multiple defender effects: synergistic coral defense by mutualist crustaceans, *Oecologia*, **169**, 1095 – 1103.

Ogden, H.E. (2017). On asymptotic validity of naive inference with an approximate likelihood, *Biometrika*, **104**, 153 – 164.

R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Longitudinal clustering of athletes' careers under informative missing data patterns

Mattia Stival[1], Mauro Bernardi[1], Manuela Cattelan[1], Petros Dellaportas[234]

[1] Department of Statistical Sciences, University of Padova, Padova, Italy
[2] Department of Statistical Science, University College London, London, UK
[3] Department of Statistics, Athens University of Economics and Business, Athens, Greece
[4] The Alan Turing Institute, London, UK

E-mail for correspondence: mattia.stival@unipd.it

Abstract: In this work we analyze the evolution in the careers of 369 Italian male middle-distance runners, born in 1988, considering their seasonal best performances in the 800, 1500, 5000 meters races during the period 2006–2019. In this context, clustering of trajectories allows to identify the possible careers of one athlete, a relevant aspect for coaches that aim at planning the future and tracking the progress of their athletes. However, differently from other disciplines, the presence of missing values for middle distance athletes is a critical aspect as they are potentially correlated with performances. On one side, drop-in and drop-out phenomena implicitly lead to a different development in each athlete's career history. On the other side, middle distance athletes can compete in different races, an aspect which is typically related to their personal attitudes. We propose a Bayesian clustering model in which both the observed trends and the presence of missing data inform on the clustering structure. Observed trends of each race are described by group-specific state space models, useful to capture longitudinal dependence across performances of the same athlete. Information on missing values is included by means of two distinct group dependent processes: the first one describes the drop-in and drop-out phenomena in the sample; the second one describes the actual participation in the competitions by the athletes, as an index of their different attitudes. Our findings suggest that athletes who are more likely to participate in different type of races have better performances during the years.

Keywords: Informative missing data; Longitudinal clustering; Sports performance analysis.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

1 Introduction

Planning the future career of young athletes is a relevant aspect of the work of coaches, whose role is to guide athletes during training so that they can perform at their best in competition and achieve their desired results. In this context, the identification of possible careers for an athlete, in terms of observed personal performance trajectories over time, is of paramount importance because it can give indications on the expected progress of different athletes over the years, thus indicating whether the training process has been carried out correctly. In this paper, we analyze a new dataset describing the careers of Italian male middle-distance athletes, born in 1988, with age ranging between 18 and 32 years old, and focus on clustering of longitudinal data. Observed data are represented in Figure 1. Clustering of longitudinal data has been extensively explored in literature (see, Frühwirth-Schnatter, 2011; Bartolucci and Murphy, 2015), as a tool for identifying different observable scenarios and describing the heterogeneity present in the data. We extend the matrix-variate state space models for clustering so that the clustering structure depends both on the observed trend and on the presence of missing data. Indeed, unlike other types of athletes and sports, middle distance runners can compete in different distances, i.e. in the 800, 1500 and 5000 meters races, as well as in other spurious races. The choice of races in which to compete is subjective and typically associated with personal attitudes, and not all athletes compete in all type of races. In this way, not only do we observe different races for each athlete over time, but the absence of a particular race can be informative on the athletes' attitudes. Beyond the variability among subjects related to the type of races performed, there is also variability in the development of athletes' careers, related for example to the number of year spent in career. These aspects are related to drop-in and drop-out phenomena, defined as the events where athletes enter and exit the observed sample, respectively. In this context, the presence and absence of data is potentially correlated with the observed performances, a critical aspect in clustering since neglecting them (or not) may lead to different conclusions and different clusters as well.

2 The model

Let $y_{pq,t}$ denote the seasonal best performance of race p for athlete q during the year t, for $p = 1, \ldots, P$, $q = 1, \ldots, Q$, and $t = 1, \ldots, T$. Suppose also that the complete set of observations were available, so that each athlete participates in all P races during the years and that no drop-ins or drop-outs are observed. Assume that athletes are divided into G different unknown groups according to the evolutionary trajectories during their careers. Let athlete q belong to group g, then its seasonal best race results

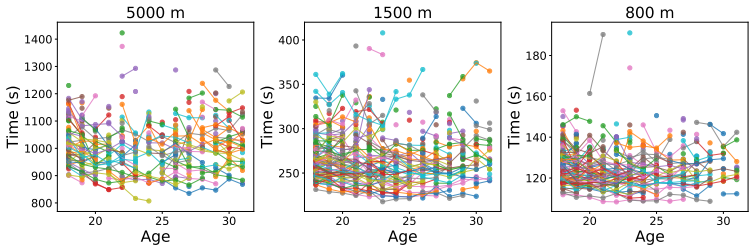


FIGURE 1. Observed seasonal best performance in 5000, 1500, and 800 meters race for $Q = 369$ Italian middle distance runners, born in 1988. Data collection was done from the annual rankings, which publicly available at `www.fidal.it`, the official website of Italian athletics federation.

over time are described by the following dynamic linear model

$$y_{pq,t} = \mathbf{z}_p^\top \boldsymbol{\alpha}_{p,t}^{(g)} + \varepsilon_{pq,t}, \quad \varepsilon_{pq,t} \sim \mathrm{N}_1(0, \sigma_p^2),$$
$$\boldsymbol{\alpha}_{p,t+1}^{(g)} = \mathbf{T}_p \boldsymbol{\alpha}_{p,t}^{(g)} + \boldsymbol{\xi}_{p,t}^{(g)}, \quad \boldsymbol{\xi}_{p,t}^{(g)} \sim \mathrm{N}_{F_p}(\mathbf{0}, \boldsymbol{\Psi}_p),$$

where $\boldsymbol{\alpha}_{p,1}^{(g)} \sim \mathrm{N}_{F_p}(\widehat{\boldsymbol{\alpha}}_{p,1|0}^{(g)}, \mathbf{P}_{p,1|0}^{(g)})$, for $p = 1, \ldots, P$, $t = 1, \ldots, T$, and $\widehat{\boldsymbol{\alpha}}_{p,1|0}^{(g)}$, $\mathbf{P}_{p,1|0}^{(g)}$ are fixed mean and variance of the latent process at the first time instant. In the above specification, the row vector $\mathbf{z}_p^\top$, which is characterized by a known structure, links the observation $y_{pq,t}$ to the column vector $\boldsymbol{\alpha}_{p,t}^{(g)}$, which describes the group-specific dynamics of the p–race for all the athletes that belong to group g. These dynamics are determined by the state transition equation, that describes a first-order autoregressive process with transition matrix $\mathbf{T}_p$, which is race-specific, known, and shared across all the groups. Moreover, this dependence is not required to be common across different races, as $\mathbf{T}_p$ may differ from $\mathbf{T}_{p'}$ for any $p \neq p'$. The error terms $\varepsilon_{pq,1}, \ldots, \varepsilon_{pq,T}$ are assumed to be serially independent and independent of both the states $\boldsymbol{\alpha}_{p,1}^{(g)}, \ldots, \boldsymbol{\alpha}_{p,T}^{(g)}$ and the disturbances $\boldsymbol{\xi}_{p,1}^{(g)}, \ldots, \boldsymbol{\xi}_{p,T}^{(g)}$, for $p = 1, \ldots, P$ and $g = 1, \ldots, G$. Given these assumptions, it is possible to write the model in a compact matrix-variate state space representation (see, e.g., Wang and West, 2009).

The model described assumes that all data are observed, i.e., that the athletes run all P races during the years and that drop-ins and drop-outs are not observed. However, this is not the case for data that describe the career trajectories of athletes, since the lack of data is part of the career itself. To include these factors as informative aspects of athletes' career, we include two more variables in the model. The variable $d_{pq,t} \in \{0, 1\}$ describes, respectively, the absence or presence of an observations for race p of athlete q during year t. The variable $d_{q,t}^\star$ represents whether the athlete q is in career during year t or not. More specifically, $d_{q,t}^\star$ is such that $d_{q,t}^\star = 0$ if athlete q has never started the career before t (included), $d_{q,t}^\star = 1$ if athlete

q is in career during t, or $d^\star_{q,t} = 2$ if athlete q has finished the career in t (included). The variable $d^\star_{q,t}$ is not decreasing in t, and describes the three possible states of athlete's career. Moreover, if $d^\star_{q,t} \in \{0, 2\}$, then $d_{pq,t} = 0$ with probability 1, for $p = 1, \ldots, P$, meaning that no races are observed since the athlete is not competing. On the contrary, there might be athletes such that $d_{pq,t} = 0$, for $p = 1, \ldots, P$, even if $d^\star_{q,t} = 1$. This is typical of athletes who, despite being in a career, decide not to compete during one specific year, but compete in the following years. Let $\mathbf{d}^\star_q = (d^\star_{q,1}, \ldots, d^\star_{q,T})$, $\mathbf{d}_{\cdot q,t} = (d_{1q,t}, \ldots, d_{Pq,t})^\top$, and $\mathbf{D}_q = \begin{bmatrix}\mathbf{d}_{\cdot q,1} & \ldots & \mathbf{d}_{\cdot q,T}\end{bmatrix}$. To make the missing data informative on the clustering structure, we assume that the likelihood associated with them is dependent on the cluster allocation S_q, and assume

$$p_{\boldsymbol{\theta}}(\mathbf{D}_q, \mathbf{d}^\star_q | S_q) = \prod_{t=1}^{T} \left[\prod_{p=1}^{P} p_{\boldsymbol{\theta}}(d_{pq,t} | d^\star_{q,t}, S_q) \right] p_{\boldsymbol{\theta}}(d^\star_{q,t} | d^\star_{q,t-1}, S_q),$$

where $p_{\boldsymbol{\theta}}(d^\star_{q,1} = 1 | d^\star_{q,0}, S_q = g) = \pi^\star_{1g}$ and $p_{\boldsymbol{\theta}}(d^\star_{q,1} = 0 | d^\star_{q,0}, S_q = g) = 1 - \pi^\star_{1g}$, with $d^\star_{q,0} = 0$ fixed for $q = 1, \ldots, Q$, as well as $p_{\boldsymbol{\theta}}(d_{pq,t} = 1 | d^\star_{q,t} = 1, S_q = g) = \pi_{pg}$, and $p_{\boldsymbol{\theta}}(d_{pq,t} = 0 | d^\star_{q,t} = 1, S_q = g) = 1 - \pi_{pg}$. We adopt a fully conditionally conjugate Bayesian approach, which allows to derive a Gibbs sampler for the estimation of the model.

3 Results

Analysis of results based on the posterior distribution can be different and with various levels of complexity. From a practical point of view, it is interesting to compare distinct groups based on the states describing the performances of distinct athletes, and hence, looking whether different missing data patterns are effectively associated with better or worse performances. An example is provided in Figure 2, that shows, in its first line, the performances in the 1500 meters race of three (out of 9) selected groups, and, in the second line, the performances in the 800 meters race of three other groups. In the 1500 meters discipline, we note that the groups differ not only in the level of performance but also with respect to the number of observations present in each graph, with the third one characterized by much longer careers. A longer career therefore appears to be effectively associated with generally better performance over time. In the 800 meters, on the contrary, differences are more marked especially for the blue group, which is characterized by worse performances and later entry into competitions. Other interesting findings of our application suggest that athletes who are more likely to participate in different disciplines rather than one single discipline have better performances during the years. In conclusion, our model shows that the presence of different missing data patterns in the data appears to be effectively associated with better and worse athletes' performances. On one side, these results highlight the importance of

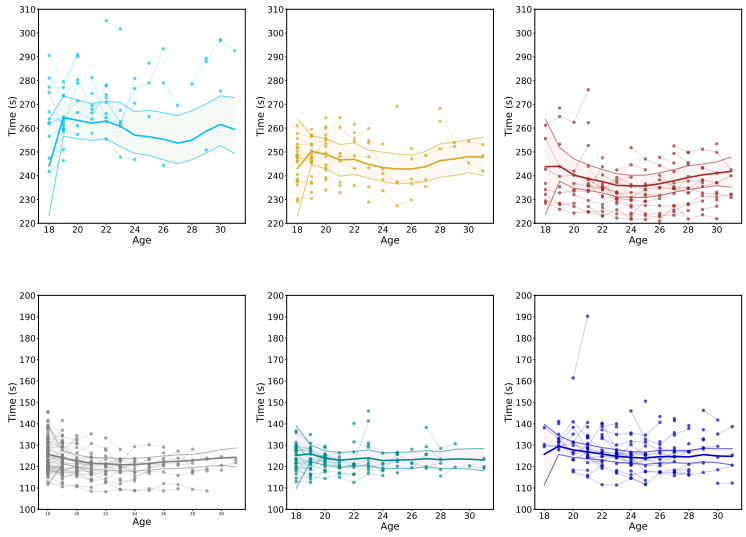


FIGURE 2. In the first line, performances on 1500 meters race for three selected groups. In the second line, performances on 800 meters race for three other groups. Thicker lines denote posterior medians of the states. Colored bands denote the respective 90% pointwise posterior credible intervals based on quantiles. Observed data are represented in the background, according to athletes' MAP cluster allocations.

starting careers at young ages. On the other side, they highlight the importance of being able to compete in different disciplines over the years. This aspect, not only allows to monitor the progress of the athletes, but also suggests possible strategies of competitions (e.g. competing in different type of races) to obtain better results. In this work, the analyses were obtained using a sample of Italian male athletes. Future research involves studying female athletes, or athletes from other countries or at the international level. Beyond that, our approach also fits well with other sports in which athletes may compete in different disciplines, such as swimming or track-cycling. From a methodological perspective, further analyses will investigate in detail the choice of using a fixed group number, rather than alternative approaches such as the use of sparse finite mixture (Malsiner-Walli et al., 2016).

Acknowledgments: This research was supported by funding from the University of Padova Research Grant 2019-2020, under grant agreement BIRD203991.

References

Bartolucci, F. and Murphy, T.B. (2015). A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. *Journal of Quantitative Analysis in Sports*, **11**(4), 193 – 203.

Frühwirth-Schnatter, S. (2011). Panel data analysis: a survey on model-based clustering of time series. *Advances in Data Analysis and Classification*, Oxford: Clarendon Press. **5**(4), 251 – 280.

Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, **26**, 303 – 324.

Wang, H. and West, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika*, **96**(4), 821 – 834.

The Feature-First Block Model

Lawrence Tray[1], Ioannis Kontoyiannis[2]

[1] Department of Engineering, University of Cambridge, UK
[2] Statistical Laboratory, University of Cambridge, UK

E-mail for correspondence: lpt30@cantab.ac.uk

Abstract: Labelled networks are an important class of data, naturally appearing in applications in science and engineering. A typical inference goal is to determine how the vertex labels (or *features*) affect the network's structure. We introduce a new generative model, the feature-first block model (FFBM), that facilitates the use of rich queries on labelled networks. We develop a Bayesian framework and devise a two-level Markov chain Monte Carlo approach to efficiently sample from the posterior distribution of the FFBM parameters. This allows us to infer if and how the observed vertex-features affect macro-structure. We apply the proposed methods to several real-world networks to extract the most important features along which the vertices are partitioned. Importantly, the whole feature-space is used automatically and features can be rank-ordered implicitly by importance. [The full version of this paper is available on arXiv as [cs.LG] 2105.13762.]

Keywords: Stochastic Block Model; Labelled Networks; Inference.

1 Introduction

Many real-world networks exhibit strong community structure, with most nodes belonging to densely connected clusters. In this work, we examine vertex-labelled networks, referring to the labels as *features*. A typical goal is to determine whether a given feature impacts graphical structure. Answering this requires a random graph model; the standard is the stochastic block model (SBM), see Peixoto (2017).
Analysing a labelled network with one of the standard SBM variants requires partitioning the graph into blocks grouped by distinct values of the feature of interest. The associated model can then be used to test for evidence of heterogeneous connectivity between the feature-grouped blocks. But this approach can only consider disjoint feature sets and the feature-grouped blocks often provide an unnatural partition.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

We would instead prefer to partition the graph into its most natural blocks and then find which of the available features – if any – best predict the resulting partition. Thus motivated, we present a novel framework for modelling labelled networks. This is not the first extension of the SBM to labelled networks, e.g. Stanley et al (2019). However, most of the current approaches are focused on leveraging feature information to partition the graph more reliably in the presence of noise. We seek instead to develop a model well suited for inferring how vertex features impact graphical structure and to report our confidence in those conclusions.

2 Feature-First Block Model

We propose a novel generative model for labelled networks, which we call the feature-first block model (FFBM), illustrated in Figure 1. Let N denote the number of vertices, B the number of blocks and D the number of features associated with each vertex. We write X for the $N \times D$ *feature matrix* containing the feature vectors $\{x_i\}_{i=1}^N$ as its rows. For the FFBM, we start with the feature matrix X and generate a random vector of block memberships $b \in [B]^N$, where we write $[K] = \{1, 2 \dots K\}$. For each vertex i, the block membership $b_i \in [B]$ is generated based on the feature vector x_i, independently between vertices, so $p(b|X, \theta) = \prod_{i \in [N]} \phi_{b_i}(x_i; \theta)$.

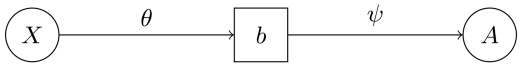


FIGURE 1. The Feature-First Block Model (FFBM)

Once the block memberships b have been generated, we then draw the adjacency matrix of the graph $A \sim \text{DC-SBM}_{\text{MC}}(b, \psi)$ from the microcanonical DC-SBM, Peixoto (2017), with additional parameters ψ. Appropriate priors are placed on the parameters θ and ψ to complete the Bayesian framework.

3 Inference

Given a labelled network (A, X), we wish to infer if and how the observed features X impact the graphical structure A. Formally, this means characterising the posterior distribution for θ, $p(\theta|A, X) \propto p(\theta) \cdot p(A|X, \theta)$. Following standard Bayesian practice, we propose an iterative Markov chain Monte Carlo (MCMC) approach to obtain samples $\theta^{(t)}$ from this posterior. First we sample $b^{(t)}$ from the block membership posterior, and then use $b^{(t)}$ to obtain a corresponding sample $\theta^{(t)}$,

$$b^{(t)} \sim p\big(b|A, X\big) \quad \text{then} \quad \theta^{(t)} \sim p\big(\theta|X, b^{(t)}\big). \tag{1}$$

Splitting the Markov chain into two levels side-steps the intractable summation over all latent $b \in [B]^N$ required to directly compute the likelihood, $p(A|X,\theta)$. The resulting $\theta^{(t)}$ samples are asymptotically unbiased in that the expectation of their distribution converges to the true posterior.

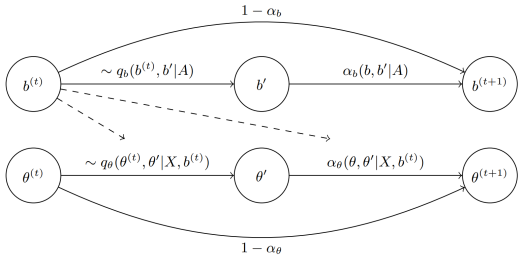


FIGURE 2. θ-sample generation.

Figure 2 shows an overview of the proposed method, with q and α denoting the Metropolis-Hastings proposal distribution and acceptance probability. Due to the formulation of the FFBM, evaluating $p(b|X)$ does not depend on X so we do not need X to sample b. And on the other level, in order to obtain samples for θ we use only b but not A, as $(\theta \perp\!\!\!\perp A)|b$.

4 Experimental results

We apply our proposed methods to a variety of real-world datasets. The inferred partitions b for all of these are given on Figure 3. To assess model performance, the average description length per entity (nodes and edges) $\bar{S}_e$ is used to gauge the SBM fit, and the vertex set $[N]$ is partitioned at random into training and test sets, $\mathcal{G}_0$ and $\mathcal{G}_1$, to assess the performance of the feature-to-block predictor. The average cross-entropy loss over each set, denoted $\bar{\mathcal{L}}_\star$, is used to gauge the quality of the fit.
For higher-dimensional datasets, we develop a novel dimensionality reduction method to select only the top D' features. We then retrain the feature-block predictor using only the retained feature set, and report the cross-entropy loss $\bar{\mathcal{L}}'_\star$ over the training and test sets for the reduced classifier.
Table 1 summarises the results for each experiment. We see that the dimensionality reduction procedure brings the training and test losses closer, indicating that the retained features are indeed well correlated with the underlying graphical partition and that the approach generalises correctly.

TABLE 1. Results averaged over $n = 10$ iterations (mean $\pm$ std. dev.).

Dataset	B	D	D'	$\bar{S}_e$	$\bar{\mathcal{L}}_0$	$\bar{\mathcal{L}}_1$	c^*	$\bar{\mathcal{L}}'_0$	$\bar{\mathcal{L}}'_1$
Polbooks	3	3	–	2.250 ± 0.000	0.563 ± 0.042	0.595 ± 0.089	–	–	–
School	10	13	10	1.894 ± 0.004	0.787 ± 0.127	0.885 ± 0.129	1.198 ± 0.249	0.793 ± 0.132	0.853 ± 0.132
FB egonet	10	480	10	1.626 ± 0.003	1.326 ± 0.043	1.538 ± 0.069	0.94 ± 0.019	1.580 ± 0.150	1.605 ± 0.106

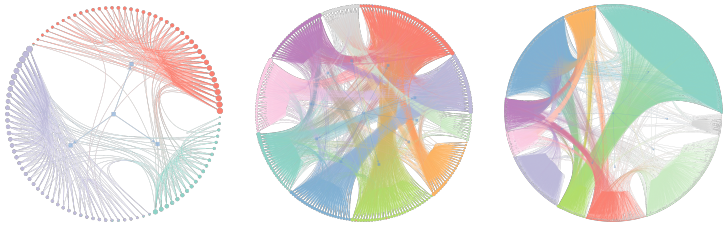


FIGURE 3. Networks laid out and coloured according to inferred block memberships. Left to right: Polbooks, Krebs (2004); Primary School, Stehle et al (2011); Facebook Egonet, Leskovec and Mcauley (2012).

5 Conclusion

The feature-first block model (FFBM) is introduced, as a new generative model for labelled networks with communities. An efficient MCMC algorithm is developed for sampling from the posterior distribution of the relevant parameters in the FFBM; the main idea is to divide up the graph into its most natural partition under the associated parameter values, and then to determine whether the vertex features can accurately explain the partition. Through applications on empirical network data, this approach is demonstrated to be effective at extracting and describing the most natural communities in a labelled network. Nevertheless, it can only currently explain the structure at the macroscopic scale. Future work will benefit from extending the FFBM to a further hierarchical model, so that the structure of the network can be explained at all scales of interest.

References

Krebs, V. (2004) . The political books network, http://www.orgnet.com/.

Leskovec, J., Mcauley, J. (2012). Learning to discover social circles in ego networks. *Advances in Neural Information Processing Systems* vol. 25

Peixoto, T.P. (2017). Nonparametric Bayesian inference of the microcanonical stochastic block model. Physical Review E 95(1).

Stanley, N. et al (2019). Stochastic block models with multiple continuous attributes. Applied Network Science, 4(1), 1-22.

Stehle, J. et al (2011). High resolution measurements of face-to-face contact patterns in a primary school. PLoS ONE 6(8), 1—13.

Tray, L., Kontoyiannis, I. (2021). The feature-first block model. arXiv preprint 2105.13762 [cs.LG].

GLMM Based Clustering of Multivariate Mixed Type Longitudinal Data

Jan Vávra[1], Arnošt Komárek[1]

[1] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

E-mail for correspondence: vavraj@karlin.mff.cuni.cz

Abstract: Several GLMMs for longitudinal data of a mixed type are joined together. Then, a mixture of such models is proposed to classify units into groups differing in evolution patterns. Data from The European Union Statistics on Income and Living Conditions database (EU-SILC) are analysed.

Keywords: Multivariate longitudinal data; GLMM; Model-based clustering; Sparse finite mixture.

1 Introduction

Fundamental purpose of the European Union Statistics on Income and Living Conditions (EU-SILC) survey is to map actual life situation within European households, their social-demographic structure, income differentiation, quality and financial burden of housing. The rotational design of this still ongoing study replaces a quarter of households each year so that each household is observed for no more than $n_i = 4$ consecutive years. Apart from the *Equivalised total disposable household income*, which is the key numeric outcome, plenty of other closely related numeric, binary, ordinal or general categorical outcomes are recorded. The observed values may depend on several potential regressors such as the size of a household, the household location and its population density, etc., but most importantly the time. During the follow-up period, the households had to withstand many unfavourable external conditions in the form of economical crisis and its consequences. Our task is to differentiate patterns in the evolution of social-economic indicators. Households intact and still prospering, households of a steady course or even households negatively impacted by the crisis are expected to be discovered.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Model

In our previous work (Vávra and Komárek, 2022), we modelled categorical outcomes by thresholding a latent numeric outcome. Here we propose a modification that utilizes *generalized linear mixed-effects models* (GLMM) instead and, moreover, simultaneously estimates the suitable number of clusters.

2.1 Generalized linear mixed-effects models

Let us denote by $Y_{i,j}^r$ the jth value ($j = 1, \ldots, n_i$) of outcome $r \in \{1, \ldots, R\}$ of type $\mathsf{t}(r) \in \{\mathsf{N}, \mathsf{B}, \mathsf{O}, \mathsf{C}\}$ (numeric, binary, ordinal or general categorical, respectively) observed within the household $i = 1, \ldots, n$ together with a set of covariates $\mathcal{C}_{i,j}$ including the time $t_{i,j}$. For each outcome r we construct a predictor $\eta_{i,j}^r$ consisting of fixed (given by coefficients $\boldsymbol{\beta}_r$) and random effects $\mathbf{b}_i^r$ specific to household i.
Suitable distributional family and corresponding GLMM is chosen for each outcome depending on its type. For numeric outcome $Y_{i,j}^r$ we assume *classical normal linear mixed-effects model*: $Y_{i,j}^r \big| \, \mathbf{b}_i^r; \mathcal{C}_{i,j} \sim \mathsf{N}\left(\eta_{i,j}^r, \tau_r^{-1}\right)$, where τ_r is the precision parameter of the error terms. *Logistic regression* is assumed for binary outcomes $Y_{i,j}^r \in \{0, 1\}$, the probability of success is parametrized by $\mathsf{logit}^{-1}\left(\eta_{i,j}^r\right)$. Logit function is also used for parametrization of cumulative probabilities of an ordinal outcome $Y_{i,j}^r \in \{1, \ldots, K_r\}$ in the following way: $p_k := \mathsf{P}\left[Y_{i,j}^r > k | \eta_{i,j}^r\right] = \mathsf{logit}^{-1}(\eta_{i,j}^r - c_{r,k})$, where $\mathbf{c}_r$ is a set of ordered intercepts for outcome $r, \mathsf{t}(r) = \mathsf{O}$. The probability of $Y_{i,j} = k$ is then just a difference of p_{k-1} and p_k. Note that by this parametrization, the *proportional odds assumption* has been employed. However, for general categorical outcome each category level $k \in \{1, \ldots, K_r\}$ is given a special predictor $\eta_{k,i,j}^r$ determined by a different set of fixed effects $\boldsymbol{\beta}_{r,k}$. The probability of attaining level k is then given by a generalization of logit function often called *softmax* function: $\mathsf{P}\left[Y_{i,j}^r = k \big| \, \boldsymbol{\eta}_{i,j}^r\right] \propto \exp\left\{\eta_{k,i,j}^r\right\}$, where the predictor for the last category has to be fixed to 0 for identifiability purposes. Here we limit ourselves with only these four model types, however, an extension to different distributional families or parametrizations would be straightforward.
Given random effects and covariates, the one individual observation $Y =$

$Y_{i,j}^r$ tied with predictor $\eta = \eta_{i,j}^r$ contributes to the likelihood by

$$p_{\mathsf{t}(r)}\left(Y|\,\mathbf{b}_i^r;\,\mathcal{C}_{i,j}\right) = \begin{cases} (2\pi)^{-\frac{1}{2}}\tau^{\frac{1}{2}}\exp\left\{-\dfrac{\tau}{2}\left(Y-\eta\right)^2\right\}, & \text{if } \mathsf{t}(r) = \mathsf{N}, \\ \left[\mathsf{logit}^{-1}\left(\eta\right)\right]^{Y}\left[1-\mathsf{logit}^{-1}\left(\eta\right)\right]^{1-Y}, & \text{if } \mathsf{t}(r) = \mathsf{B}, \\ \mathsf{logit}^{-1}\left(\eta - c_{r,Y-1}\right) - \mathsf{logit}^{-1}\left(\eta - c_{r,Y}\right), & \text{if } \mathsf{t}(r) = \mathsf{O}, \\ \dfrac{\exp\left\{\eta_Y\right\}}{\sum_{k=1}^{K_r}\exp\left\{\eta_k\right\}}, & \text{if } \mathsf{t}(r) = \mathsf{C}. \end{cases}$$

Given the random effects and the covariates, we treat models for different outcomes as independent of each other. However, real data often exhibit strong relationships among outcomes, e.g. households of high disposable income have a higher chance to afford a one week holiday away from home. Therefore, we gather all random effects $\mathbf{b}_i^r, r = 1, \ldots, R$ into a long vector $\mathbf{b}_i$ and assume a centred multivariate normal distribution with a completely general covariance matrix $\mathbf{\Sigma}$. The associations of the random effects are transferred to the marginal distribution of the outcomes.

2.2 Model-based clustering

In order to discover potential different patterns of evolution, we have to assume a certain heterogeneity within the data. We employ a method of unsupervised clustering called *model-based clustering*, which creates a mixture of G outlined models differing in the parameter values, e.g. the fixed effects $\boldsymbol{\beta}_r^{(g)}$ for outcome r specific to cluster $g = 1, \ldots, G$ describe the specific evolution in time or effect of other covariates of this cluster. Potentially, any unknown parameter (τ_r, $\mathbf{c}_r$, $\mathbf{\Sigma}$) could be set to be group-specific, which is denoted by (g) in the superscript.

Let U_i be the unknown latent group allocation indicator with a marginal distribution $\mathsf{P}\left[U_i = g\right] = w_g$, where $0 < w_g < 1$ are unknown probabilities such that $\sum_{g=1}^{G} w_g = 1$. Let f_g stand for the pdf for the distribution of all outcome values $\mathbb{Y}_i$ within the gth cluster. Then, the pdf of the marginal distribution of $\mathbb{Y}_i$ is of the mixture type: $\sum_{g=1}^{G} w_g f_g(\mathbb{Y}_i; \mathcal{C}_i)$. The Bayes rule yields the probability of belonging to cluster g given the observed data: $\mathsf{P}\left[U_i = g|\mathbb{Y}_i; \mathcal{C}_i\right] \propto w_g f_g(\mathbb{Y}_i; \mathcal{C}_i)$.

Unfortunately, the evaluation of f_g includes integration of the latent random effects $\mathbf{b}_i$, which is complicated by the different outcome types and the joint distribution of $\mathbf{b}_i$ across all of the outcomes. In particular, the contribution of a household i to the likelihood function can be expressed as

$$L_i(\boldsymbol{\theta}) = \sum_{g=1}^{G} w_g \int \prod_{r=1}^{R}\prod_{j=1}^{n_i} p_{\mathsf{t}(r)}\left(Y_{i,j}^r \,\middle|\, \mathbf{b}_i^r,\, \boldsymbol{\theta}^{(g)};\, \mathcal{C}_{i,j}\right) p\left(\mathbf{b}_i \,\middle|\, \mathbf{\Sigma}^{(g)}\right) \mathrm{d}\mathbf{b}_i,$$

where $\boldsymbol{\theta}$ stands for all unknown parameters, $\boldsymbol{\theta}^{(g)}$ for parameter values within the group g and $p(\cdot|\cdot)$ denotes the pdf of the corresponding conditional probability distribution, where subscript $\mathsf{t}(r)$ denotes the distributional family depending on the type of the outcome r. This integral can be directly evaluated only for numeric outcomes, $t(r) = \mathsf{N}, \forall r$. To elegantly avoid the integration during the estimation, we switch to the Bayesian framework with the use of *Bayesian data augmentation* principle, which treats latent variables as additional unknown model parameters. This includes the potentially missing outcome values. The prior distribution in some sense regularizes the likelihood and removes potential problems with maximization. Moreover, a suitable choice of the prior distribution over the marginal allocation probabilities $\mathbf{w}$ can solve the problem of apriori unknown number of mixture components G.

2.3 Prior and posterior distribution

Inspired by the work of Frühwirth-Schnatter and Malsiner-Walli (2019), we set the Dirichlet prior over $\mathbf{w}$ so that *sparse finite mixture* is induced. We fix the maximal number of mixture components G_{max}. Given the allocations U_i, we count the current cluster occupancy numbers $n_g = \sum_{i=1}^{n} \mathbf{1}_{(U_i=g)}$, some of which may eventually be empty. Then, the number of non-empty clusters $G_+ := G_{\mathsf{max}} - \sum_{g=1}^{G_{\mathsf{max}}} \mathbf{1}_{(n_g=0)}$ and its posterior is targeted instead. The Dirichlet prior with low parameter values then encourages $G_+ < G_{\mathsf{max}}$ with high probability.

Common distributional families are used (to achieve conjugacy) for the prior distributions of the rest of the unknown parameters. The hyperparameters have to be set carefully, since empty clusters are described by the prior only, which may have an impact on the level of sparsity.

We take an MCMC approach for the estimation of the posterior distribution by sampling a Markov chain from the full-conditioned distributions (Gibbs sampler), where the problematic full-conditionals are replaced by Metropolis proposals of multivariate normal steps with suitably chosen covariance matrix. In practise, the chain starts with G_{max} clusters, some of which are emptied in time, until the chain settles with some G_+ solution. We also apply the recommended post-sampling procedure for relabelling in case the *label-switching problem* occurs.

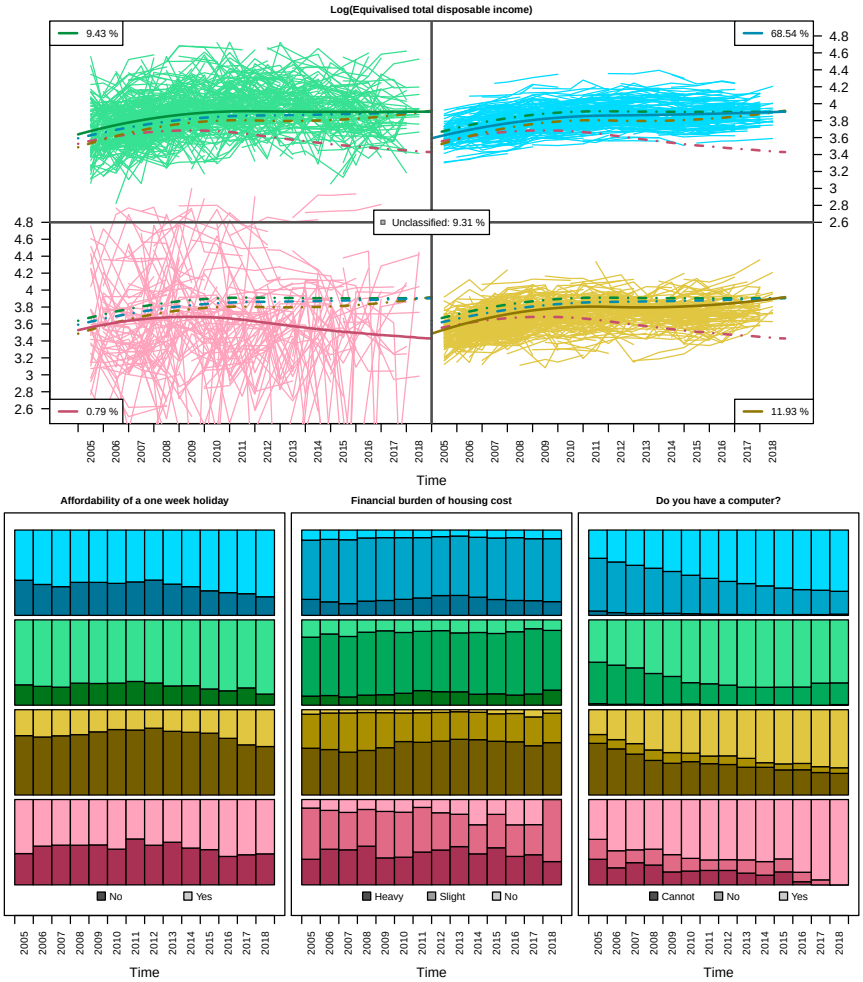


FIGURE 1. Evolution in time of a numeric (top) and categorical (bottom) outcomes grouped into $\widehat{G}_+ = 4$ colourful clusters (red, yellow, green and blue). The top plot of *Equivalised total disposable income* contains estimated median posterior curves. Ratios of observable levels of categorical outcomes (bottom) are displayed for each cluster and year separately.

3 Application to the EU-SILC data

We limit the analysis to the data of $n = 23\,360$ households from the Czech Republic. One outcome per type is taken: *Equivalised total disposable income* (numeric), *Affordability of a one week holiday* (binary), *Financial burden of the total housing cost* (ordinal) and *Do you have a car?* (categorical). We primarily focus on the evolution in time, hence, a spline parametrization of the time is considered for each cluster. Additional regressors such as level of urbanization, equivalised household size, the highest ISCED level attained within the household, etc., are considered to filter out these potential effects common to all households.
The number of non-empty clusters heavily depends on the choice of group-specificity of other unknown parameters. Under a common precision parameter τ for the income, more than 10 clusters remain, some of which are very scarce and of unique trend combinations. On the other hand, the choice of group-specific precision $\tau^{(g)}$ results in only four groups, where the inner variance rather than the actual trend is the determining factor for clustering, see Figure 1. The four discovered groups do not completely fulfil our prior expectations, yet, each of them could be uniquely characterized: declining and extremely volatile with outliers (red), poor but steady (yellow), prosperous but volatile (green), average and steady majority (blue).

Acknowledgments: This research was supported by the Czech Science Foundation (GAČR) grant 19-00015S and the Charles University, project GA UK No. 298120.

References

Vávra, J. and Komárek A. (2022). Classification Based on Multivariate Mixed Type Longitudinal Data with an application to the EU-SILC database. Accepted to *Adv Data Anal and Classif.*

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Adv Data Anal and Classif*, **13**(1), 33 – 64.

Adjusting for spatial confounding using eigendecomposed CAR models

Massimo Ventrucci[1], Garritt Page[2]

[1] Department of Statistical Sciences, University of Bologna, Italy
[2] Department of Statistics, Brigham Young University, USA

E-mail for correspondence: massimo.ventrucci@unibo.it

Abstract: Observational epidemiological studies analyzed via spatial regression methods are often affected by spatial confounding. We formulate an adjustment strategy based on spectral methods, by considering eigendecomposed conditional autoregressive models. We model exposure and confounding variable using different commonly used spatial models, to study the extent to which bias due to spatial confounding can be adjusted under different specifications.

Keywords: BYM; Spatial causal inference; Coherence; Conditionally autoregressive prior

1 Introduction

It is now common for environmental and epidemiological studies to be spatially varying in addition to being observational. A fundamental task is to estimate the effect of a treatment variable (or exposure variable) on a response variable. As an example, Wu et al (2020) noticed that many co-morbidities associated with COVID-19 had connections to being exposed to higher concentrations of ambient fine particular matter ($PM_{2.5}$); see Figure 1. Due to this, they conducted a study to determine if an increase in $PM_{2.5}$ resulted in a higher COVID-19 mortality rate. They found that an increase of 1 $\mu g/m^3$ in ambient fine particulate matter ($PM_{2.5}$) is associated with a 15% increase in the COVID-19 mortality rate.
A key assumption necessary to endow their analysis with a causal interpretation is that no unmeasured (spatial) confounders have been excluded from the model. This assumption, generally speaking, is impossible to verify. However, it may be possible to remove the effects of unmeasured confounders that are spatially structured under certain assumptions. Guan et al (2020) provided a framework where this phenomena can be studied. They propose modeling the exposure and unmeasured confounder in the spectral domain with a joint Gaussian which permitted deriving the coherence function and determine the assumptions necessary to establish a causal interpretation of exposure. Guan et al (2020) focus on the Leroux parametrization of a CAR model. In this work we explore the same idea but on

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

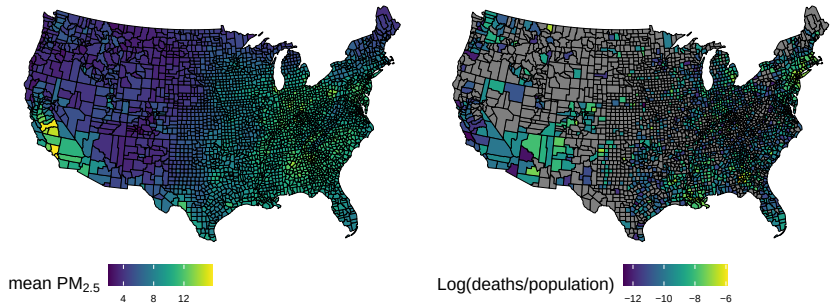


FIGURE 1. **$PM_{2.5}$ exposure and COVID-19 mortality by US county**: (right) Average $PM_{2.5}$ ($\mu g/m^3$) over 2000-2016 and (left) log COVID-19 mortality rate (i.e., log(deaths/population)) through May 12, 2020 (counties with no deaths are shaded gray).

a much broader scale by considering two other popular and commonly employed areal data spatial models. Interestingly, we show that the ability to recover a causal estimate of exposure depends on the model selected to fit the data.

2 Eigendecomposed CAR models

Let $\mathbf{Y} = (Y_1, ..., Y_n)^{\mathrm{T}}$, $\mathbf{X} = (X_1, ..., X_n)^{\mathrm{T}}$ and $\mathbf{Z} = (Z_1, ..., Z_n)^{\mathrm{T}}$ be, respectively, the response, the exposure and the missing confounder at spatial units $1, \ldots, n$. We use the generic spatial regression model

$$\mathbf{Y} = \beta_0 \mathbf{1} + \beta_x \mathbf{X} + \beta_z \mathbf{Z} + \varepsilon. \tag{1}$$

If $\mathbf{Z}$ is observed and model (1) correct, then it is straight forward to estimate the causal effect β_x. In most studies $\boldsymbol{Z}$ is unobserved. Because of this, a reasonable approach would be to model $\boldsymbol{Z}$ and $\boldsymbol{X}$ jointly. We do this following the approach in Guan et al (2020) using spatial processes on the spectral domain. To illustrate the spectral framework consider that both $\boldsymbol{X}$ and $\boldsymbol{Z}$ in model (1) follow an intrinsic CAR model, in the case of $\boldsymbol{Z}$

$$\boldsymbol{Z} \sim \text{Normal}\left(\mathbf{0}, \sigma_z^2 \boldsymbol{\Gamma} \boldsymbol{W}^{-} \boldsymbol{\Gamma}^{\mathrm{T}}\right) \tag{2}$$

where $\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{R}$ is the spectral decomposition of the structure matrix $\boldsymbol{R}$, specifying adjacency relationships between regions. Matrix $\boldsymbol{\Gamma}$ contains eigenvectors and $\boldsymbol{W}$ is a diagonal eigenvalue matrix with kth diagonal element $\omega_k \geq 0$, ordered so that $\omega_1 \leq ... \leq \omega_n$. Note in model (2), $\mathrm{V}(\boldsymbol{Z}) = \sigma_z^2 \boldsymbol{R}^{-}$, where $\boldsymbol{R}^{-} = \boldsymbol{\Gamma}\boldsymbol{W}^{-}\boldsymbol{\Gamma}^{\mathrm{T}}$ and $\boldsymbol{W}^{-}$ has diagonal elements $\{1/\omega_1, \ldots, 1/\omega_n\}$.

We can project $\boldsymbol{X}$ and $\boldsymbol{Z}$ into the spectral domain by applying the transform $\boldsymbol{X}^* = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{Z}^* = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{Z}$. We assume the pairs (X_k^*, Z_k^*) are independent across k, and Gaussian with mean zero and covariance

$$\operatorname{cov}\begin{pmatrix} X_k^* \\ Z_k^* \end{pmatrix} = \begin{pmatrix} \sigma_x^2 f_x(\omega_k) & \rho\sigma_x\sigma_z f_{xz}(\omega_k) \\ \rho\sigma_x\sigma_z f_{xz}(\omega_k) & \sigma_z^2 f_z(\omega_k) \end{pmatrix}, \tag{3}$$

where σ_x^2 and σ_z^2 are variance parameters, $f_x(\omega_k) > 0$ and $f_z(\omega_k) > 0$ are variance functions that determine the covariance of $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively, and scalar $\rho \in [-1, 1]$ and cross spectral density $f_{xz}(\omega_k)$ determine the dependence between $\boldsymbol{X}$ and $\boldsymbol{Z}$. Note, under the intrinsic CAR model assumption (2), $f_j(\omega_k) = \omega_k^{-1}$, $j \in \{x, z\}$.
Finally, by marginalizing the response $\boldsymbol{Y}^* = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{Y}$ over the unknown $\boldsymbol{Z}^*$ we get

$$Y_k^* \mid X_k^* \sim \text{Normal}\left(\beta_0 M_k + \beta_x X_k^* + \beta_z \alpha(\omega_k) X_k^*, \tau^2(\omega_k) + \sigma^2\right), \forall k \tag{4}$$

where M_k is the sum of the kth column of $\boldsymbol{\Gamma}$. According to the type of CAR model that we start from we have different expressions for the terms $\alpha(\omega_k)$ and $\tau^2(\omega_k)$ in (4). However, in general it can be shown that the regression coefficient for X^* is a function of the spatial resolution ω, namely $\beta(\omega_k) = \beta_x + \beta_z\alpha(\omega_k) \neq \beta_x$. Thus, the size of spatial confounding bias is driven by $\alpha(\omega_k)$. In particular, $\alpha(\omega_k) = \gamma(\omega_k)\frac{\sigma_z\sqrt{f_z(\omega_k)}}{\sigma_x\sqrt{f_x(\omega_k)}}$, where $\gamma(\omega_k) = \rho\frac{f_{xz}(\omega_k)}{\sqrt{f_x(\omega_k)f_z(\omega_k)}}$ is the coherence function, which determines the correlations between the two spectral processes.
Given that $\alpha(\omega)$ is not known in practice, the bias cannot in general be identified and removed. Guan et al (2020) propose an identification strategy based on two assumptions:

1. $\gamma(\omega_k) = \rho$ (*parsimonious coherence model*);
2. $\alpha(\omega_k) \to 0$ for large ω_k, i.e. the cross-spectral density decreases to zero faster than the spectral density of X, implying a decrease in confounding in higher frequency or local variations (*unconfoundedness at high-frequencies*).

While assumption (1) is relevant to study the impact of the properties such as smoothness of the spatial processes on the confounding, assumption (2) leads to employing spline based methods to adjust for spatial confounding. Resolution varying coefficient models, where $\beta(\omega)$ is modelled as a linear combination of B-spline basis functions defined over the eigenvalue domain ω, have been proposed in Guan et al (2020). The idea is that under assumption (2) of unconfoundedness at high-frequencies, the fitted value $\hat{\beta}(\omega_n)$, where ω_n is the largest eigenvalue, can be assumed as an unbiased estimate of the exposure effect β_x in (1).
Interestingly, for an intrinsic CAR model it turns out that $\alpha(\omega_k) = \rho\sigma_z/\sigma_x$ which is constant as a function of ω_k. Notice then that it can never go to zero at high-frequencies, which means that this model cannot provide adjustment for spatial confounding under assumption (2).

2.1 Besag York and Mollié model

Let's consider other CAR models that are based on a mix of a spatially structured and an unstructured component, known as BYM (Besag York and Mollié, 1991). Dean et al (2001) and Leroux et al (2000) proposed different types of BYM models. Under Dean parametrization (we show the formula for $\boldsymbol{Z}$ only)

$$\text{Var}(\boldsymbol{Z}) = \sigma_z^2 \boldsymbol{\Gamma} \left[(1-\lambda_z)\boldsymbol{I}_n + \lambda_z \boldsymbol{W}^- \right] \boldsymbol{\Gamma}^{\text{T}}. \tag{5}$$

The parameter λ_z in (5) indicates the proportion of variance attributed to spatially structured random effects.
Under Leroux, the covariance is $\text{Var}(\boldsymbol{Z}) = \sigma_z^2 \boldsymbol{\Gamma} \left[(1-\lambda_z)\boldsymbol{I}_n + \lambda_z \boldsymbol{W} \right]^{-1} \boldsymbol{\Gamma}^{\text{T}}$, where λ_z is simply a spatial smoothing parameter with no more interpretation in terms of explained variance. The conditional model in (4) takes a different form according to the kind of BYM considered (assuming we have defined the same BYM for both $\boldsymbol{X}$ and $\boldsymbol{Z}$). It can be shown that under the parsimonious coherence model (assumption (1)), for Dean model

$$\alpha(\omega_k)_D = \rho \frac{\sigma_z}{\sigma_x} \sqrt{\frac{1-\lambda_z+\lambda_z/\omega_k}{1-\lambda_x+\lambda_x/\omega_k}},$$

while for Leroux

$$\alpha(\omega_k)_L = \rho \frac{\sigma_z}{\sigma_x} \sqrt{\frac{1-\lambda_x+\lambda_x\omega_k}{1-\lambda_z+\lambda_z\omega_k}}.$$

We can notice that if $\lambda_z > \lambda_x$ then $\alpha(\omega_k)$ decreases as ω_k gets larger. Therefore, reducing bias due to spatial confounding is possible under assumption (2) of unconfoundedness at high-frequencies both using Dean and Leroux BYM models, but only when $\boldsymbol{Z}$ is smoother than $\boldsymbol{X}$. Interestingly, preliminary investigation suggests that $\alpha(\omega)_D > \alpha(\omega)_L$ when $\lambda_z > 1 - \lambda_x$.

3 Discussion

Modelling exposure and missing confounder in the spectral domain allows us to study the extent to which spatial confounding bias can be adjusted under CAR models. By looking at the eigendecomposed version of each CAR, we were able to highlight how the smoothness properties of both exposure and missing confounder have an impact in reducing spatial confounding bias. Further, the BYM permits recovering the causal estimate β_x under the spectral framework by Guan et al (2020) (assumptions (1) and (2)) when $\lambda_z > \lambda_x$, while it is not clear that the intrinsic CAR provides any adjustment.

References

Besag J., York J., and Mollié A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**(1): 1–20.

Guan, Y., Page, G. L., Reich, B. J., Ventrucci, M., and Yang, S. (2020). A spectral adjustment for spatial confounding. *arXiv preprint arXiv:2012.11767*.

Dean C., Ugarte M., and Militino A. (2001). Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics* **57**(1): 197–202.

Leroux B.G., Lei X., and Breslow N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. *In Statistical Models in Epidemiology, the Environment, and Clinical Trials.*: 179–191.

Wu X, Nethery R, Sabath M, Baun D, and Dominici F. (2020). Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Science Advances*, **6**, eabd4049.

Hierarchical graphical modelling of microbiome interactions in related environments

Veronica Vinciotti[1], Ernst Wit[2]

[1] Department of Mathematics, University of Trento, Italy
[2] Università della Svizzera italiana, Lugano, Switzerland

E-mail for correspondence: veronica.vinciotti@unitn.it

Abstract: Unraveling interactions between microbial communities is of vital importance in understanding how microbes influence human health. Rich sources of microbiome data have been generated by the latest sequencing experiments, measuring microbial abundances under a variety of environmental conditions, such as at different body sites or across different time points. In this paper, we model the complexity of these data, and of the underlying dependency structure, via a Gaussian copula graphical model. Heterogeneity in the data is captured both at the individual microbial level, via marginal distributions that are linked parametrically with external covariates, and at the dependency level, with a hierarchical prior on the graph. We develop an efficient Bayesian structural learning procedure for parameter inference and, inspired by the microbiome application, we propose a latent space network prior for capturing structural relatedness across multiple environments.

Keywords: Microbiome; Copula graphical models; Bayesian structural learning.

1 Microbiome data: sparse, discrete, compositional, heterogeneous

Interactions between microbes are fundamental in shaping the structure and functioning of the human microbiota, and their malfunctioning has been linked to a number of medical conditions. Learning these interactions is thus of great interest and is made possible by a rich source of microbiome data that has been recently generated and made available from large consortia, such as the data from the Human Microbiome Project (HMP Consortium, 2012) that we use in this paper.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

In particular, we consider the abundances of 87 Operating Taxonomic Units (OTUs) - these are the higher level microbial communities in which microbes are clustered - measured in 4503 micriobiomes from healthy individuals and across 13 different body sites. Graphical modelling approaches for network inference from microbiome data are made difficult by the complexity of the data. Indeed, some typical features are:

- High-dimensionality: the raw data contains the abundances of 10730 OTUs, though many of these are extremely rare
- Discreteness: data generated from 16S variable region V3-5 sequencing technologies are in the form of counts
- Zero inflation: a high percentage of zeros. As an indication, Figure 1 (left) is a boxplot of the percentage of zeros for each OTU, split by body sites.

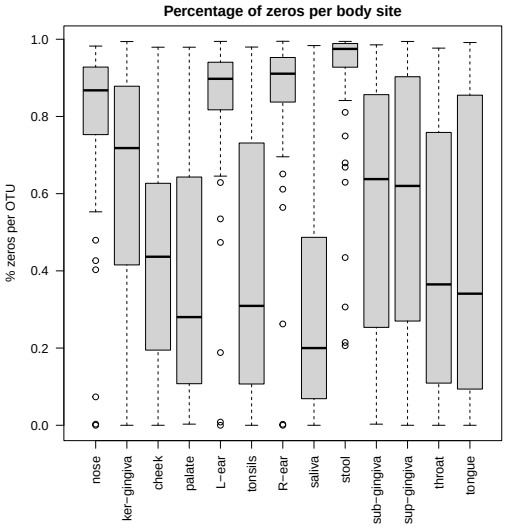


FIGURE 1. Percentages of zeros across the 87 OTUs split by body sites

- Compositionality: sequencing depths change significantly between samples due to experimental effects.

2 Hierarchical graphical model

In this section we define a hierarchical graphical model for network inference from heterogeneous data. Let $\mathbf{Y} = (Y_1, \ldots, Y_p)$ be the random vector of

interest, e.g. the abundances of p OTUs. We assume $\mathbf{Y}$ to be distributed according to some graphical model (GM),

$$\mathbf{Y}|G \sim GM(G),$$

relative to some conditional independence graph G, which itself is distributed according to some prior model,

$$G \sim P(G; \boldsymbol{\theta}),$$

for some vector of parameters $\boldsymbol{\theta}$. Depending on the context of interest, there are various inference tasks that could be considered. For example, given some data on $\mathbf{Y}$, one can learn about the drivers of the conditional independence graph, $\boldsymbol{\theta}|$data. Or given some data on $\mathbf{Y}$, one can use the prior to capture particular structures in the underlying conditional independence graph. For example,

1. If the data is a time-series on $\mathbf{Y}$, then consider a temporal prior on the graph, $G_t|G_{t-1}, \boldsymbol{\theta}$.
2. If the data concerns a number of related conditions on $\mathbf{Y}$, consider priors that capture relatedness of the conditions.

The type of graphical model and the type of hierarchical prior depend on the situation under consideration. As for the graphical model, we are particularly interested in the Gaussian copula graphical model, due to its easy mathematical formulation and its high applicable format, particularly for data that are not Gaussian such as the count microbiome data. Thus, we consider:

$$P(Y_1 \leq y_1, \ldots, Y_p \leq y_p) = \Phi_p(\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F_p(y_p))|\mathbf{R}),$$

where Φ_p is the cumulative distribution function of a p-dimensional multivariate normal with a zero mean vector and correlation matrix $\mathbf{R}$, while Φ is the standard univariate normal distribution function. Here, the dependency structure is captured by the inverse of the correlation matrix $\mathbf{K} = \mathbf{R}^{-1}$, typically called the precision or concentration matrix. Indeed, the zero patterns in this matrix define the conditional independence graph in the latent Gaussian space, following from the theory of Gaussian graphical models (Lauritzen, 1996). As for the hierarchical prior, we consider a latent space network model, i.e. $G \sim LSM(\mathbf{z})$ for some vector of latent variables $\mathbf{z}$. Inspired by the microbiome application, we define this model with the aim of capturing relatedness between multiple environments, as we explain in the next section.

3 Gaussian copula model for microbiome data

Microbiome data are heterogenous in many ways. For the data that we consider, two external covariates are of interest and may help in capturing this

heterogeneity. The first one is the library size. This is a normalizing factor that accounts for the compositionality of the data and that is typically estimated (offline), for a given sample, by the geometric mean of pairwise ratios of OTU abundances of that sample with all other samples (Cougoul et al, 2019). The second one contains the information about the body site where the biological sample was collected. This results in a categorical variable with 13 levels. The method that we propose can account for the heterogeneity in the data in two ways. Firstly, by linking the marginal distributions F_j to the covariates, and secondly, by allowing the dependency structure $\mathbf{K}$ to depend on the covariates.
As for the marginals, we adopt discrete Weibull (DW) regressions, i.e, we assume that

$$F_j(y_j|\mathbf{X}=\mathbf{x}) = 1 - q_j(\mathbf{x})^{(y_j+1)^{\beta_j(\mathbf{x})}}, \text{ for } y_j = 0, 1, \ldots,$$

with

$$\log\Big(\frac{q_j(\mathbf{x})}{1-q_j(\mathbf{x})}\Big) = \mathbf{x}^t\boldsymbol{\theta}_j, \quad \log(\beta_j(\mathbf{x})) = \mathbf{x}^t\boldsymbol{\gamma}_j,$$

where $\boldsymbol{\theta}_j$ and $\boldsymbol{\gamma}_j$ denote the regression coefficients linking the Y_j marginal component of the model to the external covariates $\mathbf{x}$. As in Vinciotti et al (2022), we find that DW fits significantly better than the commonly used negative Binomial distribution also on these data. For example, Figure 2 shows the fitting of a DW regression model with library size and body site used as covariates, for each OTU. For both distributions, we consider the addition of a zero-inflated component, via a constant zero-inflation parameter, if it produces a smaller BIC compared to the non-zero inflated version of the model.
The expectation is that the underlying network may vary across the different body sites, possibly with a high similarity between the networks among the different conditions. For this reason, a second level of heterogeneity is introduced by setting a latent space prior on the graph. This will capture the tendency of two nodes to be connected in a particular condition, mediated by the vicinity of that condition with other conditions. In particular, we consider the following latent space model

$$P(G_{ijk}=1) = \Phi\Big(\mathbf{z}_i^t\mathbf{z}_j \sum_{k\neq k'} \mathbf{c}_k^t\mathbf{c}_{k'}\Big), \tag{1}$$

where $G_{ijk}=1$ denotes the presence of an edge between node Y_i and node Y_j in body site k, $\mathbf{z}_1, \ldots, \mathbf{z}_p \in \mathbb{R}^2$ are the latent space variables for each node and $\mathbf{c}_1, \ldots, \mathbf{c}_B \in \mathbb{R}^2$ the latent space variables for each condition. In both cases, we opt for a 2D-latent space.
We have developed an efficient Bayesian inferential procedure, that consists of the following steps:

1. Fitting of marginal regression models, with OTU count abundances as response variable and library size as covariate. The fitted DW

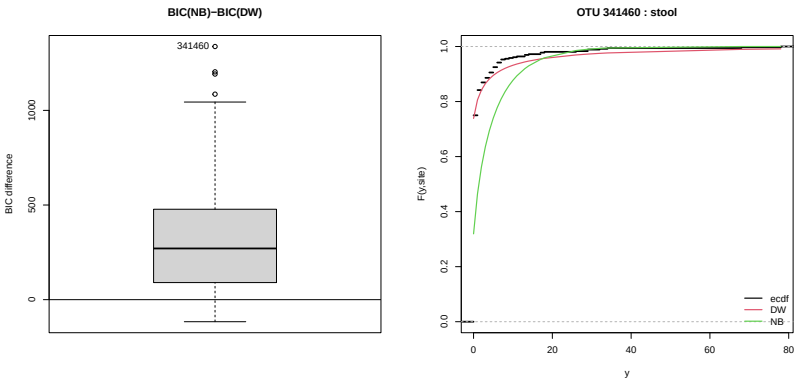


FIGURE 2. Left: Difference of BIC between negative Binomial and discrete Weibull marginals. Right: Comparison of empirical cdf, fitted DW and NB marginals for the abundance of OTU 341460 within the *stool* body site (averaged across the normalizing factor).

marginals define the regions for sampling the latent W_j where the Gaussian graphical model resides:

$$\mathcal{D}_F(\mathbf{y}) = \left\{\mathbf{w} \in R^{n\times p} : \Phi^{-1}\big(\hat{F}_j(y_{ij}-1)\big) < w_{ij} < \Phi^{-1}\big(\hat{F}_j(y_{ij})\big)\right\}$$

2. Given the current graph G, the latent space model $(\mathbf{z}, \mathbf{c})$ is fitted via a Bayesian probit Gibbs sampling, returning updated posterior graph probabilities that are used as prior for the next step

3. Given the sampled $\mathbf{w}$ and the latent variables, the edges are independent. We make use of this in the Bayesian structural learning procedure for sampling the next graph G, using the efficient search algorithm of Mohammadi and Wit (2015).

Iterating step 2 and 3 provides samples from the posterior distribution.

4 Simulation

We conclude with a simulation study showing the potential of the proposed method. We simulate data with $B = 15$ conditions and $p = 20$ variables. The simulation of the data involves the following steps:

1. Setting the latent node variables $\mathbf{z}$, by taking a sample from independent standard normals. For the latent condition variables $\mathbf{c}$, we draw these in three groups of five from normal distributions, with mean (-1,-2), (0,0) and (2,2), respectively, and standard deviations 0.2. In this way, we create three groups of similar conditions.

2. Fitting a probit model to calculate the edge probabilities for each condition given $\mathbf{z}$ and $\mathbf{c}$, and using these probabilities to sample a graph G for each condition via independent Bernoulli draws

3. Sampling a precision matrix $\mathbf{K}$ associated to the graph in each condition via a G-Wishart distribution, $\mathbf{K} \sim W_G(3, \mathbb{I}_p)$, and thus sampling multivariate Gaussian data of size $n = 100$ for each condition.

Figure 3 shows the results of the simulation after 100k MCMC iterations. The left plot shows how the posterior edge probabilities, either estimated from the MCMC chain of graphs or from the estimated latent space (eq. 1), are quite close to the true ones used to simulate the graph. The right plot shows the estimated latent variables of the conditions ($\mathbf{c}_k$), with the three known groups of conditions represented by the three different colours and only partly recovered.

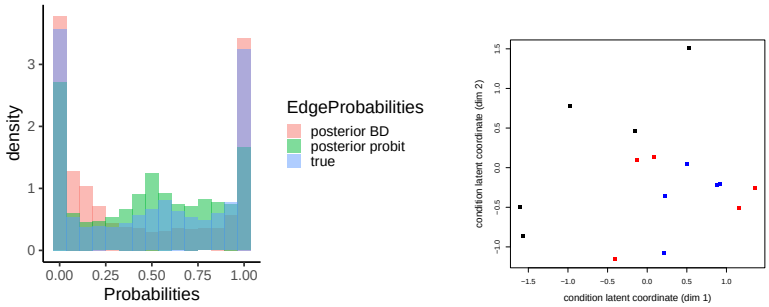


FIGURE 3. Results of the simulation study. Left: Posterior edge probabilities compared with true probabilities. Right: Posterior latent variables for the 15 conditions ($\mathbf{c}_k$) coloured according to the three known groups.

References

Cougoul, A., X. Bailly, and Wit, E. (2019). MAGMA: inference of sparse microbial association networks. *bioRxiv*: **538579**.

HMP Consortium (2012). A framework for human microbiome research. *Nature*, **486**, 215 – 221.

Lauritzen, S. (1996). Graphical Models. Oxford: Clarendon Press.

Mohammadi, R. and Wit, E. (2015). Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian analysis*, **10(1)**, 109 – 138.

Vinciotti, V., Behrouzi, P. and Mohammadi, R. (2022). Bayesian structural learning of microbiota systems from count metagenomic data. *arXiv:* **2203.10118**.

Estimating quality-adjusted life-years: Assessing the bias induced for a terminal decline quality of life model

Alexandra Welsh[1], Deborah A. Costain[1], Andrew C. Titman[1]

[1] Dept. of Mathematics and Statistics, Lancaster University, UK

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

Abstract: The quality-adjusted life-year (QALY) is a summary measure used to evaluate the effectiveness of medical treatments in terms of both quality and length of life. One method used to estimate QALYs is the area under the time-utility curve (AUC). However, this approach may induce bias, due to its inability to capture the dependency between the quality of life measures and the survival time. A simulation study is conducted to assess the bias induced when estimating QALYs using the AUC method, using a terminal decline data pattern including censored individuals and missing responses.

Keywords: Quality-adjusted life-years; Joint longitudinal-survival models.

1 Introduction

To allocate healthcare resources effectively, the financial cost and health outcomes associated with an intervention must be evaluated. Cost-effectiveness analyses should use outcomes which incorporate the impact of the treatment on both the length of life and health-related quality of life (HQoL). The quality-adjusted life-year (QALY) is one such summary measure; one QALY is equivalent to one year of life in perfect health.
A patient's HQoL can be assigned to a health state, with a utility value to indicate its desirability; usually between 0 = Death and 1 = Perfect health. QALYs are calculated as the product of the time spent in a HQoL state and its corresponding utility value.
QALYs can be estimated using an area under the curve (AUC) method, where the longitudinal HQoL data points are linearly interpolated between observation times, with the value 0 taken after the time of death. However,

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

summary measures such as the AUC may result in biased estimates, especially in the presence of missing data (Bell et al., 2014). The AUC method does not take into account the dependence between the HQoL observations and the survival process, which may be one cause of bias.
In our previous study (Welsh et al., 2020), we assessed the bias induced when using the AUC approach to estimate QALYs for subjects with linear HQoL responses and dependent survival data. The QALY estimates for the random intercept (RI) and random intercept random slope (RIRS) HQoL models were slightly biased when data was complete, but when including censoring or missing responses the bias increased significantly.
Building upon our previous results, we aim to determine under which circumstances the AUC approach for estimating QALYs can be biased when subjects' survival is incorporated into a more complex piecewise linear HQoL model. A simulation study is conducted on terminal decline longitudinal HQoL data with varying censoring and missingness mechanisms.

2 Simulation Study

2.1 Methodology

For each study parameter set, 1000 dependent HQoL and survival datasets were simulated. The data was generated for $n = 100$ subjects per iteration, with longitudinal HQoL responses denoted by Y_{ij}, where $i = 1, \ldots, n$ and $j = 0, 1, \ldots, 10$ are the subject and time indices, respectively.
The data was simulated using the joint longitudinal-survival model framework introduced by Wulfsohn and Tsiatis (1997). It incorporates a linear mixed effects model and a Cox proportional hazards model, as shown:

$$Y_{ij} = (\beta_0 + \nu_{0i}) + (\beta_1 + \nu_{1i})t_j + \epsilon_{ij}\,, \tag{1}$$

$$h_i(t\,;\boldsymbol{\nu}_i) = h_0(t)\exp(\gamma_0(\beta_0 + \nu_{0i}) + \gamma_1(\beta_1 + \nu_{1i}))\,. \tag{2}$$

In Equation 1, $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ are the fixed effects and subject-specific random effects, respectively, and ϵ_{ij} is an independent error term. In Equation 2, $h_0(t)$ is the baseline hazard function, and $\boldsymbol{\gamma}$ are the dependence parameters such that each subject's event time is affected by their personal HQoL intercept and slope values.
To simulate data with a terminal decline pattern, this framework was applied multiple times. Firstly, the pre-progression longitudinal observations were generated using Equation 1, and the subject-specific intercepts and slopes were used in Equation 2 to simulate progression and initial survival times. If the survival time was less than the progression time for a subject, the subject was deemed to have died before progressing to terminal decline and their survival time was recorded. For the remaining subjects a post-progression slope, dependent on their progression time, and a final survival time, dependent on said post-progression slope, were then generated. The

final survival times were then recorded as event times, and the longitudinal HQoL data was adjusted to reflect the post-progression slope.
The pre-progression HQoL and survival models and the progression model retained the same parameters throughout the study. Two post-progression HQoL and survival models were used: a "fast decline" and a "slow decline" model, which allowed the average number of terminal decline HQoL observations to vary. Within these models, only the post-progression dependence was varied, to create independent, "weak" dependence, or "strong" dependence. The parameter choices used are shown in Table 1.

	Variable	**Value(s)/Distribution**	
Pre-Prog.	HQoL fixed effects	(0.8, -0.01)	
	HQoL random effects	$\text{MVN}_2(\mathbf{0}, \mathbf{\Sigma})$ $\sigma_0 = 0.05, \sigma_1 = 0.001, \rho = 0.2$	
	Error	$\text{N}(0, 0.01^2)$	
	Baseline hazard	$\text{Weib}(1.2, 24.28)$	
	Dependence	(0.2, -1)	
Prog.	Baseline hazard	$\text{Weib}(1.2, 6.38)$	
	Dependence	(0.02, -1)	
		SD	**FD**
Post-Prog.	HQoL fixed effect	-0.05	-0.11
	HQoL random effect	$\text{N}(0, 0.005^2)$	$\text{N}(0, 0.011^2)$
	Dependence	{0, -2.5, -5}	{0, -0.55, -1.1}
	Baseline hazard	$\text{Weib}(0.75, 3.36)$	$\text{Weib}(0.65, 0.73)$

TABLE 1. The variables used in the terminal decline HQoL model. Prog. = Progression. SD = slow decline and FD = fast decline.

Three censoring mechanisms were considered in the study: uniform, early and late, with 34%, 56% and 25% of subjects censored, respectively.
The impact of missing HQoL data was also considered. Response missingness was either random or dependent on the value of the previous HQoL observation, and thus followed a missing completely at random (MCAR) or missing at random (MAR) data pattern, respectively. Both patterns resulted in 20% missing observations. Imputation through last observation carried forward (LOCF), and no imputation were considered to handle missing responses.
We estimated the QALYs gained using the method developed by Glasziou et al. (1998). The mean QALY restricted to time L is defined as

$$\text{QALY}_L = \int_0^L P(t)Q(t)\,\mathrm{d}t\,,$$

where $P(t)$ is the proportion of subjects alive at time t, and $Q(t)$ is the

mean HQoL of those subjects at time t. The Kaplan-Meier estimator is used to approximate $P(t)$; $Q(t)$ is estimated by interpolating the HQoL for each individual and combining to yield a group mean function.
To estimate $Q(t)$, longitudinal responses for each subject are required at all observation times and at distinct censoring and survival times up to and including their own event time. For those individuals who experienced the event, three methods were considered to estimate the HQoL at their recorded survival time: LOCF, extrapolation based on a linear regression model, and linear interpolation between the last observation and 0. Each censored subject required a response at time $\max(t) = 10$; this response was estimated for the individual using either LOCF, or extrapolation based on a linear mixed effects model. From this point, responses could be linearly interpolated for each subject at all times necessary.

2.2 Study Results

The study was completed using four scenarios of increasing complexity: complete uncensored data; complete censored data; uncensored data with missing responses; censored data with missing responses. For comparability the results are reported as proportional bias, equal to the difference of the model and population QALYs divided by the population QALY.
Due to the method of estimating the population QALY, using interpolation to 0 to estimate death time observations consistently lead to a lower proportional bias than using LOCF or extrapolation at death time when varying any other data or QALY estimation model choices. For the slow decline data, this resulted in QALY underestimates.
Data with fast post-progression decline resulted in a higher estimate of QALYs than for the slow decline data, across all other data and model variables. When combined with LOCF or extrapolation to estimate death time observations, this resulted in a large overestimate. However, the median proportional bias when interpolation to 0 was used at death time was close to 0; the true HQoL data may not have been modelled accurately, but the QALY under- and over-estimation compensate for one another.
Boxplots of the proportional bias for the complete data with uniform censoring are shown in Figure 1; there was no significant difference in results between the uniform, early, and late censoring patterns. The method used to estimate the HQoL at $\max(t)$ for censored subjects has very little impact on the bias induced, across both post-progression dependence parameter choice and post-progression slope.
Boxplots of the proportional bias for the uncensored data with MCAR responses are shown in Figure 2; there was no significant difference in results between the MCAR and MAR response patterns. The effect of the missing response imputation method on the proportional bias induced is noticeable: the bias is greater when LOCF is used to impute missing observations, with median differences ranging from 0.0034 to 0.0280.

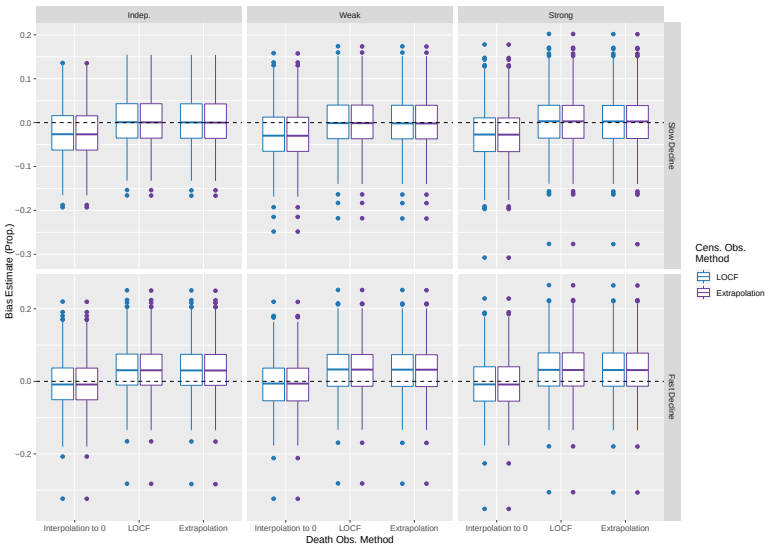


FIGURE 1. The proportional bias induced during QALY estimation for uniformly censored data. Column facets refer to the dependence parameter choice.

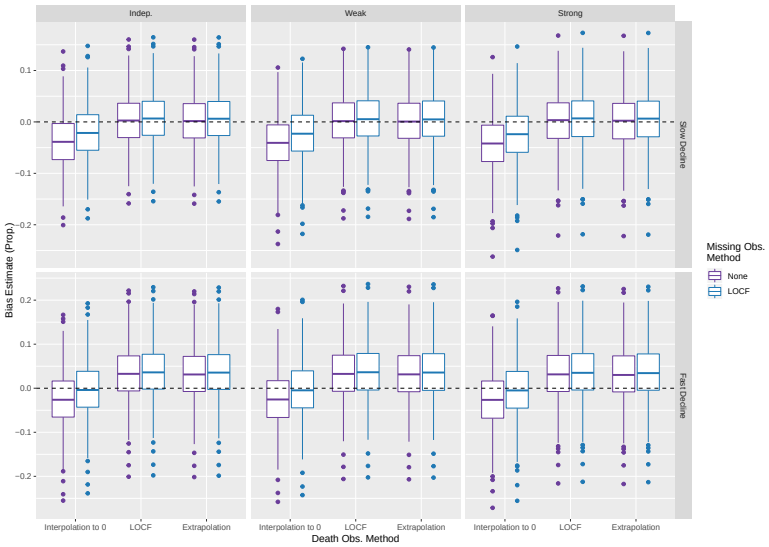


FIGURE 2. The proportional bias induced during QALY estimation for MCAR data. Column facets refer to the dependence parameter choice.

3 Discussion and Future Work

Combining the current study and that presented in Welsh et al. (2020), we have assessed the bias induced by the AUC method for estimating QALYs when using three data patterns for the longitudinal HQoL responses: RI, RIRS, and terminal decline. To expand fully upon the study by Bell et al. (2014), there are two data patterns remaining to be simulated: a plateau and a temporary decline. When completed, we will have a better understanding of where current QALY estimation methods using the AUC may be lacking when used with piecewise linear HQoL data models.
One approach proposed as an alternative to AUC for the estimation of QALYs is joint longitudinal-survival modelling (Rizopoulos, 2012). By fitting a joint model to the data, QALYs can be estimated by integrating the model over the survival times. Li et al. (2013) proposed a joint model, which makes use of a "reverse" time scale, and applied it to QALY estimation.
In our future work, we aim to investigate the potential benefits of using joint modelling approaches, rather than the AUC method, to estimate QALYs from dependent longitudinal HQoL and survival data.

Acknowledgments: Special thanks to the Economic and Social Research Council (ESRC) for their financial support of this project [ES/P000665/1].

References

Bell, M. L., King, M. T., and Fairclough, D. L. (2014). Bias in area under the curve for longitudinal clinical trials with missing patient reported outcome data: Summary measures versus summary statistics. *SAGE Open*, **4**(2), 2158244014534858.

Glasziou, P. P., Cole, B. F., Gelber, R. D., Hilden, J., and Simes, R. J. (1998). Quality adjusted survival analysis with repeated quality of life measures *Statistics in Medicine*, **17**, 1215 – 1229.

Li, Z., Tosteson, T.D., Bakitas, M.A. (2013). Joint modeling quality of life and survival using a terminal decline model in palliative care studies. *Statistics in Medicine*, **32**, 1394 – 1406.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data with Applications in R.* CRC Press.

Welsh, A., Costain, D., and Titman, A. (2020). Bias induced during the estimation of quality-adjusted life-years. *Paper presented at 35th International Workshop on Statistical Modelling, Bilbao, Spain.*

Wulfsohn, M. S., and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**(1), 330 – 339.

Bayesian Boosting for Simultaneous Estimation and Selection of Fixed and Random Effects in High-Dimensional Mixed Models

Boyao Zhang[1], Colin Griesbach[1], Elisabeth Bergherr[1]

[1] Chair of Spatial Data Science and Statistical Learning, Faculty of Business and Economics, Georg-August-Universität Göttingen, Göttingen, Germany

E-mail for correspondence: boyao.zhang@uni-goettingen.de

Abstract: Selection of relevant fixed and random effects without prior choices made from possibly insufficient theory is important in mixed models. Inference with current boosting techniques suffers from biased estimates of random effects and the inflexibility of random effects selection. This paper proposes a new inference method "BayesBoost" combining Bayesian methods and gradient boosting which performs estimation and selection of fixed and random effects in mixed models simultaneously. The method introduces a novel selection strategy for random effects, which allows for computationally fast selection of random slopes even in high-dimensional data structures. Additionally, the new method not only overcomes the shortcomings of Bayesian inference in giving precise and unambiguous guidelines for the selection of covariates by benefiting from boosting techniques, but also provides Bayesian ways to construct estimators for the precision of parameters such as variance components or credible intervals, which are not available in conventional boosting frameworks. The effectiveness of the new approach can be observed via simulation and in a real world application.

Keywords: Bayesian inference; Boosting; Linear mixed models; Probing; Variable selection.

1 Introduction

Linear mixed models (LMMs) are widely used in longitudinal data analysis as they incorporate random effects to deal with group-specific heterogeneity. Bayesian statistics can be used to make inference for LMMs, but it lacks unambiguous ways to perform variable selection. The model technique,

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

boosting and especially componentwise gradient boosting is famous for the straightforward variable selection procedure. But the outcomes brought by boosting are solely estimation points and standard parametric hypothesis test is impossible without bootstrap or other resampling techniques.
We therefore propose a novel inference method combining Bayesian inference and boosting technique, which benefits from the shrinkage and variable selection properties of boosting and from the uncertainty estimates of Bayesian inference.

2 Methods

2.1 Model specification

For clusters or individuals $i = 1, \ldots, m$ with $n = \sum_{i=1}^{m} n_i$, where n_i denotes the replicates of the i-th individual, consider the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

and its predictor $\boldsymbol{y} = \boldsymbol{\eta} = \sum_{k=1}^{p} \boldsymbol{\eta}_k$, with $\boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{Z}_k\boldsymbol{\gamma}_k$, where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are $n \times (p+1)$- and $n \times (q+1)$-dimensional design matrices, $\boldsymbol{\beta}$ is a vector of fixed effects with intercept, $\boldsymbol{\gamma}$ is a vector of cluster-specific random effects with random intercept and $\boldsymbol{\epsilon}$ is a vector of errors.
We assume the independency between $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ with positive definite covariance matrices. The covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ are block-diagonal with $\boldsymbol{R} = \text{blockdiag}(\sigma^2\boldsymbol{\Sigma}_{n_1}, \cdots, \sigma^2\boldsymbol{\Sigma}_{n_m})$, $\boldsymbol{G} = \text{blockdiag}(\boldsymbol{Q}, \cdots, \boldsymbol{Q})$, where $\boldsymbol{\gamma}_i \sim N(\boldsymbol{0}, \boldsymbol{Q})$ with $(1+q) \times (1+q)$-covariance matrix $\boldsymbol{Q}$. For i.i.d. errors, $\boldsymbol{R}$ simplifies to $\boldsymbol{R} = \sigma^2\boldsymbol{I}$.

2.2 Bayesian boosting inference method

The additive predictor contains the fixed and random parts, and each part can be estimated separately by treating the others as an offset. Specifically, the first step is to estimate fixed effects following a componentwise gradient boosting routine, and the second step is to make Bayesian inference by setting the estimated fixed effects as offsets.
In componentwise gradient boosting, the negative gradient vector with a loss in boosting iteration s is fitted by each base-learner $h(X_k)$ for $k = 1, \ldots, p$. Select the best-fitting k^*-th base-learner based on their contribute to the model and update fixed effects $\hat{\boldsymbol{\beta}}^{[s]} = \hat{\boldsymbol{\beta}}^{[s-1]} + \nu\hat{\boldsymbol{\beta}}^{[s]}_{k^*}$, where ν denotes a step-length or learning rate.
After obtaining the estimated fixed effects, the full Bayesian inference for the parameters of interest is based on the posterior distribution for the parameters of interest is based on the posterior distribution

$$p(\boldsymbol{\gamma}, \boldsymbol{G}, \boldsymbol{R}|\tilde{\boldsymbol{y}}) \propto p(\tilde{\boldsymbol{y}}|\boldsymbol{\gamma}, \boldsymbol{G}, \boldsymbol{R})p(\boldsymbol{\gamma}|\boldsymbol{G})p(\boldsymbol{G})p(\boldsymbol{R}),$$

with $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{[s]}$ treating fixed effects as an offset term, and $\boldsymbol{\gamma}|\boldsymbol{G}$, $\boldsymbol{R}$ and $\boldsymbol{G}$ are assumed to be independent. In general, it cannot be displayed in a closed form, such that the full Bayesian inference is usually conducted through MCMC simulation.
Usually, we set a Gaussian prior for the random effects, the full conditional distribution is then a Gaussian $N(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ with parameters

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}, \quad \boldsymbol{\mu}_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\left(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{[s]}\right)\right).$$

The covariance matrix $\boldsymbol{R}$ is dominated by σ^2, and a weakly informative inverse gamma prior $\sigma^2 \sim IG(a, b)$ with small a and b is commonly proposed. The full conditional density of σ^2 is thus an inverse gamma $IG(\tilde{a}, \tilde{b})$ with

$$\tilde{a} = a + \frac{n}{2}, \quad \tilde{b} = b + \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{[s]} - \boldsymbol{Z}\boldsymbol{\gamma}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{[s]} - \boldsymbol{Z}\boldsymbol{\gamma}\right).$$

Analogously, we assume an inverse Wishart prior $IW(v_0, \boldsymbol{\Lambda}_0)$ for the individual covariance matrix $\boldsymbol{Q}$ and get the posterior distribution, which is again an inverse Wishart with

$$v = v_0 + m, \quad \boldsymbol{\Lambda} = \boldsymbol{\Lambda}_0 + \boldsymbol{\gamma}^T\boldsymbol{\gamma}.$$

Bayesian inference for these unknown parameters are thus made according to the MCMC samples drawn from the full conditional distributions.

2.3 Random effects selection and early stopping

The componentwise gradient boosting routine discussed above contains already the fixed effects selection procedure, i.e. select the best-fitting k^*-th covariate according to their improvement to the model. Moreover, we can benefit from the selected covariate and make the random effects selection possible by comparing the model improvement of the fixed effect and that of the random effect of the selected k^*-th covariate. Note that the random effects structure as well as their corresponding design matrix $\boldsymbol{Z}$ should reconstructed by considering the random effects candidates. Accounting for the computing efficiency, we assume that only the variables that already have fixed effects can have random effects. Therefore, the new random effects structure is constructed by taking the best-fitting variable into account, whose random effect, however, is not selected into the model.
Due to the stochasticity of MCMC simulation, the common information criterion is not suitable for determining the stopping iteration. We suggest to use probing to prevent overfitting, the main idea of which is adding artificial non-informative variables to the data to benefit from the presence of variables that are known to be independent from the outcome. Practically, the algorithm stops when any of these permuted non-informative variables selected into the model.
The proposed algorithm is shown below:

Algorithm 1 Bayesian Boosting for Linear Mixed Models

Initialization
for Boosting iterations $s = 1$ to m_{stop} **do**
  Compute the negative gradients $\boldsymbol{u}^{[s]} = \frac{\partial}{\partial \boldsymbol{\eta}} \rho(\boldsymbol{y}, \boldsymbol{\eta}) \Big|_{\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}^{[s-1]}(\boldsymbol{X})}$
  Fit the negative gradient vector $\boldsymbol{u}^{[s]}$ separately to every base-learner $\hat{h}_k^{[s]}(x_k)$ for $k = 1, \ldots, p$, and in linear case $\hat{h}_k^{[s]}(x_k) = \boldsymbol{X}_k \hat{\boldsymbol{\beta}}_k$
  Select the component k^* that best fits the negative gradient vector
  Update fixed effects $\hat{\boldsymbol{\beta}}^{[s]} = \hat{\boldsymbol{\beta}}^{[s-1]} + \nu \hat{\boldsymbol{\beta}}_{k^*}^{[s]}$
  for MCMC iteration $t = 1$ to T **do**
    Draw samples for the random effect $\gamma^{(t)}$, variance $\sigma^{2(t)}$ and covariance $\boldsymbol{Q}^{(t)}$
  end for
  Compute posterior modes as estimates
  if $\text{MSE}_{k^*,\text{random}} < \text{MSE}_{k^*,\text{fixed}}$ **then**
    Keep the random effect $\boldsymbol{\gamma}_{k^*}$
  end if
end for

3 Simulation & Results

Since no other approaches can perform both random effect selection and uncertainty estimation simultaneously, the effectiveness of our proposed method is showed in three simulations. The basic specification for all three simulations is a random effect model

$$\boldsymbol{y}_i = 1 + 2\boldsymbol{x}_{i1} + 4\boldsymbol{x}_{i2} + 3\boldsymbol{x}_{i3} + 5\boldsymbol{x}_{i4} + \boldsymbol{\gamma}_{i0} + \boldsymbol{\gamma}_{i1}\boldsymbol{x}_{i3} + \boldsymbol{\gamma}_{i2}\boldsymbol{x}_{i4} + \boldsymbol{\varepsilon}_i,$$

with $(\boldsymbol{\gamma}_{i0}, \boldsymbol{\gamma}_{i1}, \boldsymbol{\gamma}_{i2}) \sim N(\boldsymbol{0}, \boldsymbol{Q})$ where

$$\boldsymbol{Q} = \begin{pmatrix} \tau^2 & \tau^* & \tau^* \\ \tau^* & \tau^2 & \tau^* \\ \tau^* & \tau^* & \tau^2 \end{pmatrix}.$$

1. Estimation accuracy: By comparing to the gradient boosting method for LMMs (`grbLMM`), we see no obvious differences in estimation accuracy but large improvement in false-positives of fixed effects, see Table 1, while the improvement can partially be granted to the usage of probing.

2. Random effects selection: Since `grbLMM` cannot perform random selection, we evaluate our method separately to see the performance of estimation and find that there exists no obvious decrease in estimation accuracy, since very low false-positives of random effects are observed and there is no occurrence of false-negatives, see Figure 1.

TABLE 1. Mean value of 100 simulation runs with respect to each model evaluation metric between `grbLMM` and BayesBoost in the random slope setup.

τ	p	grbLMM					BayesBoost				
		$\text{MSE}_{\boldsymbol{\beta}}$	$\text{MSE}_{\boldsymbol{Q}}$	MSE_{σ^2}	$\text{MSE}_{\boldsymbol{\gamma}}$	FP	$\text{MSE}_{\boldsymbol{\beta}}$	$\text{MSE}_{\boldsymbol{Q}}$	MSE_{σ^2}	$\text{MSE}_{\boldsymbol{\gamma}}$	FP
0.4	10	0.020	0.013	0.002	4.482	0.46	0.026	0.010	<.001	3.529	0.10
	25	0.022	0.012	0.003	4.530	0.27	0.029	0.010	<.001	3.545	0.04
	50	0.023	0.012	0.003	4.551	0.18	0.030	0.011	<.001	3.588	0.02
	100	0.025	0.012	0.003	4.536	0.10	0.031	0.011	<.001	3.644	0.01
	500	0.027	0.011	0.003	4.453	0.03	0.040	0.010	<.001	3.660	<.01
0.8	10	0.072	0.124	0.002	6.923	0.44	0.083	0.133	<.001	6.332	0.11
	25	0.073	0.121	0.003	7.015	0.28	0.087	0.136	<.001	6.482	0.04
	50	0.074	0.119	0.003	6.956	0.17	0.087	0.134	<.001	6.620	0.02
	100	0.078	0.094	0.003	7.060	0.11	0.085	0.118	<.001	6.635	0.01
	500	0.082	0.124	0.003	6.953	0.04	0.100	0.139	<.001	6.832	<.01
1.6	10	0.280	1.829	0.002	16.970	0.41	0.321	2.053	<.001	15.669	0.09
	25	0.277	1.808	0.002	16.605	0.29	0.316	2.007	<.001	15.940	0.04
	50	0.294	1.435	0.002	17.124	0.19	0.302	1.796	<.001	15.950	0.02
	100	0.299	1.852	0.003	16.682	0.14	0.310	1.931	<.001	15.006	0.01
	500	0.320	1.804	0.003	17.658	0.04	0.361	1.773	<.001	17.280	<.01

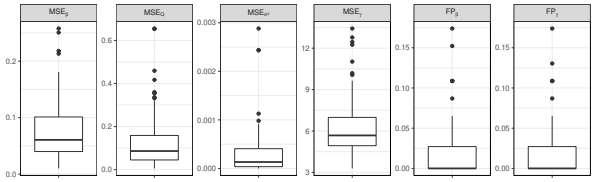


FIGURE 1. Boxplot of each model evaluation metric estimated by BayesBoost summarizing the outcomes of 100 simulation runs for the random slope setup with $\tau = 0.8$ and $p = 50$.

3. Performance of uncertainty estimation: Since the outcomes provided by boosting approaches contain no uncertainty information, we compare our method to BayesX. According to the results, no obvious differences in the uncertainty estimation for random effects are observed, see Figure 2. This indicates the good performance of the proposed algorithm in uncertainties.

We also apply our method to a real-world data [Schelldorfer et al. 2011], the riboflavin data, and find for the first time, that the gene *YXLD-at* has not only fixed effect, but also informative random effect.

4 Conclusion

In the proposed algorithm, uncertainty estimation for random effects is possible inside the boosting framework due to the usage of Bayesian inference and the selection ability of current boosting approaches is extended to

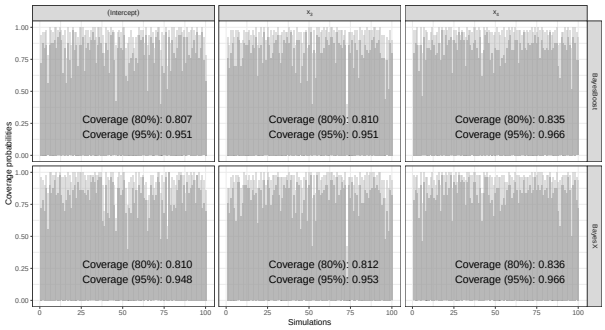


FIGURE 2. Coverage probabilities of the 80%- and 95%-intervals of each random effect in 100 simulation runs for both by BayesBoost and BayesX. For each run, the coverage probability summarizes the percentage of true effects covered by the corresponding interval. The dark and light grey bars in each simulation indicate the 80%- and 95%-interval respectively. The overall coverage rate among all 100 runs are labeled with the corresponding values.

random effects by comparing the contribution of selected variable as fixed and random effect to the model.
According to simulations and a real world data analysis, we find that the outcome produced by the proposed approach is very competitive to other existing methods.

References

Fong, Y., Rue, H., and Wakefield, J. (2010). *Bayesian inference for generalized linear mixed models*. Biostatistics, 11(3):397-412.

Griesbach, C., Säfken, B., and Waldmann, E. (2021) *Gradient boosting for linear mixed models*. The International Journal of Biostatistics, 17(2): 317-329.

Thomas, J., Hepp, T., Mayr, A., and Bischl, B. (2017) Probing for sparse and fast variable selection with model-based boosting. Computational and Mathematical Methods in Medicine.

Schelldorfer, J., Bühlmann, P., and de Geer, S. v. (2011) Estimation for high-dimensional linear mixed-effects models using L1-penalization. Scandinavian Journal of Statistics, 38(2):197-214.

Mean and median bias reduction in generalized linear models with large data sets

Patrick Zietkiewicz[1], Ioannis Kosmidis[1,2]

[1] Department of Statistics, University of Warwick, Coventry, UK
[2] The Alan Turing Institute, London, UK

E-mail for correspondence: patrick.zietkiewicz@warwick.ac.uk

Keywords: Iteratively reweighted least squares; Incremental QR decomposition; Mean and median bias reduction.

1 Introduction

Mean and median bias reduction (BR) (see Kosmidis et al., 2020), are becoming increasingly popular in the estimation of generalized linear models (GLMs). Recent work by Sur and Candès (2019) and Kosmidis and Firth (2021) also illustrates that mean BR may be particularly effective in high-dimensional problems with $p/n \to \kappa \in (0, 1)$ where n and p are the number of observations and parameters respectively. In this work, we develop and present new algorithms to estimate GLMs with mean and median BR for arbitrarily large data sets without encountering memory issues. The algorithms come from adjusting particular quantities in the iteratively reweighted least squares (IWLS) algorithm for BR (see, Kosmidis et al., 2020, Section 2) so that at each iteration the least squares optimisation problem can be solved with incremental QR decompositions. The key insight for mean BR is reusing quantities from the previous IWLS iteration, and for median BR, complementing that with a double least squares computation at each iteration. The method gives exact, not approximate, estimates and opens the door for using mean and median BR on large n and large p data sets.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Incremental algorithms for bias reduction

2.1 Introduction to GLMs

Suppose that $y = (y_i, \ldots, y_n)^{\mathrm{T}}$ are observations of independent random variables $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ each with probability density

$$f_{Y_i}(y; \theta_i, \phi) = \exp\left\{\frac{y\theta_i + b(\theta_i) - c_1(y)}{\phi/m_i} - \frac{1}{2}a\left(-\frac{m_i}{\phi}\right) + c_2(y)\right\}$$

for some sufficiently smooth functions $b(\cdot)$, $c_1(\cdot)$, $a(\cdot)$ and $c_2(\cdot)$, and fixed observation weights $m_1, \ldots, m_n$. The expected value and variance of Y_i are $\mathbb{E}[Y_i] = \mu_i = b'(\theta_i)$ and $\mathrm{Var}[Y_i] = \phi b''(\theta_i)/m_i = \phi V(\mu_i)/m_i$, respectively where $b'(\theta_i)$ and $b''(\theta_i)$ are the first two derivatives of $b(\theta_i)$ and $V(\mu_i)$ is the variance function. The dispersion parameter ϕ allows shrinking or inflating the contribution of the mean. A GLM links the mean μ_i to a linear predictor η_i through a monotone, sufficiently smooth link function $g(\mu_i) = \eta_i$ with $\eta_i = \sum_{t=1}^{p} \beta_t x_{it}$ where x_{it} is the (i,t)th component of a model matrix X, and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$. An intercept parameter is typically included in the linear predictor, in which case $x_{i1} = 1$ for all $i \in \{1, \ldots, n\}$. In order to estimate β we first compute the derivative of the log-likelihood, given by expression (1)

$$s_\beta = \frac{1}{\phi} X^{\mathrm{T}} W D^{-1}(y - \mu) \tag{1}$$

where $\mu = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, $W = \mathrm{diag}\{w_1, \ldots, w_n\}$, $D = \mathrm{diag}\{d_1, \ldots, d_n\}$ and $w_i = m_i d_i^2 / v_i$ is the ith working weight, with $d_i = d\mu_i/d\eta_i$ and $v_i = V(\mu_i)$. For brevity we will ignore the estimation of ϕ.

2.2 Incremental least squares with QR decomposition

Suppose we wish to regress $Y \in \mathbb{R}^n$ on $X \in \mathbb{R}^{n \times p}$ with known weights $W \in \mathbb{R}^{n \times n}$ using least squares and we have the QR decomposition

$$W^{1/2} X = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0_{(n-p)\times p} \end{bmatrix} \tag{2}$$

where $Q_1 \in \mathbb{R}^{n \times p}$, $Q_2 \in \mathbb{R}^{n \times (n-p)}$, R_1 is an upper triangular $p \times p$ matrix, and $0_{(n-p)\times p}$ is a $(n-p) \times p$ matrix of zeros. The least squares estimate $\hat{\beta} \in \mathbb{R}^p$ is given by back-solving $Q_1^{\mathrm{T}} W^{1/2} Y = R_1 \hat{\beta}$. Miller (1992) describes how this entire process can be done incrementally, i.e. requiring only one observation y_i, $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and $w_i = (w_{i1}, \ldots, w_{in})^{\mathrm{T}}$ at a time. This is owed to the fact that we can generate the QR decomposition (2) incrementally. The major benefit of an incremental approach is that we do not require to have the entire data set in memory at any one time. As a result, we can perform least squares on arbitrarily large data sets.

2.3 Mean and median bias reduction for GLMs

We call s_β, from expression (1), the score function with respect to the coefficients $\beta \in \mathbb{R}^p$ of a GLM. The maximum likelihood (ML) estimate is found by solving $s_\beta = 0_p$ with respect to β where 0_p is a $p-$vector of zeros. Firth (1993) proposed adjusting the score equation $s_\beta = 0_p$ to

$$s_\beta + X^{\mathrm{T}} W \xi = 0_p$$

which results in mean BR estimates where $\xi = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$ and $\xi_i = h_i d'_i/(2d_i w_i)$, $d'_i = d^2\mu_i/d\eta_i^2$ and h_i is the "hat" value for the ith observation, obtained as the ith diagonal element of the matrix $H = X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W$. In a similar vein, Kenne Pagui et al. (2017) propose a different adjustment

$$s_\beta + X^{\mathrm{T}} W (\xi + Xu) = 0_p$$

which gives median BR estimates where $u = (u_1, \ldots, u_p)^{\mathrm{T}}$ and

$$u_t = \left[(X^{\mathrm{T}}WX)^{-1}\right]_t^{\mathrm{T}} X^{\mathrm{T}} \begin{bmatrix} \tilde{h}_{t,1}\{d_1 v'_1/(6v_1) - d'_1/(2d_1)\} \\ \vdots \\ \tilde{h}_{t,n}\{d_n v'_n/(6v_n) - d'_n/(2d_n)\} \end{bmatrix} \tag{3}$$

where $[B]_t$ denotes the tth row of matrix B as a column vector, $v'_i = dV(\mu_i)/d\mu_i$ and $\tilde{h}_{t,i}$ is the ith diagonal element of $XK_tX^{\mathrm{T}}W$ where $K_t = [F^{-1}]_t[F^{-1}]_t^{\mathrm{T}}/[F^{-1}]_{tt}$ with $F = X^{\mathrm{T}}WX$ and $[B]_{tt}$ denotes the (t,t)th element of the matrix B. Solving $s_\beta = 0_p$ for GLMs can be done using the IWLS algorithm (Green, 1984), which solves a least squares problem at each iteration until the estimates converge. The least squares problem can be solved using the QR decomposition, which in turn can be done incrementally requiring only the data for a single (or small block of) observation at a time, as described in Section 2.2. For the case of GLMs this is useful since all scalar quantities involved in IWLS (e.g. linear predictors, working weights and working variates) depend only data from the corresponding observations. We will call a quantity with this property *local*, and without it *non-local*. This means we could, at each iteration of the IWLS algorithm, solve the least squares problem by: reading in one observation, performing a computation, and replace the observation with the next one. In this way ML estimation of GLMs can be performed with data sets of arbitrary size in n.

This is not directly possible for mean and median BR because the quantities h_i $(i = 1, \ldots, n)$ and u force the ith working observation in the least squares problems at each iteration of the IWLS algorithms to involve quantities which depend on all observations, i.e. non-local. This means we cannot solve the problem as for ML by dealing with one observation at a time, and thus the naive IWLS implementations of mean and median BR in Kosmidis et al. (2020) cannot handle data sets with arbitrarily large n.

2.4 IWLS for mean and median bias reduction

Expression (4) gives the update step for β at the $(j+1)$th iteration in the three implementations of the IWLS algorithm in this work, namely ML, mean and median BR. The braces in (4) indicate the relevant working variates. For example $z^{(j)} = (z_1^{(j)}, \ldots, z_n^{(j)})^{\mathrm{T}}$ with $z_i^{(j)} = \eta_i^{(j)} + (y_i - \mu_i^{(j)})/d_i^{(j)}$ is the working variate for ML, and $z^{(j)} + \phi^{(j)}\xi^{(j)}$ the working variate for mean BR. The superscript (j) indicates a quantity that is evaluated at the jth iteration.

$$\beta^{(j+1)} \leftarrow (X^{\mathrm{T}}W^{(j)}X)^{-1}X^{\mathrm{T}}W^{(j)}(\overbrace{\underbrace{z^{(j)}}_{\text{ML}} + \phi^{(j)}\xi^{(j)}}^{\text{mean BR}} + \phi^{(j)}Xu^{(j)}) \qquad (4)$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{median BR}}$$

2.5 Incremental algorithms

We will deal with the quantities that prevent incremental least squares for mean and median BR, the "hat" values and u vector, separately. First, the "hat" values. We note that in order to solve the least squares problem at iteration $(j-1)$ we would have available the QR decomposition $W^{1/2(j-1)}X = Q_1^{(j-1)}R_1^{(j-1)}$. The first idea is to substitute $H^{(j)}$ with $H^{(j-1)} = Q^{(j-1)}Q^{\mathrm{T}(j-1)}$ at iteration j, so we now have $\xi_i^{*(j)} = h_i^{(j-1)}d_i^{'(j)}/(2d_i^{(j)}w_i^{(j)})$ $(i = 1, \ldots, n)$. Since $j/(j-1) \to 1$ as $j \to \infty$, the algorithm has the same stationary point as the IWLS algorithms in Kosmidis et al. (2020). Now, for median BR, instead of using $u_t^{(j)}$ as defined in expression (3) we reframe expression (3) as the following least squares problem

$$U^{*(j)} = \left(X^{\mathrm{T}}W^{(j-1)}X\right)^{-1}X^{\mathrm{T}}W^{(j)}\begin{bmatrix}\tilde{h}_{t,1}^{(j-1)}\{d_1^{(j)}v_1^{'(j)}/(6v_1^{(j)}) - d_1^{'(j)}/(2d_1^{(j)})\}/w_1^{(j)}\\ \vdots \\ \tilde{h}_{t,n}^{(j-1)}\{d_n^{(j)}v_n^{'(j)}/(6v_n^{(j)}) - d_n^{'(j)}/(2d_n^{(j)})\}/w_n^{(j)}\end{bmatrix}$$

for $(t = 1, \ldots, p)$ and extract $u_t^{*(j)} = \left[U^{*(j)}\right]_t$ where we use $\tilde{h}_{t,i}^{(j-1)}$ and the QR decomposition $W^{1/2(j-1)}X = Q_1^{(j-1)}R_1^{(j-1)}$ from the previous iteration in order to make all quantities involved local. With these two adjustments we modify the update (4) to give expression (5)

$$\beta^{(j+1)} \leftarrow (X^{\mathrm{T}}W^{(j)}X)^{-1}X^{\mathrm{T}}W^{(j)}(\overbrace{\underbrace{z^{(j)}}_{\text{ML}} + \phi^{(j)}\xi^{*(j)}}^{\text{mean BR}} + \phi^{(j)}Xu^{*(j)}) \qquad (5)$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{median BR}}$$

Since all quantities in expression (5) are local it is now possible to estimate GLMs using mean and median BR using data sets with an arbitrarily large number of observations, without encountering memory problems.

3 Demonstration

In Figure 1 we demonstrate these algorithms on the high-dimensional simulation set up used in Figure 2 of Sur and Candès (2019). We fit a logistic regression model using ML, mean and median BR, data has $n = 4000$ observations, $p = 800$ predictor variables and $X_{it} \sim N(0, 1/n)$. We see mean and median BR achieve numerically similar estimates centred around the true values, whilst the ML estimates exhibit a larger variance and bias in absolute value as estimates fan away from the non-zero coefficients.

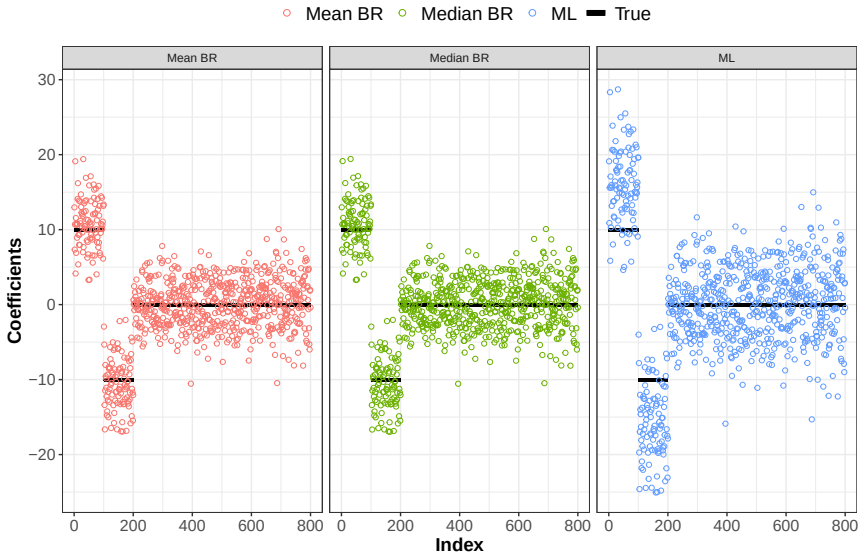


FIGURE 1. Simulation of mean and median bias reduction for logistic regression in the left and right panels respectively. $n = 4000$ observations, $p = 800$ predictor variables where $\beta_1 = \cdots = \beta_{100} = 10$, $\beta_{101} = \cdots = \beta_{200} = -10$ and $\beta_{201} = \cdots = \beta_{800} = 0$ and $X_{it} \sim N(0, 1/n)$.

References

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80** (1) 27 – 38.

Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society Statistical Methodology Series B*, **46** (2) 149 – 192.

Kenne Pagui, E.C., Salvan, A., Satori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika*, **104** (4) 923 – 938.

Kosmidis, I., Kenne Pagui E. C., and Sartori, N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing*, **30** (1) 43 – 59.

Kosmidis, I., and Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, **108** (1) 71 – 82.

Miller, A. J. (1992). Algorithm AS 274: Least squares routines to supplement those of Gentleman. *Applied Statistics*, **41** (2) 458 – 478.

Sur, P., and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, **116** (29) 14516 – 14525.

Part III

Comparing recovery sample designs to test for the presence of MNAR

Adetola Adedamola Adediran[1], Jack Noonan[2], Robin Mitra[2], Stefanie Biedermann[3]

[1] School of Mathematical Sciences, University of Southampton, United Kingdom
[2] School of Mathematics, Cardiff University, United Kingdom
[3] School of Mathematics and Statistics, The Open University, United Kingdom

E-mail for correspondence: a.a.adediran@soton.ac.uk

Abstract: Missing data is known to be an inherent and pervasive problem in the process of data collection. The specific problem of data Missing Not At Random (MNAR) is known to be one of the most complex and challenging problems to handle in this field. One major issue is the fact MNAR missingness is an untestable assumption. Unless one has the ability to recover some of the missing values, for example, through a follow up survey, then MNAR cannot be identified. In this research, using a test for MNAR, we compare how effectively four follow up designs detect the presence of MNAR.

Keywords: Missing data; Missing not at Random; Selection Model

1 Introduction

In order to correctly make inferences in the presence of missing data, Rubin (1976) classified missing data problems into Missing Data Mechanisms (MDM). These mechanisms are Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). One can show that in terms of inference, the first two types are essentially ignorable. However, MNAR cannot be ignored and is an untestable assumption based on the original incomplete data (Little and Rubin (2002)). If not correctly accounted for, MNAR can lead to significant bias in analysis and potentially incorrect conclusions. If it is possible to follow-up or recover some of the missing values, then statistical tests can be constructed (Carpenter and Kenward (2007)) to detect the presence of MNAR.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

Since (typically) only a small percentage of missing observations can be recovered due to cost and time constraints, there is a demand to recover missing values in a manner that complements such MNAR tests. The benefits of considering MNAR in the design stage of an experiment were demonstrated in (Lee, Mitra and Biedermann (2018)). However, the key difference between this research and previous work is our inability to design the experiment beforehand. Instead, we only have control to carefully construct, or 'design', the follow up sample.
In what follows, we consider tests for MNAR along the lines proposed in Carpenter and Kenward (2012) and investigate how different constructions of the follow up sample (we will simultaneously use the word design) will affect the Type I error and the power of the test. In Section 2, we introduce the methodology. The main numerical results are presented in Section 3 and concluding remarks are provided in Section 4.

2 Method and Analysis

Consider the simple regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon\,, \tag{1}$$

where $\varepsilon \sim N(0, \sigma^2)$ and β_0, β_1 and σ^2 are potentially unknown parameters. Suppose there is the possibility for missing values to be present in the model, where for simplicity we assume missing values are constrained to Y and the covariate X is always observed. Let M denote an indicator random variable that equals one if Y is missing and zero if Y is present. The MDM's of Rubin (1976) are determined by the conditional density $f(M|X, Y, \theta)$, where θ are some fixed parameters of the distribution. For MCAR, missingness does not depend on X and Y, i.e.

$$f(M|X, Y, \theta) = f(M|\,\theta)\,.$$

For MAR, the probability of missingness is only dependent on observed values, i.e.

$$f(M|X, Y, \theta) = f(M|\,X, \theta)\,.$$

For MNAR, the dependence on Y cannot be ignored. Under MCAR and MAR, the MDM can often be ignored and unbiased inferences can be made without needing to incorporate the MDM into the inferences (Rubin (1976); Heitjan and Basu (1996)). However, a MNAR mechanism cannot be ignored and must be incorporated into the analysis to make appropriate inferences. Combining this with the fact MNAR is an untestable assumption based on the original data, analysing MNAR data is substantially more complicated than other types of missing data. This is likely why the MAR assumption (or the even stronger assumption of MCAR) is the most common assumption when faced with missing data in practical applications. ,

Consider a realisation of size n from the model in (1), resulting in the vectors $\mathbf{Y} := (Y_1, \ldots, Y_n)$ and $\mathbf{X} := (X_1, \ldots, X_n)$. Suppose n_{miss} values in $\mathbf{Y}$ are missing but we know their corresponding covariate values in $\mathbf{X}$. Now assume one has the ability to recover n^* cases from the n_{miss} missing values, with $n^* \leq n_{miss}$. Let $\mathbf{Y}^*$ denote the augmented response data; that is the Y_i values that were originally observed and the n^* of the recovered Y_i. A test for MNAR (from Carpenter and Kenward (2012) page 15) formulated in what is referred to as the Selection Model framework (SMF) is as follows. For $Y_i^* \in \mathbf{Y}^*$, fit the model

$$logitPr(M_i = 1) = \alpha_0 + \alpha_1 Y_i^* + \alpha_2 X_i \,. \tag{2}$$

Under the null hypothesis of MAR, we have $\alpha_1 = 0$, otherwise we conclude MNAR.

In this research, we explore how different follow up designs (or ways of choosing what missing response values to recover based only on their corresponding observed covariate values) affect the power of the SMF test for detecting MNAR. In this paper, we will consider the following four follow up designs: 1) *Random*, which involves a random selection across the covariate space of the missing cases; 2) *Highest*, where the n^* highest values of the covariates corresponding to the missing values are selected; 3) *Smallest*, where n^* values with the smallest covariates are selected; 4) *Half Highest Half Smallest Values* which selects the n^* values such that (approximately) half of the recovered responses have the highest covariate values and half have the smallest covariate values.

3 Simulation study

In the first example of this numerical study section, we generate points according to the simple linear regression model:

$$Y_i|(X_i = x_i) \quad \sim \quad N(1 + 2x_i, 4)\,,$$

with $X_i \sim N(5, 1)$, for $i = 1, \ldots, 1000$. When introducing missingness into the model, under a MAR mechanism we use

$$P(M_i = 1) = 1/(1 + \exp(3 - 0.42 \cdot x_i))\,. \tag{3}$$

Under a MNAR mechanism, we will use

$$P(M_i = 1) = 1/(1 + \exp(3 - 0.19 \cdot y_i))\,. \tag{4}$$

The choice of these parameters results in (on average) 300 missing values being simulated in Y, or approximately 30% missingness in the response. We then apply the SMF test in (2), performing a suitable hypothesis test on α_1, and subsequently replicate the process $1,000$ times to obtain empirical

summaries for Type 1 error (under MAR) and power (under MNAR) using the different designs.

In Figure 1 (Left), we plot empirical Type 1 error against the recovered sample size for the four designs considered; here we use the missing mechanism provided in (3). In this figure, we see all designs approximately produce a Type 1 error close to the pre-specified value of 0.05. The *Highest* design appears to struggle slightly controlling Type 1 error; its error is above 0.05 when fewer than 67% of the missing values are recovered. The *Smallest* design gives Type 1 errors above 0.05 when less than 50% of the missing values are recovered. Selecting *Half Highest Half Smallest* appears to produce Type 1 errors values very close but slightly below 0.05. The *Random* design produces Type 1 errors in line with expectation, where the volatility around 0.05 appears to be from simulation variance.

Type 1 error appears only approximately close to 0.05 for most designs due to the following reason. By fitting the logit model in (2), one immediately assumes the missing mechanism (MAR in this case) is of the *expit* form $1/(1+exp(\beta_0-\beta_1 x))$. Whilst this is true in the example considered in (3), there is no guarantee that when fitting the logit model to the augmented data with $n^* < n_{miss}$ that the true model is also of the required *expit* form. This appears to only be true for a *Random* design. Nevertheless, as Figure 1 (Left) demonstrates (and later Figure 2 (Left)), this issue does not seem to drastically impact results and meaningful analysis with the test in (2) can still be obtained if one accepts this slight caveat.

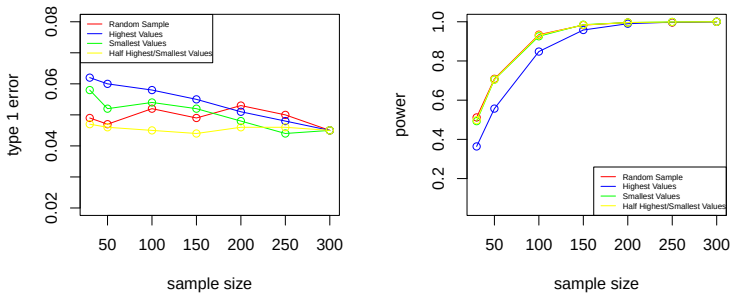


FIGURE 1. Left: Type 1 error using SMF. Right: power using SMF

In Figure 1 (Right), we plot power against recovered sample size using the MNAR model in (4) for the four designs considered. At $n^* = 30$, the power for all designs is less than or equal to 0.512. The *Highest* design has the lowest power. The *Random* design seems to have the highest power from n^* from $30-100$. For $n^* = 150$ and above, the *Smallest* design and *Half highest/smallest* design seem to have the highest power.

In the second example of this numerical study section, we generate points

according to the model:

$$Y_i|(X_i = x_i) \quad \sim \quad N(1 + x_i, 4)\,,$$

with $X_i \sim N(1, 4)$, for $i = 1, \ldots, 1000$. When introducing missingness into the model, under a MAR mechanism we use

$$P(M_i = 1) = 1/(1 + \exp(0.4 + x_i))\,. \tag{5}$$

Under a MNAR mechanism, we will use

$$P(M_i = 1) = 1/(1 + \exp(0.5 + x_i - 0.1 \cdot y_i))\,. \tag{6}$$

The choice of these parameters results in approximately 30% missingness in the response variable. In Figure 2 (Left) we plot the empirical Type 1 errors against the recovered sample size. The style of this figure is the same as Figure 1 (Left) and a similar phenomenon is seen for the MAR mechanism in (5); that is, Type 1 errors seem approximately close to the pre-specified error of 0.05. The *Smallest* design seems to suffer more than the other designs in terms of Type 1 error. In Figure 2 (Right), we depict power against the recovered sample size using the MNAR mechanism of (6). Figure 2 is extremely insightful and provides a proof of concept example that was not so obviously seen in Figure 1 (Right). It demonstrates that for this particular MNAR mechanism, the *Random* design can be significantly improved on by the *Highest* design across all recovery sample sizes.

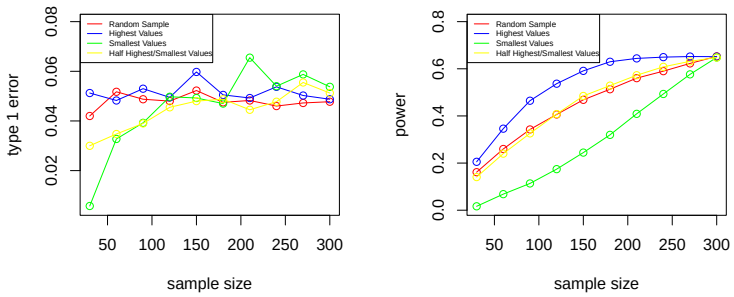


FIGURE 2. Left: Type 1 error using SMF. Right: power using SMF

Another important observation from Figure 2 (Right) is that for the *Highest* design, there is little benefit in recovering more than 180 of the missing responses (or 60% of the missing vales) should one encounter this MNAR mechanism. If a particular cost is associated with each recovery, a significant saving could be made with a minimal loss in power.

4 Conclusion

This research has demonstrated a proof of concept example by showing that different follow up designs have different capabilities of identifying the presence of MNAR. Most importantly, we have seen an example where the *Random* design can be beaten in terms of power. This provides motivation to theoretically study the SMF test. A careful design of the follow up sample has the potential to provide significant benefits to the identification of MNAR over a random follow up.

References

Carpenter, J.R. and Kenward, M.G. (2007). *Missing data in randomised controlled trials: a practical guide.* Health Technology Assessment Methodology Programme, Birmingham.

Carpenter, J.R. and Kenward, M.G. (2012). *Multiple imputation and its application.* John Wiley and Sons.

Heitjan, D.F. and Basu, S. (1996). Distinguishing 'missing at random' and 'missing completely at random'. *The American Statistician* , 50(3):207-213.

Little, R.J. and Rubin, D.B. (2002). *Statistical analysis with missing data.* John Wiley and Sons.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3):581-592.

Lee, K.M and Mitra, R. and Biedermann, S. (2018) Optimal design when outcome values are Not Missing at Random *Statistica Sinica* 27(4): 1821-1838.

Reliable generalized linear latent variable models estimation via simulated maximum likelihood

Giuseppe Alfonzetti[1], Ruggero Bellio[2]

[1] University of Padua, Italy
[2] University of Udine, Italy

E-mail for correspondence: giuseppe.alfonzetti@phd.unipd.it

Abstract: The estimation of generalized linear latent variable models has been tackled via several approaches, but it remains an open problem. In this contribution, we propose an estimator based on randomized quasi-Monte Carlo simulated maximum likelihood, which allows for an accurate approximation of maximum likelihood while remaining scalable to moderate settings. A simulation study is carried out for comparison purposes with the estimation via Expectation-Maximization, which is considered among the most reliable estimation approaches.

Keywords: latent variables; simulated maximum likelihood; importance sampling

1 Introduction

Generalized linear latent variable models (GLLVM) are a widely used framework to incorporate latent variables within regression problems, allowing for responses of different types in the same fashion as multivariate generalized linear models. There is a clear connection with the mixed model literature, since latent variables can be seen as a special case of random effects related to each observed unit.
Maximum likelihood estimation of GLLVM is well known to be troublesome since it involves integrals with no closed-form solution to be evaluated at each unit in the sample. As reviewed in Bartholomew et al. (2011), the standard approach is to rely on algorithms based on Expectation-Maximization (EM), with the E-step approximated through a Gaussian quadrature procedure. This typically does not allow to accurately scale model estimation

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

on latent spaces with dimensions greater than two, or slightly higher in the case of adaptive quadrature; see for example Schilling and Bock (2005). Because of that, several methods have been proposed in the literature to extend GLLVM estimation on larger datasets, but no generally reliable estimation method is available for practitioners yet.
Adapting applications in the generalized linear mixed models literature (e.g. Skaug, 2002) we propose a simulated maximum likelihood approach (SML), implemented as an explicitly parameter dependent importance sampling procedure, as outlined in Brinch (2012). The estimation is therefore based on a set of random points common to all the iterations. A further improvement considered here is to switch to randomized quasi-Monte Carlo integration, as in Jank (2006); see also Lemieux (2009) for an overview.

2 Simulated maximum likelihood

Let n be the sample size, p the number of observed variables and q the dimension of the latent space. The observed dataset y is the realization of the $n \times p$ random matrix Y, such that $Y = (Y_1^{\mathrm{T}}, \ldots, Y_n^{\mathrm{T}})$, with $Y_i = (Y_{i1}, \ldots, Y_{ip})$ and $i = 1, \ldots, n$. At the same time, u is defined as the realization of the $n \times q$ random matrix U such that $U = (U_1^{\mathrm{T}}, \ldots, U_n^{\mathrm{T}})$ and $U_i = (U_{i1}, \ldots, U_{iq})$, with $U_i \overset{iid}{\sim} \mathcal{N}_q(0, \Sigma)$ and Σ is a correlation matrix. With $\phi_q(u_i; 0, \Sigma)$ we refer to the density associated with $\mathcal{N}_q(0, \Sigma)$ evaluated at u_i. Let Λ be a $p \times q$ matrix which plays a similar role of a loading matrix in normal factor models, such that $\Lambda = (\lambda_1^{\mathrm{T}}, \ldots, \lambda_p^{\mathrm{T}})$ and $\lambda_j = (\lambda_{j1}, \ldots, \lambda_{jq})$, and $\alpha = (\alpha_1, \ldots, \alpha_p)$ a vector of intercepts, for $j = 1, \ldots, p$. Then, the interest lies in the estimation of the free parameters in α, Λ and Σ, which are collected in the d-dimensional vector Ψ.
The model assumes the observed variables to be locally independent given the latent space, such that the joint likelihood of the data and the latent variables for the i-th subject is

$$\mathcal{L}(\Psi; y_i, u_i) = \phi_q(u_i; 0, \Sigma) \prod_j^p p(y_{ij} | u_i; \eta_{ij}), \tag{1}$$

where $p(y_{ij}|u_i; \eta_{ij})$ is assumed to be a member of the exponential family, and its canonical parameter depends on the linear predictor $\eta_{ij} = \alpha_j + \lambda_j^{\mathrm{T}} u_i$ through an appropriate link function. It follows that, since the values of u_i are not known, the likelihood function requires to compute the integral

$$\mathcal{L}(\Psi; y_i) = \int_{\mathbb{R}^q} \phi_q(u_i; 0, \Sigma) \prod_j^p p(y_{ij} | u_i; \eta_{ij}) du_i. \tag{2}$$

Aside from the special case where $p(y_{ij}|u_i; \eta_{ij})$ is normal, (2) has no closed-form solution and must be approximated.

Our proposal simulates the integral in (2) through an importance sampling procedure, which suitably extends the first-order Laplace approximation to the integral (2). More precisely, the importance distribution is chosen as the normal density implied by a second-order Taylor's expansion of (1) around its maximum with respect to u_i. The generation of new samples from the importance distribution relies on a quasi-Monte Carlo procedure, based on randomized Halton sequences (see Lemieux 2009, Ch. 5). Namely, the approximation of (2) is obtained using R independent sequences of length M, such that the simulated likelihood is computed via

$$\mathcal{L}_{IS}(\Psi; y) = \prod_i^n \frac{1}{R} \sum_r^R \frac{1}{M} \sum_m^M \frac{\mathcal{L}(\Psi; Y_i = y_i, U_i = v_i^{(m)(r)})}{\phi_q(v_i^{(m)(r)}; \hat{u}_i(\Psi), H_i(\Psi)^{-1})}. \tag{3}$$

Here, $\hat{u}_i(\Psi)$ is the maximizer of $\ell(\Psi; y_i, u_i)$ given the data and the parameter vector, $H_i(\Psi)$ is the related matrix of negative second derivatives evaluated at $\hat{u}_i(\Psi)$ and $v_i^{(m)(r)}$ is drawn from $\mathcal{N}_q(\hat{u}_i(\Psi), H_i(\Psi)^{-1})$ for $m = 1, \ldots, M$, $r = 1, \ldots, R$ and $i = 1, \ldots, n$. Note that, following the explicitly parameter dependent construction outlined in Brinch (2012), the importance sample $v_i^{(m)(r)}$ needs to be expressed as

$$v_i^{(m)(r)} = \hat{u}_i(\Psi) + C_i(\Psi)^{\intercal} \Phi_q^{-1}(h^{(m)(r)}), \tag{4}$$

where $C_i(\Psi)$ is the lower Cholesky decomposition of $H_i(\Psi)^{-1}$, $\Phi_q(\cdot)$ the cumulative distribution function of a q-dimensional standard normal, and $h^{(m)(r)} \in [0, 1]^q$ is the m-th element of the r-th halton sequence generated following the algorithm in Owen (2017). Note that, as expected, when $R = 1$, the simulation of (2) corresponds to a deterministic quasi-Monte Carlo procedure. On the other hand, despite the linear increase in R of the computational cost, using $R > 1$ provides the possibility to assess the variance of the importance sampling estimate of the integral in (2) via its sample estimate over the R independent sequences.

3 A simulation study

The results of a small-scale simulation study are reported to highlight the effectiveness of the proposed method even in simple settings, where the EM-based estimation procedure relying on Gaussian quadrature integration of latent variables is usually considered the standard option. Binary data are generated from a true model with logit link, $p = 8$, $q = 2$ and $n \in \{100, 250, 500, 1000\}$. The loading matrix has a simple structure, with a single loading per row and four loadings in each column. For the sake of simplicity, the latent variables are assumed to be uncorrelated. For EM-based estimation, the implementation in `mirt` (Chalmers, 2012) is used, with 140 Gaussian quadrature points per dimension. On the other hand,

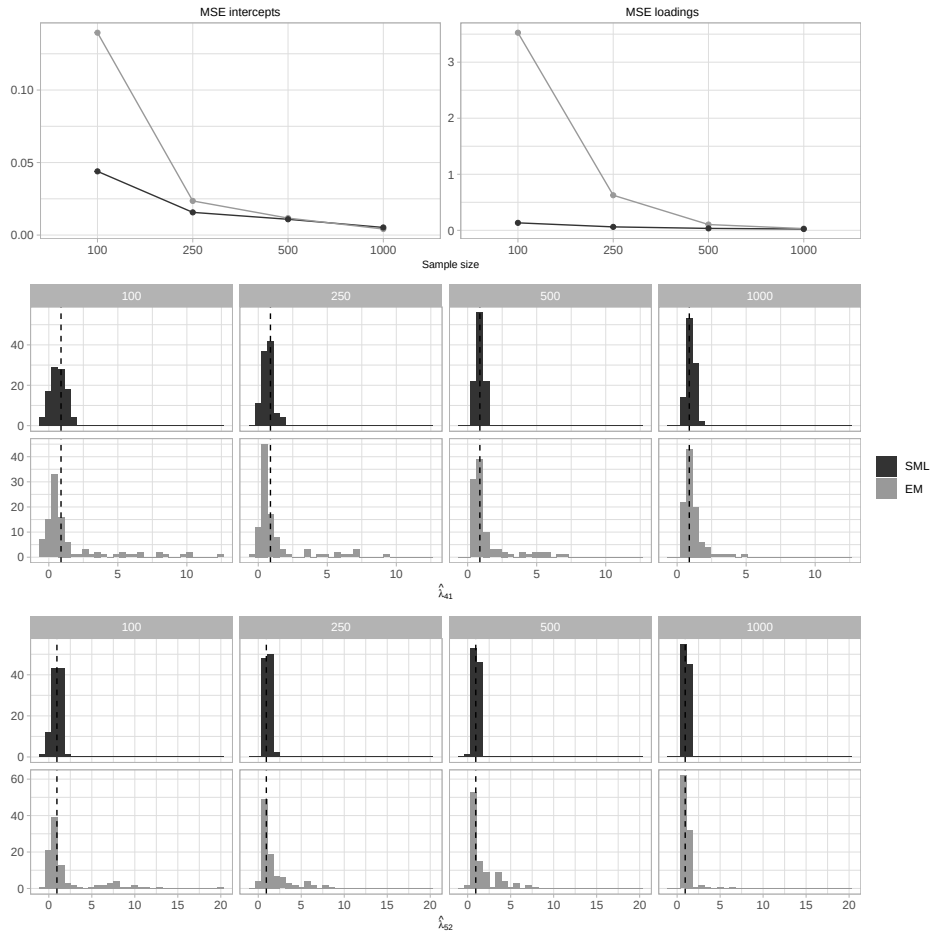


FIGURE 1. Top: Median Monte Carlo average MSE for intercepts and loadings. Center and bottom: respectively, Monte Carlo distribution of $\hat{\lambda}_{41}$ and $\hat{\lambda}_{52}$, for $n \in \{100, 250, 500, 1000\}$. Dashed lines represent the true value of the parameter.

SML estimation is performed through custom code written in `Rcpp` (Eddelbuettel and François, 2011), and the importance sampling dimensions are set to $M = 250$ with $R = 5$.
Figure 1 shows results for a Monte Carlo study with 100 replications. The plot at the top highlights the difference between the two methods in terms of mean square error (MSE). In particular, it tracks the median of the MSE for estimated intercepts and loadings across different sample sizes. While the two methods converge to the same estimates on larger sample sizes, SML clearly outperforms EM-based estimation on small-to-moderate samples. The lower panels show a more detailed representation of the distribution of individual estimators. The estimates $\hat{\lambda}_{41}$ and $\hat{\lambda}_{52}$ are presented as an example, but all the remaining ones behave similarly. The histograms point out that SML estimates are more concentrated around the true parameter value, while EM-based estimation exhibits heavier tail behavior.

4 Discussion and ongoing work

The estimation of GLLVM has been proven to be generally challenging, and at the time of writing one can safely say there is no available software that estimates such models in a reliable and free of convergence issues fashion. Here we propose a simulated maximum likelihood approach that is competitive with a quadrature-based EM approach even in a simple setting. Since SML estimation does not rely on a quadrature procedure, it does not suffer from the exponential complexity in the dimension of the latent space that affects the standard EM estimation approach. The accuracy of the method can be improved to an arbitrary level by choosing a higher simulation sample size, without the curse of dimensionality of quadrature-based methods. Therefore, the estimation can be potentially carried out also in more complex settings, providing practitioners with a new reliable tool to carry out approximate maximum likelihood estimation. An essential step for this aim is the availability of statistical software and, to this end, an R package implementing SML estimation for some notable GLLVM is currently under development.
Other investigations currently in progress concern the comparison of the results with further recent proposals, such as dimension-wise quadrature (Bianconcini et al., 2017), variational approximations (Hui et al., 2017), and stochastic EM algorithms (Zhang et al., 2020).

References

Bartholomew, D., Knott, M. and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. New York: John Wiley & Sons.

Bianconcini, S., Cagnone, S. and Rizopoulos, D. (2017). Approximate likelihood inference in generalized linear latent variable models based on the dimension-wise quadrature. *Electronic Journal of Statistics*, **11(2)**, 4404–4423.

Brinch, C. (2012). Efficient simulated maximum likelihood estimation through explicitly parameter dependent importance sampling. *Computational Statistics*, **27(1)**, 13 – 28.

Chalmers, R.P. (2012). mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software*, **48(6)**, 1 – 29.

Eddelbuettel, D. and François, R. (2011). Rcpp: seamless R and C++ integration. *Journal of Statistical Software*, **40(8)**, 1 – 18.

Hui, F. C. K., Warton, D. I, Ormerod, J. T., Haapaniemi, V. and Taskinen, S. (2017). Variational approximations for Generalized Linear Latent Variable Models, *Journal of Computational and Graphical Statistics*, **26(1)**, 35–43.

Jank, W. (2006). Efficient simulated maximum likelihood with an application to online retailing. *Statistics and Computing*, **16(2)**, 111 – 124.

Lemieux, C. (2009) *Monte Carlo and Quasi-Monte Carlo Sampling*. New York: Springer.

Owen, Art B. (2017) A randomized Halton algorithm in R. *arXiv: 1706.02808*.

Schilling, S. and Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, **70(3)**, 533–555.

Skaug, H. (2002). Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. *Journal of Computational and Graphical Statistics*, **11(2)**, 458 – 470.

Zhang, S., Chen, Y. and Liu, Y. (2020). An improved stochastic EM algorithm for large-scale full-information item factor analysis. *British Journal of Mathematical and Statistical Psychology*, **73, 44–71.**

Bayesian mixtures of discretely observed multi-state models

Rosario Barone[1], Andrea Tancredi [2]

[1] University of Rome Tor Vergata, Italy
[2] University of Rome La Sapienza, Italy

E-mail for correspondence: rosario.barone@uniroma2.it

Abstract: In this paper we propose a clustering technique for discretely observed continuous-time models in order to take account of groups of individuals having similar process realizations. In fact, fitting standard parametric models in presence of heterogeneity between population groups may produce biased inferences for relevant process features. To model individual heterogeneity we consider both finite mixtures and Dirichlet process mixture (DPM) of different multi-state models. We base our algorithms on the whole reconstructed trajectories with the reconstruction step conducted by the uniformization technique usually employed for the generation of Markovian multi-state processes. We present MCMC inference for Markov, semi-Markov and in-homogeneous Markov models with an application to a real dataset.

Keywords: Dirichlet process mixtures; Multi-state Markov models; Uniformization.

1 Introduction

In a panel data set with individuals observed in time, clustering techniques may be useful for finding groups of similar individuals. With exception for the Markov mixtures proposed by Luo et al. (2021), all the approaches for handling heterogeneity in multi-state models are related to finite mixtures and consider completely observed processes, see for example Fruhwirth-Schnatter and Pamminger (2010). In this paper we propose a general framework for tackling the clustering problem for different classes of of discretely observed multi-state processes. In general, note that inference for generalizations of Markov models may present computational difficulties when observations are at discrete time points, so that the process is not completely

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

observed. In fact, the likelihood function is not available and approximation methods are required. Barone and Tancredi (2022) reconstruct the likelihood function by simulating the trajectories between the observed points with a Metropolis-Hastings step based on Markovian proposals drawn from the uniformization algorithm of Hobolth and Stone (2009). Here, we extend their approach in the context of mixture models. For the sake of brevity we present the case of mixtures of semi-Markov models but our approach can be efficiently applied to the case of in-homogeneous Markov models and also to the simpler case of Markov models.

2 Semi-Markov multi-state models

Let us consider a continuous time process $Y(\cdot) = \{Y(t), t \geq 0\}$ with discrete state space $\mathcal{S} = \{1, \dots S\}$. We assume that the process $Y(\cdot)$ is semi-Markov. This is equivalent to say that the instantaneous transition rates $q_{rs}(t, \mathcal{F}_t)$, conditionally on the past history of the process, depend only on the time spent in the current state, i.e.

$$q_{rs}(t, \mathcal{F}_t) = \lim_{\delta t \to 0} \frac{P\{Y(t+\delta t) = s | X(t) = r, T^* = t - u\}}{\delta t}$$

where T^* denotes the entry time in the last state assumed before time t. Hence, semi-Markov models can be obtained by defining the transition functions $q_{rs}(u)$ and setting

$$P\{Y(t+\delta t) = s | Y(t) = r, T^* = t-u\} = \begin{cases} q_{rs}(u)\delta t + o(\delta t) & s \neq r \\ 1 - \sum_{l \neq r} q_{rl}(u)\delta t + o(\delta t) & s = r \end{cases}$$

Notice that a semi-Markov process $Y(t)$ can be also defined as the result of a state sequence generated by a Markov chain with transition probabilities p_{rs} and sojourn times having distribution functions F_{rs}, that is depending only on the departure and arrival states. The density of trajectory y on the interval $[0, T]$ can be generally written as

$$p_\theta(y) = p_\theta(s, z) = \left(\prod_{i=1}^{n} p_{s_{i-1} s_i} q(z_i - z_{i-1}; \phi_{s_{i-1}}) e^{-\int_0^{z_i - z_{i-1}} q(u; \phi_{s_{i-1}},) du} \right) \times e^{-\int_0^{T - z_n} q(u; \phi_{s_n}) du},$$

where $z = (z_1, \dots, z_n)$ is the sequence of jump times, $s = (s_1, \dots, s_n)$ is the sequence of visited states and $\theta \in \Theta$ is the vector with all the process parameters.

3 Dirichlet Process Mixture of semi-Markov models

In this section we introduce the notation of the DPM model. Let $p_\theta(y)$ be the probability density function of a semi-Markov trajectory $y(t) = (s, z)$.

Let G be a probability distribution defined on the parameter space Θ. We define the density function of an infinite mixture of semi-Markov models p_G with respect to the mixing measure G as

$$p_G(s,z) = \int p_\theta(s,z) dG(\theta).$$

By assuming a $DP(M, G_0)$ on the mixing measure G, we get a DPM of semi-Markov models. Let $y_i(t) = (s,z)_i$, for $i = 1, \ldots, N$, be N fully observed paths on $[0, T_i]$. Note that N represents the number of sample individuals. We may rewrite the model in a hierarchical form:

$$\begin{aligned} y_i(t)|\theta_i &\overset{ind}{\sim} p_{\theta_i} \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim \mathrm{DP}(MG_0). \end{aligned}$$

To extend the fitting of the DPM of semi-Markov models to the the case of discretely observed trajectories, that is when the exact jump times are unknown and the density function $p_\theta(y)$ is not available we use the algorithm proposed by Barone and Tancredi (2022) to make MCMC inference for parametric semi-Markov models. In fact, this algorithm reconstructs the trajectories between the discretely observed points for each observed individual via a Metropolis-Hastings step based on a Markovian approximation of the semi-Markov process. Hence by inserting this step in the algorithm for DPM of fully observed continuous time semi-Markov process we can naturally handle also the discretely observed case.

4 Application

As a real data application, we analyze the progression of coronary allograft vasculopathy (CAV) with a data set available with the R package `msm`, see Jackson (2011). The data provides the disease status (CAV-free (1), mild CAV (2) and moderate or severe CAV (3)) observed approximately each year after transplant for a set of 622 subjects followed up until their most recent visit if alive at the end of the observation period or until death (state (4)). Death times are exactly observed. To specify the semi-Markov model we assume Weibull sojourn times. The parameter set is $\theta = (p, \gamma, \alpha)$ where γ and α are the vector with the rate and shape parameters of the Weibull sojourn times and p is the matrix with the transition probabilities. For the Dirichlet process we chose a precision parameter $M = 1$ and defined the centering measure to be the product between Dirichlet distributions for the rows of p, Gamma distributions for the rate parameters and log Normal distributions for the sahpe parameters.In Table 1 and Figure 1 we show the results by reporting some posterior summaries for the model parameters. Note that we indicate with Ψ the cluster indexes.

TABLE 1. CAV data: DPM of semi-Markov.

	ψ	γ_1	γ_2	γ_3	α_1	α_2	α_3
$E(\cdot\|Y,\psi)$	1	0.13	0.24	0.25	1,47	1.46	1.10
$SD(\cdot\|Y,\psi)$	1	0.01	0.02	0.03	0.10	0.18	0.12
$q_{0.025}(\cdot\|Y,\psi)$	1	0.12	0.19	0.20	1.30	1.16	0.88
$q_{0.975}(\cdot\|Y,\psi)$	1	0.14	0.29	0.30	1.65	1.85	1.34
$E(\cdot\|Y,\psi)$	2	6.27	1.42	1.30	0.66	0.89	4.45
$SD(\cdot\|Y,\psi)$	2	3.62	1.45	1.01	0.19	1.32	4.45
$q_{0.025}(\cdot\|Y,\psi)$	2	0.16	0.02	0.08	0.39	0.17	0.37
$q_{0.975}(\cdot\|Y,\psi)$	2	13.08	4.30	3.81	1.10	4.95	15.85
	ψ	p_{12}	p_{14}	p_{23}	p_{24}		
$E(\cdot\|Y,\psi)$	1	0.71	0.29	0.72	0.28		
$SD(\cdot\|Y,\psi)$	1	0.04	0.04	0.08	0.08		
$q_{0.025}(\cdot\|Y,\psi)$	1	0.64	0.21	0.56	0.12		
$q_{0.975}(\cdot\|Y,\psi)$	1	0.79	0.36	0.88	0.44		
$E(\cdot\|Y,\psi)$	2	0.71	0.29	0.16	0.84		
$SD(\cdot\|Y,\psi)$	2	0.36	0.36	0.23	0.23		
$q_{0.025}(\cdot\|Y,\psi)$	2	0.02	0.00	0.00	0.13		
$q_{0.975}(\cdot\|Y,\psi)$	2	1.00	0.98	0.87	1.00		

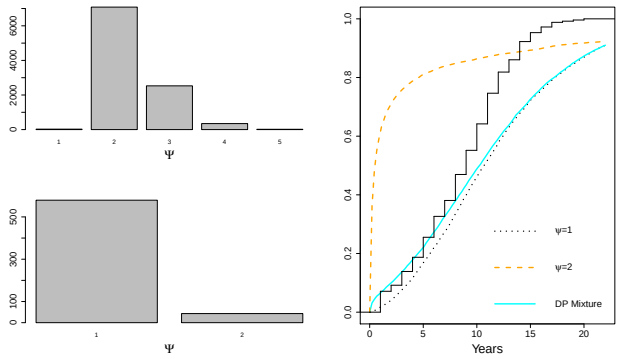


FIGURE 1. Maximum number of mixture components observed for each iteration (top left) distribution of the observations across the estimated components (bottomn left); death time cumulative posterior predictive distributions (right).

References

Barone, R. and Tancredi, A. (2022) Bayesian inference for discretely observed continuous time multi-state models. *Statistics in Medicine*

Fruhwirth-Schnatter, S. and Pamminger, C. (2010) Model-based clustering of categorical time series. *Bayesian Analysis* 5, 345-368

Hobolth, A. abd Stone, E. (2009) Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *AoAS* 3, 1204-1231

Jackson, C. (2011) Multi-state models for panel data: the msm package for R *Journal of Statistical Software* 38, 1 -29

Luo, Y., Stephens, D., and Buckeridge D. (2021). Bayesian clustering for continuous- time hidden Markov models. *Canadian Journal of Statistics*

Boosting for variance components in mixed models

Michela Battauz[1], Paolo Vidoni[1]

[1] Department of Economics and Statistics - University of Udine, Italy

E-mail for correspondence: michela.battauz@uniud.it

Abstract: The boosting algorithm was originally proposed in the machine learning literature as a means to obtain improved ensemble classification procedures. This idea has been developed in the statistical setting, with the aim of fitting regression models using a sequential procedure which performs also variable selection. In this paper a new component-wise boosting algorithm is applied for selecting variance components in linear mixed models. This represents the novelty of the proposal since so far the focus has been on the fixed part of the model.

Keywords: Random Effects; Regularization; Statistical Learning.

1 Introduction

Mixed models are widely used to account for the correlations among observations nested in groups or collected over time. When variable selection is an issue, the boosting approach proposed by Tutz and Groll (2010) can be considered, even if the main attention is on the fixed effects, while the random effects should be pre-specified. The purpose of our proposal is to deal with the random part of the model and to develop a data-driven method to automatically select which variables have a random effect. This objective poses new challenges, since the boosting methods proposed in the literature can not be directly applied. In fact, both the gradient boosting (Friedman, 2001) and the likelihood-based boosting (Tutz and Binder, 2006) approaches are based on the gradient of the objective function, which, in this case, is null in the starting point of the algorithm. As a result, these algorithms are not able to move from the initial point and then alternative strategies should be defined. In this paper, we propose a new component-wise boosting algorithm based on directions of negative curvature, besides

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

the classical Newton direction. After introducing the model and the likelihood, we describe the new algorithm and we present the results of a preliminary simulation study which confirms the validity of the approach.

2 Model specification and likelihood function

The mixed effects model for group i can be expressed as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, M,$$
$$\mathbf{b}_i \sim N(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}), \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

where $\mathbf{y}_i$ is a vector of responses with dimension n_i, $\mathbf{X}_i$ is a matrix of predictors with dimension $n_i \times p$, $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effects, $\mathbf{Z}_i$ is a matrix of predictors with dimension $n_i \times q$ and $\mathbf{b}_i$ is a q-dimensional vector of random effects. Using the Cholesky factorization of the covariance matrix $\boldsymbol{\Sigma} = \mathbf{C}^T\mathbf{C}$, the random effects can be expressed as $\mathbf{b}_i = \mathbf{C}^T\mathbf{u}_i$ and the model can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{C}^T\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, M,$$
$$\mathbf{u}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

which is convenient for computational purposes, when some components in $\boldsymbol{\Sigma}$ are zero. Let $\boldsymbol{\theta}$ be the vector of parameters that determine $\mathbf{C}$. Similarly to Bates and DebRoy (2004), the likelihood function for the data in group i is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \int \frac{1}{(2\pi\sigma^2)^{(n+q)/2}} \exp\left(\frac{||\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{C}^T\mathbf{u}_i||^2 + \mathbf{u}_i^T\mathbf{u}_i}{-2\sigma^2}\right) d\mathbf{u}_i,$$

and the profile log-likelihood for $\boldsymbol{\theta}$ is

$$\ell_p(\boldsymbol{\theta}) = -\frac{1}{2}\log\left(|\mathbf{C}\mathbf{Z}_i^T\mathbf{Z}_i\mathbf{C}^T + \mathbf{I}|\right) - \frac{n_i}{2}\left[1 + \log\left(\frac{2\pi r_{yy}^2}{n}\right)\right],$$

where r_{yy} derives from the following decomposition:

$$\begin{pmatrix} \mathbf{C}\mathbf{Z}_i^T\mathbf{Z}_i\mathbf{C}^T + \mathbf{I} & \mathbf{C}\mathbf{Z}_i^T\mathbf{X}_i & \mathbf{C}\mathbf{Z}_i^T\mathbf{y}_i \\ \mathbf{X}_i^T\mathbf{Z}_i\mathbf{C}^T & \mathbf{X}_i^T\mathbf{X}_i & \mathbf{X}_i^T\mathbf{y}_i \\ \mathbf{y}_i^T\mathbf{Z}_i\mathbf{C}^T & \mathbf{y}_i^T\mathbf{X}_i & \mathbf{y}_i^T\mathbf{y}_i \end{pmatrix} = \mathbf{R}_e^T\mathbf{R}_e,$$

with

$$\mathbf{R}_e = \begin{pmatrix} \mathbf{R}_{ZZ} & \mathbf{R}_{ZX} & \mathbf{r}_{Zy} \\ \mathbf{0} & \mathbf{R}_{XX} & \mathbf{r}_{Xy} \\ \mathbf{0} & \mathbf{0} & r_{yy} \end{pmatrix}.$$

Furthermore, the matrix $\mathbf{C}$ is written as $\boldsymbol{\Gamma}\mathbf{D}$, where $\boldsymbol{\Gamma}$ is upper triangular with ones on the diagonal and $\mathbf{D}$ is a diagonal matrix, so that a zero element on the diagonal of $\mathbf{D}$ determines that the variance of the corresponding random effect is zero. This parameterization was adopted also in Bondel et al. (2010).

3 The boosting algorithm

Starting from a model with variance components all set equal to zero, the algorithm, at each step, updates the parameter in $\boldsymbol{\theta}$ that leads to the largest decrease of the objective function, which is the negative profile log-likelihood. To this end, two alternative directions are computed at each step: a negative curvature direction and a Newton direction. The negative curvature direction corresponds to the eigenvector associated to the minimum negative eigenvalue (if any) of the Hessian matrix. Since only one parameter is considered at a time, the eigenvalue is simply given by the value on the diagonal of the Hessian matrix and the eigenvector is equal to 1. The decrease of the objective function along each direction is evaluated on the basis of its quadratic approximation. The direction so obtained is then multiplied by a small step-length in order to produce a weak learner. The sequential update of $\boldsymbol{\theta}$ continues until a suitable stopping criterion is satisfied.
The algorithm is a special instance of the optimization method proposed in Gould et al. (2000) and a similar approach was employed to deal with the estimation of factor analysis models for binary data in Battauz and Vidoni (2022). The procedure is implemented in R and C++. Computational methods similar to Bates and DebRoy (2004) have been employed for an efficient evaluation of the profile log-likelihood function and its derivatives, which is essential to obtain a fast procedure.

4 A preliminary simulation study

The performance of the proposal was investigated through a simulation study. The settings are similarly to Bondell et al. (2010). The data are generated form a model with 3 random effects with covariance matrix

$$\begin{pmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{pmatrix}$$

and σ^2=1. Three cases are considered: 1. $M = 30$, $n_i = 5$ and 4 potential variables, 2. $M = 60$, $n_i = 10$ and 4 potential variables, 3. $M = 60$, $n_i = 5$ and 6 potential variables. For each of them, 200 datasets were generated. The independent variables are generated from a uniform $(-2, 2)$ distribution. As stopping rule, we used the conditional AIC (Vaida and Blanchard, 2005). In case 1. our boosting algorithm correctly identifies the variance component equal to zero in 85% of the replications. In case 2. this rate is equal to 90%, and it reaches 94% in case 3. However, the algorithm tends to underestimate the variance components that are different from zero and, although the variability of the estimates obtained with boosting is lower than the variability of the maximum likelihood estimates, the root mean square error tends to be higher for the former.

5 Conclusions

The proposal of this paper revealed effective in detecting which variables have a random effect. However, the specification of the stopping criterion and the study of the performance of the boosting estimators are issues that require further investigation.

References

Battauz, M., and Vidoni, P. (2022). A likelihood-based boosting algorithm for factor analysis models with binary data. *Computational Statistics and Data Analysis*, **168**, 107412.

Bates, D. M., and DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, **91**, 1 – 17.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*, **66**, 1069 – 1077.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189 – 1232.

Gould, N. I. M., Lucidi, S., Roma, M., and Toint, PH.L. (2000). Exploiting negative curvature directions in linesearch methods for unconstrained optimization. *Optimization Methods and Software*, **14**, 75 – 98.

Tutz, G., and Binder, H. (2006). Generalized Additive Modeling with Implicit Variable Selection by Likelihood-Based Boosting. *Biometrics*, **62**, 961 – 971.

Tutz, G., and Groll, A. (2010). Generalized Linear Mixed Models Based on Boosting. In: Kneib T., Tutz G. (Eds.) *Statistical Modelling and Regression Structures*. Berlin: Physica-Verlag.

Vaida, F., and Blanchard, S. (2005). Conditional Akaike Information for Mixed-Effects Models. *Biometrika*, **92**, 351 – 370.

A Model of Individual BMI Trajectories

Laurens Bogaardt[1], Anoukh van Giessen[1], Susan Picavet[1] and Hendriek Boshuizen[1]

[1] National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands.

E-mail for correspondence: laurens.bogaardt@rivm.nl

Abstract: A model of BMI is an important building block of health simulations aimed at estimating goverment policy effects with regard to obesity. We created a model of BMI which shows realistic behaviour at an individual level but which also generates representative, population level distributions. The model is constructed by combining two datasets. Firstly, the population level distribution is extracted from a large, cross-sectional dataset. In addition, longitudinal data is used to model how individuals move along typical trajectories over time.

Keywords: BMI, Statistical Model, Individual Trajectories, Micro-Simulation

1 Introduction

Overweight and obesity pose significant health risks in many countries (Dai et al. 2020). Consequently, governments spend much effort to curb the increasing trends (Van Rinsum et al. 2018). Investigating what course of action is most fruitful is often done using health simulations (Levy at al. 2011). A model of BMI is the first step in such analyses. A common modelling strategy is to assign to each individual a percentile within the population level distribution, assuming this relative position stays fixed over their lifetime (McPherson et al. 2007, OECD 2019). A more realistic model can have the percentile of each individual fluctuate over time as well.
We created a model of BMI which shows realistic behaviour at an individual level but which also generates representative, population level distributions using data of the adult population of the Netherlands. We cleverly combine cross-sectional data with longitudinal data. The cross-sectional dataset provides representative information about the population level distribution of BMI. At an individual level, BMI fluctuates over time, following typical trajectories. These trajectories are modelled using the longitudinal data.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Population Level Distribution

To describe the population level distribution of BMI, we require a cross-sectional dataset representative of the Netherlands and large enough to accommodate various stratifications of the population. For government policy simulations, stratifying by sex, education level and age is usually adequate. The Public Health Monitor dataset is a Dutch cross-sectional dataset based on a large, health-related questionnaire (GGD'en, CBS en RIVM 2012). This questionnaire was administered in 2012 by the Community Health Services, Statistics Netherlands and the National Institute for Public Health and the Environment. To deal with non-normal BMI values, we use the flexible sinh-arcsinh normal distribution defined by four parameters; μ, σ, ν and τ (Jones and Pewsey 2009, Jones and Pewsey 2019). We can incorporate age into this analysis by making use of the *GAMLSS* package in R (Rigby and Stasinopoulos 2005, Stasinopoulos and Rigby 2007, R Core Team 2021). This allows us to model each of the four distribution parameters as functions of the predictors. Initially, we fitted splines to all four of the parameters, for both sexes and all three education levels. This indicated an approximately quadratic relationship with age for the μ and σ parameters, whereas the education level mostly impacted their intercept. The ν and τ parameters barely differed by age or education. So we repeated the analysis for a restricted, parametric model, given by equation 1.

$$\begin{aligned} \mu &= \mu_{education} + \mu_{age} \times age + \mu_{age^2} \times age^2 \\ \sigma &= \sigma_{education} + \sigma_{age} \times age + \sigma_{age^2} \times age^2 \\ \nu &= \nu_{intercept} \\ \tau &= \tau_{intercept} \end{aligned} \tag{1}$$

The resulting coefficients are listed in table 1.

TABLE 1. The coefficients for the population level BMI distribution.

	$\mu^{low}_{educa.}$	$\mu^{mid}_{educa.}$	$\mu^{high}_{educa.}$	μ_{age}	μ_{age^2}	$\nu_{intercept}$
Male	18.05	17.75	17.20	0.2573	-0.002137	0.2626
Female	19.00	18.29	17.61	0.1560	-0.001174	0.4436
	$\sigma^{low}_{educa.}$	$\sigma^{mid}_{educa.}$	$\sigma^{high}_{educa.}$	σ_{age}	σ_{age^2}	$\tau_{intercept}$
Male	2.220	1.924	1.677	0.03491	-0.0003327	0.7680
Female	2.728	2.398	1.960	0.03460	-0.0002926	0.8302

To examine the model's goodness of fit, we can first group individuals according to whether they have underweight, normal weight, overweight or obesity. Then we compare the prevalences from the predicted values to the Public Health Monitor 2012 data, as shown in figure 1. The fit seems good.

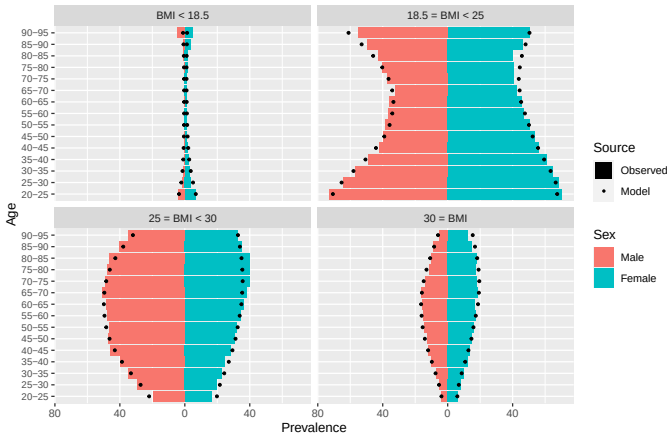


FIGURE 1. The observed and fit prevalences of BMI categories by sex and age.

3 Individual Trajectories

Next, we want to understand individual trajectories. Our approach is to reduce the BMI values in our longitudinal data to z-scores following a transformation based on the population level distribution and using values for μ, σ, ν and τ appropriate for the individual's sex, education level and age. Subsequently, we assume that individuals' z-scores follow stochastic trajectories over their lifetimes and that their BMI values are those z-scores back-transformed to the population level distribution.

This requires longitudinal data with multiple measurements over a relatively long time. The Doetinchem Cohort Study provides such data. This study has followed a group of individuals from the municipality of Doetinchem in the Netherlands for the past 30 years (Verschuren et al. 2008). Its aim is to study lifestyle factors and biological risk factors on aspects of health. The participants underwent a health examination about every 5 years since 1987. A key feature of this panel is that characteristics such as BMI were measured by research assistants instead of being self-reported.

From section 2, we know the BMI distribution stratified by sex, education level and age using the four parameters μ, σ, ν and τ. This distribution implies a transformation to z-scores. All the BMI measurements from our longitudinal studies can be transformed using the specific parameters associated with the individuals' sex, education level and age. This procedure removes the dependencies on these characteristics from the z-scores.

The individual trajectories are assumed to contain long-term, medium-term and short-term effects which can be described by a mixed effects model. The long-term effects are operationalised as the random intercept, which indicates the tendency to belong to the upper or lower percentiles of the BMI distribution. In the percentile-method, this random intercept is the

only effect which determines the trajectories. Our model generalises this method by including medium-term effects which are represented by an autoregressive process (AR1). This process assumes that, at each time period, a random shock occurs which either pushes the BMI z-score up or down. The effect of a shock decays exponentially over time, similar to how habits wax and wane. The overall effect is the sum of all previous shocks, which results in a meandering BMI value with temporal autocorrelation. Short-term effects are modelled as additional uncorrelated error representing daily fluctuations in weight. Models with these three components are described in detail in Diggle (1994) and Verbeke and Molenberghs (2000). Equation 2 shows our model for the vector Z_i of BMI z-scores of individual i. Here, J is the matrix of only ones and I is the identity matrix. It was fit to our longitudinal data using the *nlme* package in R (Pinheiro et al. 2021).

$$\begin{aligned} Z_i &= \beta_0 + \beta_1 \times age + RI_i + AR1_i + \epsilon_i \\ RI_i &\sim \mathcal{N}(0, \sigma^2_{intercept} \times J) \\ AR1_i &\sim \mathcal{N}(0, \Sigma) \text{ where } \Sigma_{tt'} = \sigma^2_{correlated} \times \rho^{|t-t'|}_{temporal} \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2_{uncorrelated} \times I) \end{aligned} \tag{2}$$

We know that Doetinchem is not a completely representative sample, so it will have a different population level distribution then the one obtained in section 2 which we used to transform the BMI values into z-scores. Consequently, this procedure need not result in values with zero mean and unit standard deviation; some bias may remain. Ultimately, we want a model which suits the entire population, so which produces true z-scores. A simple solution is to model the bias by including a fixed intercept and fixed slope in our statistical analysis. We are not interested in the values of this intercept and slope, but by including the terms, the other parameters are not compromised. The estimated values are listed in table 2.

TABLE 2. The parameters of the BMI z-score trajectories.

	$\sigma_{intercept}$	$\rho_{temporal}$	$\sigma_{correlated}$	$\sigma_{uncorrelated}$
Male	0.0020	0.9878	0.1535	0.1657
Female	0.0013	0.9887	0.1463	0.2120

To provide some feeling for the model, we can generate z-scores for a few individuals and visually compare these to values from the Doetinchem Cohort Study. Although it depends on the precise random samples which are drawn, figure 2 shows that, on the face of it, the generated z-scores and the observed data are similar. This give credence to our idea that modelling z-scores using a mixed effects model with an AR1 process yields realistic BMI trajectories. Transforming these z-scores to BMI values can be done using the appropriate values for μ, σ, ν and τ, which depend on the individual's sex, education level and age following equation 1.

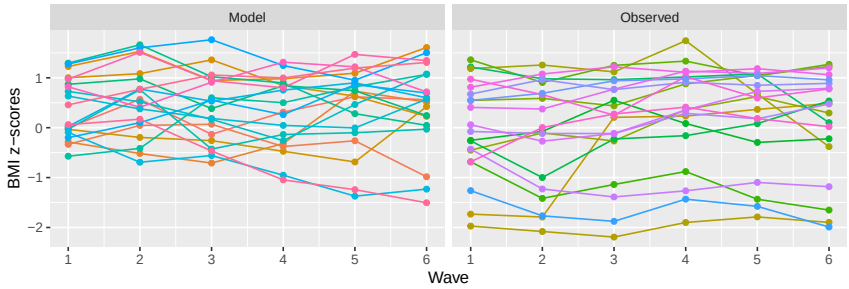


FIGURE 2. A sample of observed and model generated z-scores by wave.

4 Conclusion

To sum up, we fitted a population level BMI distribution, stratified by sex, education level and age, to a cross-sectional dataset of the Netherlands in 2012. Subsequently, we modelled individual trajectories as the z-scores of this distribution by making use of a mixed effects model with long-term, medium-term and short-term effects. Some limitations do remain. For one, the BMI values used to fit the model are self-reported. Previous research has shown that there can be a discrepancy between self-reported and measured BMI values and that individuals with a high BMI tend to underreport their weight (Olfert 2018). Another important limitation concerns the application of government interventions on BMI. When a lifestyle intervention is simulated, the direct effect on BMI must be modelled, including how its impact wanes over time. It would be incorrect to assume the temporal autocorrelation found in our prediction model implies something about the speed at which intervention effects decay.

The methods outlined here may be extended in various ways. First of all, the set of co-variates can be expanded by any other predictor which has some association with BMI. The sole requirement is that this predictor is found in both the cross-sectional and the longitudinal data. Secondly, the BMI trajectories of children and adolescents could be included. Likewise, the method of fitting a flexible distribution at the population level and modelling longitudinal z-scores as a mixed effects model could be applied to other continuous variables such as daily sugar intake or blood pressure. And finally, a generalisation which predicts multiple risk factors simultaneously can be made, which would be a great addition to analyses of government intervention on lifestyle choices.

References

Dai, H., Alsalhe, T.A., Chalghaf, N., Riccò, M., Bragazzi, N.L. and Wu, J. (2020). The Global Burden of Disease Attributable to High BMI in 195 Countries and Territories. *PLOS Medicine*, **17**, 1 – 19.

Diggle, P.J., Liang, K-Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

GGD'en, CBS en RIVM (2012). Gezondheidsmonitor Volwassenen.

Jones, M.C. and Pewsey, A. (2009). Sinh-Arcsinh Distributions. *Biometrika*, **96**, 761 – 780.

Jones, M.C. and Pewsey, A. (2019). The Sinh-Arcsinh Normal Distribution. *Significance*, **16**, 6 – 7.

Levy, D.T., Mabry, P.L., Wang, Y.C., Gortmaker, S., Huang, T.T.K., Marsh, T., Moodie, M. and Swinburn, B. (2011). Simulation Models of Obesity. *Obesity Reviews*, **12**, 378 – 394.

McPherson, K., Marsh, T. and Brown, M. (2007). Tackling Obesities: Future Choices – Modelling Future Trends in Obesity and the Impact on Health. 1 – 76.

OECD (2019). SPHeP-NCDs Documentation.

Olfert, D.M., Barr, M.L., Charlier, C.M., Famodu, O.A., Zhou, W., Mathews, A.E., Byrd-Bredbenner, C. and Colby, S.E. (2018). Self-Reported vs. Measured Height, Weight, and BMI in Young Adults. *Int. J. Environ. Res. Public Health*, **15**, 1 – 9.

Pinheiro, J., Bates, D., DebRoy S., Sarkar D. and R Core Team (2021). nlme: Linear and Nonlinear Mixed Effects Models.

R Core Team (2021). R: A Language and Environment for Statistical Computing.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society: Series C*, **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. *R Foundation for Statistical Computing.*

Van Rinsum, C., Gerards, S., Rutten, G., Philippens, N., Janssen, E., Winkens, B., Van de Goor, I. and Kremers, S. (2018). The Coaching on Lifestyle (CooL) Intervention for Overweight and Obesity. *Int. J. Environ. Res. Public Health*, **15**, 1 – 27.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer.

Verschuren, W., Blokstra, A., Picavet, H. and Smit, H. (2008). Cohort Profile: The Doetinchem Cohort Study. *International Journal of Epidemiology*, **37**, 1236 – 1241.

Probabilistic load forecasting via dynamic quantile regression

Cristian Castiglione[1], Mauro Bernardi[1]

[1] Department of Statistical Sciences, University of Padova, Padova, Italy

E-mail for correspondence: cristian.castiglione@phd.unipd.it

Abstract: Reliable short-term predictions play a crucial role when it comes to supporting the decision-making process of any quickly evolving industry, especially in the energy sector, where supply, demand and prices are characterized by high volatility and structural market changes. We then propose a new dynamic quantile regression model for estimating and forecasting the short-term evolution of power load consumption in the US. A state-space representation is considered in order to disentangle different signal components, such as smooth trends, cycles, stationary fluctuations and, possibly, non-linear covariate effects. Taking a Bayesian perspective, the parameters and latent states are estimated with an efficient variational Bayes approximation. Several quantile predictions are then combined to describe the future distribution of the power load in a probabilistic forecasting vein, that not only provides a pointwise estimate but also delivers a rich description of the future uncertainty. The novelty of our approach is to combine quantile regression, additive models and state-space models in a unified framework able to completely characterize the underlying determinants of the power load consumptions.

Keywords: Additive models; State-space models; Probabilistic load forecasting; Quantile regression; Variational Bayes.

1 Data

We consider the dataset proposed in the Global Energy Forecasting Competition 2014 (Hong, et al., 2016), which collects the US power load consumption (`load`) in megawatt-per-hour (MWh) from January 2005 to December 2011, along with some additional environmental and temporal covariates. Here we only take into account the variables selected by Galiard, et al. (2016) to be the most relevant for prediction purposes, that are the atmospheric temperature ($\texttt{temp}_t$) in Celsius scale (C°), the smoothed tem-

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

perature ($\texttt{stemp}_t$), given by $\texttt{stemp}_t = 0.05 \cdot \texttt{temp}_t + 0.95 \cdot \texttt{temp}_{t-1}$, the day of the year ($\texttt{dayyear}_t$), the day of the week ($\texttt{dayweek}_t$), and a trend variable ($\texttt{trend}_t$) correcting for possibly non-stationary components. As it is common in the energy forecasting literature, we proceed by modelling separately each half an hour of the day, and, in particular, we restrict our analysis to the time interval 11:30–12:00 a.m. so that to obtain equispaced daily observations.

Our main goal is to propose a flexible model able to predict the unobserved future distribution of the power load considering the non-linear effect of temperature, cycles and trends. In literature, this challenge has been faced either following a time series approach, for example using SARIMA models (Hong, et al., 2016) and with a semiparametric regression approach using GAMs (Galiard, et al., 2016; Fasiolo, et al., 2021). We then combine these two methods within a quantile regression framework by specifying a non-Gaussian dynamic linear model written in state-space form, so that to handle the non-linear effect of the covariates through an additive specification and to model the non-stationary trend over time as the realization of a stochastic latent Markov process.

2 Model and inference

We propose to predict the short-term evolution of the τ-quantile of the power load consumption through the following Bayesian dynamic linear model (Yu and Moyeed, 2001)

$$y_t = \mu_t + x_t^{\mathrm{T}}\beta + \varepsilon_t, \quad \varepsilon_t \sim AL(\tau, 0, \sigma_\varepsilon^2), \quad t = 1, \ldots, n, \tag{1}$$

where y_t is a real response variable (the rescaled power load) observed at time t, μ_t is a stochastic trend evolving over time, x_t is a $d \times 1$ vector of exogenous covariates, β is a $d \times 1$ vector of regression parameters and ε_t is a independent error component distributed according with an asymmetric-Laplace (AL) distribution with shape $\tau \in (0, 1)$, location 0 and scale σ_ε^2. The stochastic trend μ_t is modelled as a second order continuous random walk specified through the dynamical linear equation:

$$\begin{aligned} &\mu_{t+1} = \mu_t + \delta_t\dot{\mu}_t + \eta_t, \\ &\dot{\mu}_{t+1} = \dot{\mu}_t + \dot{\eta}_t, \end{aligned} \qquad \begin{bmatrix} \eta_t \\ \dot{\eta}_t \end{bmatrix} \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_\eta^2 \begin{bmatrix} \delta_t^3/3 & \delta_t^2/2 \\ \delta_t^2/2 & \delta_t \end{bmatrix}\right), \tag{2}$$

where μ_t is the trend level, $\dot{\mu}_t$ is the trend slope, δ_t is the length of the interval spacing two different time-points, which corresponds to $\delta_t = 1$ when no missing data are present, and $(\eta_t, \dot{\eta}_t)$ is a correlated innovation term. We then assume a diffuse initial distribution for $(\mu_0, \dot{\mu}_0)^{\mathrm{T}} \sim N_2(0, \kappa I_2)$ with $\kappa \to \infty$. Model (2) may be further enriched by exploring different specifications of the transition equation, as described e.g. by Durbin and Koopman (2012).

For the linear predictor $x_i^{\mathrm{T}}\beta$ we assume an additive model specification, being $x_i^{\mathrm{T}}\beta = x_{i1}^{\mathrm{T}}\beta_1 + \cdots + x_{iK}^{\mathrm{T}}\beta_K$, where each component $x_{ik}^{\mathrm{T}}\beta_k$ can represent either a fixed linear effect, or a basis expansion describing a possibly non-linear effect of some covariates on the response. Both x_{ik} and β_k are $d_k \times 1$ vectors, so that $d = d_1 + \cdots + d_k$.
To complete the Bayesian model specification, we assume independent multivariate Gaussian (N) prior for the regression parameters and conjugate inverse-Gamma (IG) prior for the scale parameters:

$$\begin{aligned} \beta_k|\sigma_k^2 \sim N_{d_k}(0, \sigma_k^2 R_k^{-1}), \quad \sigma_k^2 \sim IG(A_\beta, B_\beta), \\ \sigma_\eta^2 \sim IG(A_\eta, B_\eta), \quad \sigma_\varepsilon^2 \sim IG(A_\varepsilon, B_\varepsilon), \end{aligned} \tag{3}$$

In vector form, we denote $\beta = (\beta_1^{\mathrm{T}}, \ldots, \beta_K^{\mathrm{T}})^{\mathrm{T}}$ and $\sigma_\beta^2 = (\sigma_1^2, \ldots, \sigma_K^2)^{\mathrm{T}}$. Here, the constants $A_\beta, B_\beta, A_\eta, B_\eta, A_\varepsilon, B_\varepsilon > 0$ are scalar user-specified prior parameters, as well as R_k is a semi-positive definite matrix determining the prior conditional dependence structure among the elements of β_k.
In particular, we have

$$x_t^{\mathrm{T}}\beta = f_1(\texttt{temp}_t) + f_2(\texttt{stemp}_t) + f_3(\texttt{dayyear}_t) + f_4(\texttt{dayweek}_t),$$

where each non-linear function $f_k(\cdot) = x_k(\cdot)^{\mathrm{T}}\beta_k$, $k = 1, \ldots, 4$, represents a cubic B-spline expansion with associated second order differential penalization matrix R_k.

3 Estimation

The posterior distribution of the parameter vector $\theta = (\beta, \mu, \dot{\mu}, \sigma_\beta^2, \sigma_\eta^2, \sigma_\varepsilon^2)$ can be inferred in approximated form via mean field variational Bayes (Ormerod, et al., 2010; Blei, et al. 2017). We thus replace the true posterior with the factorized density

$$q(\theta) = q(\beta, \mu, \dot{\mu}, \sigma_\beta^2, \sigma_\eta^2, \sigma_\varepsilon^2) = q(\beta, \mu, \dot{\mu})\, q(\sigma_\beta^2)\, q(\sigma_\eta^2)\, q(\sigma_\varepsilon^2) \tag{4}$$

and we minimize the Kullback-Leibler divergence between $q(\theta)$ and $p(\theta|y)$ with respect to $q(\theta)$ in order to find the optimal approximation. Closed form coordinate-wise solutions can be easily derived for σ_β^2, σ_η^2 and σ_ε^2, that is

$$q^*(\psi) \propto \exp\left\{E_{-\psi}\left[\log p(\psi|\text{rest})\right]\right\} \tag{5}$$

for a generic parameter ψ, with $E_{-\psi}(\cdot)$ denoting the variational expectation over all the parameters except ψ and $p(\psi|\text{rest})$ begin the full-conditional density of ψ. Numerical optimization and Kalman filter routines (Durbin and Koopman, 2012) have to be employed for finding the optimal density of $(\beta, \mu, \dot{\mu})$ under the constrain that $q(\beta, \mu, \dot{\mu})$ belongs to the family of multivariate Gaussian distributions.

The implementation of the joint optimization scheme for all the parameters in the model thus relies on a semiparametric variational Bayes (Rohde and Wand, 2019) approach based on the so-called Knowles-Minka-Wand update (Knowles, Minka, 2011; Wand, 2014).

4 Empirical results

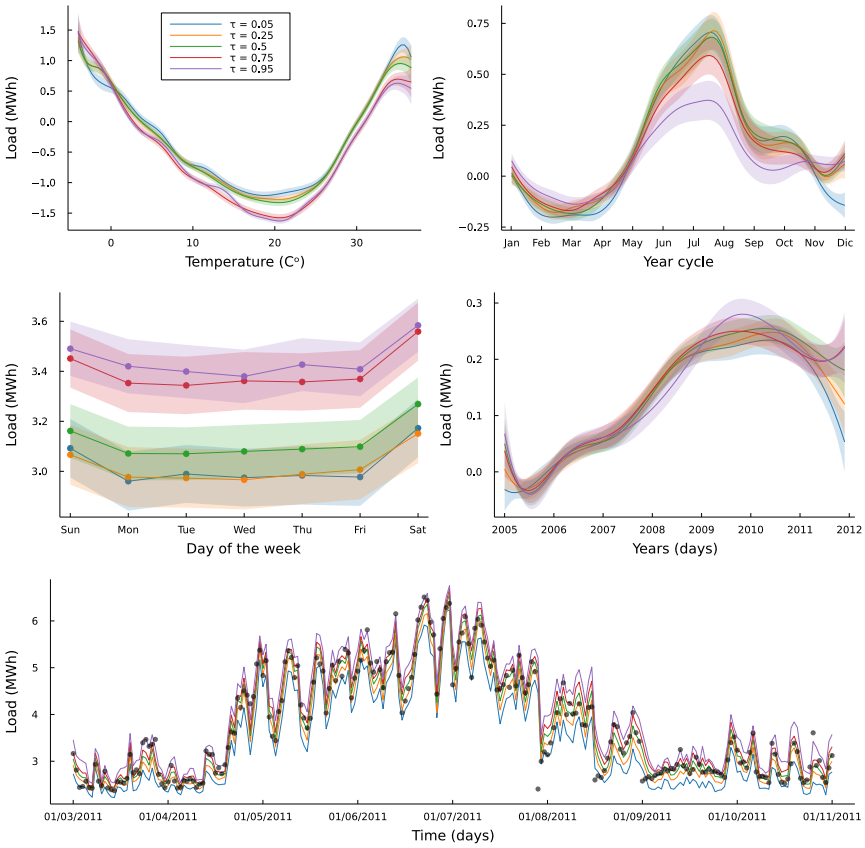


FIGURE 1. Top and middle rows: Marginal effect of temperature, year cycle, day of the week and long-term trend on the 5 considered quantiles. Bottom row: one-step-ahead predictions of the load distribution in the period going from March 1st to December 1st, 2012.

A non-parametric description of the predictive distribution of the power load can be obtained by estimating several quantile curves and then aggregating the results for approximating the cumulative density function at a given time point. We then consider the quantiles corresponding to

$\tau = 5\%, 25\%, 50\%, 75\%, 95\%$ and we use them both to understand how and how much some variables impact the power load distribution and to produce point and interval predictions.
Figure 1 shows the marginal effect of temperature, year cycle, day of the week and non-stationary trend on the load consumption. Moreover, it portrays the one-step-ahead forecasts from March to December 2011.
The temperature effect has a classical U-shape profile, with a minimum of around 20 C°, indicating a low consumption when the temperatures are moderate. Then, the residual power load increases during summer and falls during autumn with a cyclic behaviour not explained by the temperature effect. An almost constant consumption is observed during the weekdays, with a small growth on Saturday and Sunday. Looking at a long-term perspective, the load trend was increasing steeply until 2009-2010, when it flexed to a negative slope for the remaining period.
Comparing heterogeneous quantile curves, some differences arise by dividing the lower from the higher percentile levels, respectively $\tau = 5\%$, 25%, 50% and $\tau = 75\%$, 95%. This second group is characterized by a sharper U-profile in the temperature variable, a flatter cyclic fluctuation and a higher mean level in the day effect. Symmetric interpretations hold for the other group. Excluding the very first and last observed months, for which the estimates are more variable, no significant differences emerge in the trend dynamics.
The quality of the predictions is assessed by comparing the theoretical and observed quantile level, respectively, τ and $\hat{\tau}$. The later is estimated by averaging the number of occurrences in which $y_t \leq \hat{\mu}_t + x_y^{\mathrm{T}}\hat{\beta}$. For all the considered quantile levels there is no significant difference between τ and $\hat{\tau}$, suggesting a good ability of the model to recover the conditional quantile of the power load.

TABLE 1. Theoretical and observed quantile level, respectively, τ and $\hat{\tau}$.

Theoretical	Observed	Std.Dev.	C.I. (95%)
0.05	0.0467	0.0118	(0.0274, 0.0661)
0.25	0.2555	0.0244	(0.2155, 0.2954)
0.50	0.4891	0.0279	(0.4433, 0.5349)
0.75	0.7664	0.0237	(0.7276, 0.8051)
0.95	0.9470	0.0125	(0.9265, 0.9676)

References

Durbin, J., Koopman, S.J. (2012). *Time series analysis by state space method*. OUP Oxford.

Fasiolo, M., Wood, S.N., Zaffran, M., Nedellec, R., Goude, Y. (2021). Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, **116**, 1402 – 1412.

Gaillard, P., Goude, Y., Nedellec, R. (2016). Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, **32**, 1038 – 1050.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, **32**, 896 – 913.

Knowles, D., Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. *Advances in Neural Information Processing Systems*, **24**, 1701 – 1709.

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

Rohde, D., Wand, M.P. (2016). Semiparametric mean field variational Bayes: General principles and numerical issues. *Journal of Machine Learning Research*, **17**, 5975 – 6021.

Wand M.P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*, **15**, 1351 – 1369.

Yu, K., Moyeed, R.A. (2014). Bayesian quantile regression. *Statistics & Probability Letters*, **54**, 437 – 447.

On the first two size-biased picks from the normalized Inverse Gaussian prior

Annalisa Cerquetti[1]

[1] Department of Economics and Management, University of Trento, Italy

E-mail for correspondence: annalisa.cerquetti@unitn.it

Abstract: The normalized inverse Gaussian random discrete distribution has been deeply investigated in Bayesian nonparametrics as one of the possible tractable alternatives to the Ferguson-Dirichlet prior and to its two-parameter Pitman-Yor extension. Here we devise an easy to sample representation of its first two size-biased picks and discuss a potential application to prior calibration.

Keywords: Bayesian nonparametrics. Monte Carlo sampling. Normalized Inverse Gaussian prior. Size-biased permutation.

1 Introduction

The normalized Inverse Gaussian random discrete distribution arises by normalizing the ranked jumps of a 1/2-stable subordinator, conditioning on the total sum and mixing with the corresponding exponentially tilted density (see e.g. Pitman, 2003, Cerquetti, 2007). It has been extensively investigated in Bayesian nonparametrics as a tractable alternative to the Ferguson-Dirichlet prior since admits an explicit definition in terms of its finite dimensional distributions and a closed form expression of the weights appearing in the Gibbs-product form of the distribution of the corresponding exchangeable random partition. Preliminary results for its implementation in hierarchical mixture modelling are in Lijoi *et al.* (2005), a stick-breaking representation is in Favaro *et al.* (2012), slice sampling for mixture models is in Favaro and Walker (2012) and asymptotic approximations for the prediction rules of the corresponding Chinese restuarant process and for the prior on the number of clusters are respectively in Arbel and Favaro (2021) and in Bystrova *et al.* (2021).
Here we devise an easy to sample representation of the first two size-biased picks from the normalized Inverse Gaussian which provides a convenient

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

alternative to those presented in Favaro et al. (2012) and Arbel and Favaro (2021). We also illustrate an application to prior calibration.

2 Size-biased sampling of the normalized Inverse Gaussian prior

Random discrete distributions (r.d.d) can be defined by specifying the distribution of their random atoms according to different ordering. Atoms can be described in exchangeable random order, in decreasing order or in size-biased order. Given the collection $(P_j)_{j\geq 1}$ of the ranked atoms of a r.d.d. P, its size-biased permutation corresponds to the sequence $(\tilde{P}_j)_{j\geq 1}$ where $\tilde{P}_j$ denotes the random size of the jth atom discovered in the process of random sampling from P. In particular the first size-biased pick from P is the random variable $\tilde{P}_1$ taking values P_j with probability P_j, for $j = 1, 2, \ldots$. The following representation, which follows from results in Aldous and Pitman (1998), holds for the first and second size-biased picks of the normalized Inverse Gaussian prior.

Proposition 1. *Let X_1 and X_2 be independent standard Normal random variables and T an inverse Gaussian random variable of parameter $(v^{-1}, 1)$ with density*

$$f(t;\mu,1) = \sqrt{\frac{1}{2\pi t^3}} \exp\left(-\frac{(t-v^{-1})^2}{2tv^{-2}}\right), \tag{1}$$

then the first and the second size-biased pick from a normalized Inverse Gaussian random discrete distribution of parameter $b = v^2$ admit the following representation

$$\tilde{P}_1 \stackrel{d}{=} \frac{S_1}{T^{-1}+S_1} \tag{2}$$

and

$$\tilde{P}_2 \stackrel{d}{=} \frac{T^{-1}}{T^{-1}+S_1} - \frac{T^{-1}}{T^{-1}+S_2} \tag{3}$$

for $S_j := \sum_{i=1}^{j} X_i^2$.

While the densities of both (2) and (3) can be hard to handle analytically, sampling from them reduces to sampling standard Gaussian and inverse Gaussian variates and can be easily done e.g. in R (see packages VGAM or copula).

3 Application

The availability of the prior distribution induced on the number of clusters is central in Bayesian nonparametrics for prior specification and calibration. The closed form expression for the prior distribution of the number of

species K_n observed in a sample of size n is well-known for any member of the Gibbs-type family, to which the normalized Inverse Gaussian prior belongs, and corresponds to

$$Pr(K_n = k) = V_{n,k} S_{n,k}^{-1,-\alpha}. \tag{4}$$

Nevertheless, despite both the exchangeable Gibbs weights $V_{n,k}$ and the generalized Stirling numbers $S_{n,k}^{-1,\alpha}$ may be available in closed form, like e.g. for the Inverse Gaussian case, they are combinatorial in nature and usually computationally intensive to handle, becoming intractable for moderate and large sample size.

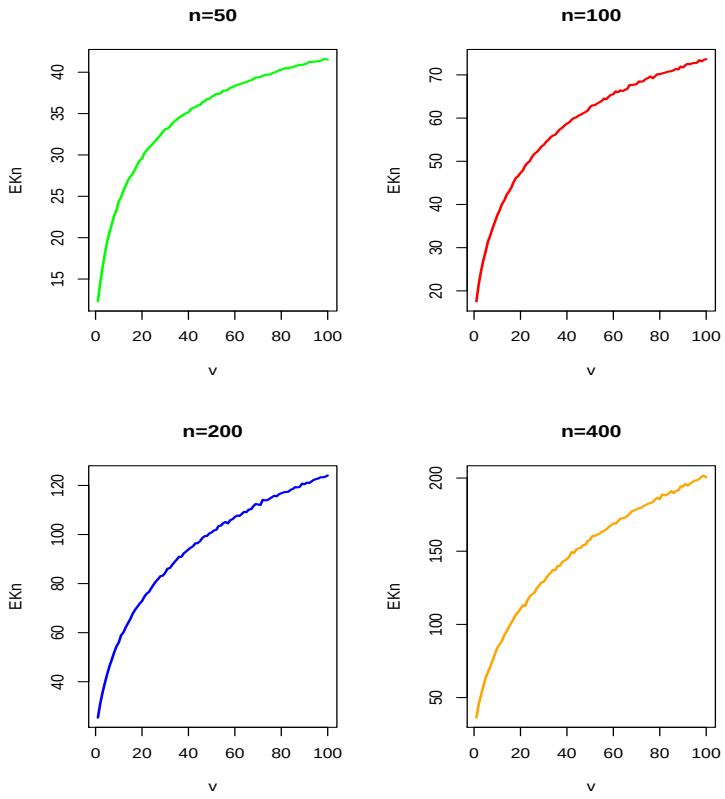


FIGURE 1. Expected number of clusters in samples of size n=50, 100, 200, 400, for increasing values of the parameter v of the normalized Inverse Gaussian prior obtained by Monte Carlo sampling from the first size-biased pick. Total system time for the four plots was 16 seconds with 30.000 Monte Carlo iterations.

Several solutions based on statistical approximations of both P or the $V_{n,k}$ weights have been proposed. See Bystrova et al. (2021) for a comprehen-

sive account. Here for the normalized Inverse Gaussian prior we propose a solution based on working directly with the moments sequence of K_n, exploiting their representation in terms of the size-biased permutation of the prior's atoms. A comprehensive treatment of this approach, providing Monte Carlo solutions to several diversity indicators estimation, will be covered in a forthcoming paper (Cerquetti, 2022). As an example here we show that first and second size-biased picks are enough to obtain fast and accurate evaluation of the behaviour of expectation and uncertainty of the number of clusters for different values of the Inverse Gaussian parameter at different sample sizes. Figure 1 shows the behaviour of $E(K_n)$ for increasing values of v and sample sizes $n = \{50, 100, 200, 400\}$.

References

Aldous, D., Pitman, J. (1998). The standard additive coalescent. *The Annals of Probability*, **26**, 4, 1703 – 1726.

Arbel, J., Favaro, S. (2021). Approximating predictive probabilities of Gibbs-type priors. *Sankhya Series A,*, **83**, 496 – 519.

Bystrova, D., Arbel, J., Kon Kam King, G., Deslandes, F. (2021) Approximating the clusters' prior distribution in Bayesian nonparametric models. In *Third Symp. on Adv. in Approximate Bayesian Inference*.

Cerquetti, A. (2007). A note on Bayesian nonparametric priors derived from exponentially tilted Poisson-Kingman models. *Statistics and Probability Letters*, **77**, 1705 – 1711..

Cerquetti, A. (2022). Moments sequence of diversity indicators under normalized Inverse Gaussian priors by Monte Carlo sampling. *In preparation*.

Favaro, S, Lijoi, A, Prünster, I. (2012) On the stick-breaking representation of normalized inverse Gaussian priors. *Biometrika*, **99**, 663 – 674.

Favaro, S. Walker, S. G. (2012). Slice Sampling σ-Stable Poisson-Kingman Mixture Models *Slice Sampling σ-Stable Poisson-Kingman Mixture Models*, **22**, 830 – 847.

Lijoi, A, Mena, R.H, Prünster, I (2005). Hierarchical Mixture Modeling With Normalized Inverse-Gaussian Priors. *Journal of the American Statistical Association*, **10**, 1278 – 1291.

Pitman, J. (2003). Poisson-Kingman partitions. In *D.R. Goldstein, editor, Science and Statistics: A Festschrift for Terry Speed*, volume 40 of Lecture Notes-Monograph Series, pages 1 – 34. IMS Hayward, California.

Adaptive P-splines via L_1-type penalty in generalized additive models

Daniele Cuntrera[1], Vito M.R. Muggeo[1]

[1] Dip.to Sc Econom, Az e Statistiche, Università di Palermo, ITALY.

E-mail for correspondence: vito.muggeo@unipa.it

Abstract: We propose a new adaptive penalty for smoothing via penalized splines. The new form of adaptive penalization is based on penalizing the differences of the coefficients of adjacent bases using penalties based on the L_1 norm. This makes possible to estimate curves with varying amounts of smoothness. Comparisons with respect to some competitors are presented.

Keywords: P-spline, adaptive smoothing, non-convex penalties, MCP, SCAD.

1 Introduction

The use of splines in statistical modelling (Eilers and Marx, 1996) has undergone a significant increase in real data analyses: for instance, in just five years, the number of R packages using splines has tripled (Perperoglou et al., 2019). However when the underlying relationship exhibits somewhat complex shapes, better results could be obtained by varying the amount of smoothness of the fitted curve, namely via *adaptive smoothing*. In this work we present a simple approach exploiting the properties of nonconvex penalties and present some comparisons via a simulation study.

2 Splines, B-splines and P-splines

Let $f(x)$ be the unknown but smooth function relating the continuous covariate x and the conditional expected value of the response $Y|x$ via the link function $g(\cdot)$, namely $E[Y_i|x_i] = f(x_i)$. The smooth function is expressed via a B-spline basis with specified degree and equally spaced knots,

$$f(x_i) = \sum_{k=1}^{K} b_k B_k(x_i) = B(x_i)^T b,$$

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

where the B_ks are the bases and b_ks the relevant coefficients. For the response values $\{y_i\}_{1,\dots,n}$, and assuming for the sake of simplicity a continuous response with identity link, the objective to be minimized is the well known least square $\sum_i (y_i - B(x_i)^T b)^2$. To bypass the issues related to the selection of the number and location of knots (affecting the basis dimension K), a penalty is added to to the fidelity term to be optimized, leading to the penalized objective $\sum_{i=1}^n (y_i - B(x_i)^T b)^2 + \lambda \sum_k (\Delta^d b_k)^2$, where λ is the tuning parameter determining how the curve has to be smoothed. The larger λ, the smoother the fitted curve, and the optimal value is usually selected by CV, AIC or BIC.
However, a unique and constant λ in the penalty implies that the amount of smoothing is fixed. Sometimes, such *constant* smoothing can lead to undesirable fits.

3 The proposal: P-spline using L_1 penalty

To allow adaptive smoothing we propose to penalize the coefficient differences $\Delta^d b_k$ via an '*unbiased*' nonconvex penalty, such as SCAD (Fan and Li, 2001) or MCP (Zhang, 2010). Unlike the naive lasso, such as $\sum_k |\Delta^d b_k|$, the '*unbiased*' penalties SCAD or MCP allow to alleviate bias in the non-null coefficients while keeping to zero the smallest (in absolute value) ones. By indicating with $p\left(|\Delta^d b_k|\right)$ a specified unbiased penalty (MCP or SCAD), the objective can be written

$$\sum_{i=1}^{n} (y_i - B(x_i)^T b)^2 + \lambda \sum_k p\left(|\Delta^d b_k|\right). \tag{1}$$

The L_1 penalty is not differentiable and therefore the usual Newton-like algorithms cannot be used. Very efficient algorithms do exist, but alternatively the *local quadratic approximation* (Fan and Li, 2001) allows to attain the final solution by optimizing iteratively the following objective

$$\sum_{i=1}^{n} (y_i - B(x_i)^T b)^2 + \lambda \sum_k w_k \left(\Delta^d b_k\right)^2, \tag{2}$$

namely a weighted ridge penalty with weights w_k depending on the penalty derivative evaluated at the previous solution $\tilde{b}$, i.e. $w_k = p'(|\Delta^d \tilde{b}_k|)$.
In addition to the well known SCAD and MCP, we also consider a new 'unbiased' nonconvex penalty, the so-called CDF to be discussed in a different paper presented in this workshop (Cuntrera et al., 2022).
Regardless of the penalty (SCAD, MCP or CDF), the optimal λ can be selected by CV, BIC/AIC or iteratively via the algorithm proposed by Schall (1991). Owing to weights w_k, even a unique λ value can lead to vary the amount of smoothing.

4 Simulation study

We contrast our proposal (using the non-convex SCAD, MCP, and CDF penalties in (2)) with the spatially-adaptive P-splines (SOP) of Rodríguez-Álvarezez et al. (2019), and the traditional P-splines with constant smoothing. We use the Doppler function as the true signal μ_i, that is a typical example in the literature on adaptive smoothing, and generate $y_i = \mu_i + \sigma\epsilon_i$ where $x_i = i/n$, $n = 400$, $\epsilon_i \sim N(0, 1)$, and different values of σ in (0.05, 0.25) to assess how performance varies as noise increases. For all settings, we carried out 300 trials. Performance is evaluated by means of the Mean Integrated Squared Error (MISE) measuring the difference between true μ_i and fitted $\hat{\mu}_i$. For all methods, we use a third-degree B-spline with 50 bases and penalty based on the 3rd order differences and λ selected via the Schall algorithm separately for each method. SOP was implemented via the R package `SOP` version 1.0.

As shown in Figure 1, the MISE curve of classical P-spline, i.e. not accounting for adaptive smoothing is the highest. Among the adaptive smoothing strategies, MCP/SCAD (indistinguishable results, red line) perform better than SOP (green line) when the signal-to-noise is lower. Interestingly, adaptive smoothing via the new CDF penalty (light blue) performs the best over all the σ values tested, expect for $\sigma = 0.05$ where the SCAD penalty performs quite slightly better. At higher values of σ, the CDF penalty and SOP tend to have the same values of MISE.

The fitted curves averaged across the 300 simulations are reported for the scenario with $\sigma = 0.15$ in the left panel of Figure 1 along with the true signal (black line). It is worth noting the curve obtained using constant smoothing shows evident undersmoothing problems that guarantee good fits only on the right side of the penalty. Curves fitted by using adaptive smoothing show a remarkable improvement: they improve the results on the left side, while maintaining a good fit on the right side. The CDF penalty seems closer to the true signal, especially in the first part.

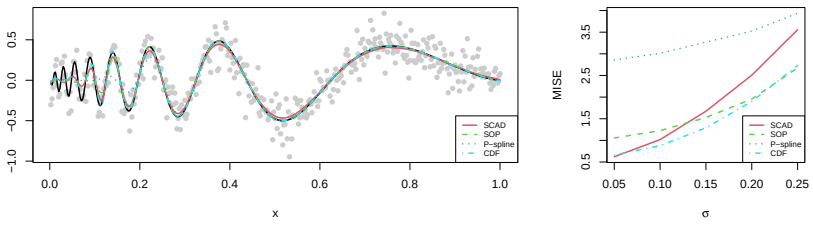


FIGURE 1. Simulation results. Left panel: True signal (black line), data (with $\sigma = 0.15$) and fitted curves (averaged across replicates). Right panel: MISE for different method and σ.

5 Conclusion

We have presented a new approach for dealing with adaptive P-splines for smoothing. Our proposal relies on the non-convex penalties, including the new CDF penalty, which favour adaptive smoothing even with a unique tuning parameter value. The simulation results show that the obtained results are very competitive, and sometimes even better than the alternative SOP method. Implementation in GLM is straightforward, but a possible more challenging task is extension to multidimensional smoothing via tensor product of bases.

References

Cuntrera, D., Muggeo, V. M. R., Augugliaro, L. (2022). Variable selection with quasi-unbiased estimation: the CDF penalty. *Proceedings of the 36th IWSM, Trieste, 18-22-july 2022*.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, **11(2)**, 89 – 121.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Statist Assoc*, **96**, 1348 – 1360.

Rodríguez-Álvarez, M. X., Durban, M., Lee, D-J. and Eilers, P. H. (2019). On the estimation of variance parameters in non-standard generalised linear mixed models: Application to penalised smoothing. *Statist and Comp*, **29**, 483 – 500.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719 – 727.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann of Statist*, **38**, 894 – 942.

Hawkes processes on networks for crime data

Nicoletta D'Angelo[1], David Payares[2], Giada Adelfio[1], Jorge Mateu[3]

[1] Department of Economics, Business and Statistics, University of Palermo, Italy
[2] Department of Earth Observation Science, University of Twente, Netherlands
[3] Department of Mathematics, Universitat Jaume I, Spain

E-mail for correspondence: nicoletta.dangelo@unipa.it

Abstract: Motivated by the analysis of crime data in Bucaramanga (Colombia), we propose a spatio-temporal Hawkes point process model adapted to events living on linear networks. We first consider a non-parametric modelling strategy, for both the background and the triggering components, and then we include a parametric estimation of the background based on covariates, and a non-parametric one of the triggering effects. Our network model outperforms a planar version, improving the fitting of the self-exciting point process model.

Keywords: Covariates; Crime data; Hawkes processes; Linear networks; Spatio-temporal point processes.

1 Introduction

Point processes are stochastic processes defining a natural and convenient formal tool to describe the process of discrete events that occur in a continuous space, time or a space-time domain; spanning many scientific branches, examples of application are forest fires, crimes, earthquakes, diseases, tree locations, animal locations or communication network failures.
A number of papers have dealt with the analysis of crime data using self-exciting point process theory. In particular, several papers have proposed a Hawkes-type point process modelling framework for crime data, as this type of data is usually clustered. As crime events are naturally constrained to occur on the streets structure of a city, in this paper, we advocate the use of the theory of spatio-temporal point processes on linear networks. In detail, we analyse robbery crimes occurred in the city of Bucaramanga (Colombia) in 2018.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

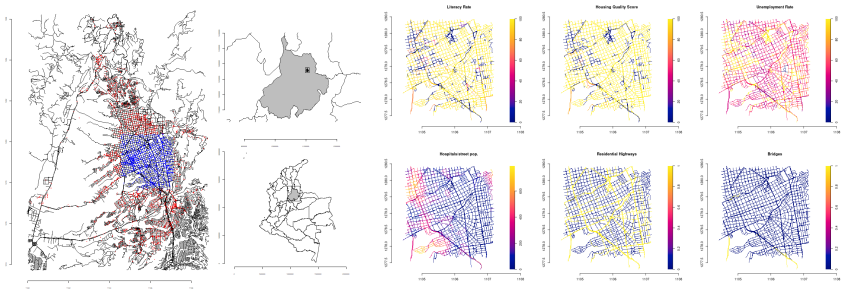


FIGURE 1. (a) Armed robberies (in red) in Bucaramanga, and in the city's downtown (in blue). In black, the segments of the streets of Bucaramanga city. (b) Some socio-economic, demographic and environmental spatial covariates used.

The proposed model includes external covariates in the purely spatial background component. We find that our proposed model, accounting also for some spatial covariates in the background component, fits better than the planar counterpart on the Euclidean space, allowing us to better interpret the results.
The paper is structured as follows. In Section 2 we introduce the proposed spatio-temporal Hawkes models. In Section 3 we present the data and carry out model fitting and diagnostics. Section 4 contains the conclusions.

2 A Hawkes model on linear networks with covariates

Point processes can be formally specified in several ways, for instance, by considering the joint distribution of the counts of points in arbitrary sets or by defining a complete intensity function. To model events that are clustered, self-exciting point processes are often used. Examples include Hawkes models (Hakwes, 1971). The conditional intensity function of a linear self-exciting process is defined as the sum of two non-negative functions: a background that describes the large-scale variation of the intensity, and a triggered component, which describes its small-scale variation due to the interaction with the events in the past.
We refer to a self-exciting model, following the semi-parametric specification proposed by Zhuang and Mateu (2019) for a spatio-temporal Hawkes process. Therefore, the first considered model (`Model 1`) is as follows

$$\lambda(t,x,y) = \mu_0\mu_t(t)\mu_w(t)\mu_b(x,y) + A\int_{-\infty}^{t-}\int\int_X g(t-s)h(x-u,y-v)N(\mathrm{d}u\times\mathrm{d}v\times\mathrm{d}s), \tag{1}$$

where we estimate the two relaxation coefficients A and μ_0, normalise to 1 the average values of $\mu_t(t)$, $\mu_w(t)$ and $\mu_b(x,y)$, and define the probability density functions g and h.

TABLE 1. Comparative results for the three fitted models

	$\hat{\mu}$	$\hat{A}$	$\log(L)$
`Model 1`	0.204	0.639	-706.47
`Model 2`	0.292	0.014	-353.35
`Model 3`	0.085	0.016	-319.70

Because of the nature of data, spatio-temporal point processes on linear networks are then referred to. Formally, a linear network $L = \cup_{i=1}^{n} l_i \subset \mathbb{R}^2$ is commonly taken as a finite union of line segments $l_i \subset \mathbb{R}^2$ of positive length. The distance between two locations in the network L is usually computed by the shortest-path distance d_L which is the minimum of the length of all possible paths between the two points. To fit (1) on the underlying spatial network, the main issue is to chose estimators for the spatial components $\hat{\mu}_b(x, y)$ and $\hat{h}(x-u, y-v)$, taking properly into account the underlying network structure. The second spatio-temporal Hawkes process model on the linear network L that we propose (`Model 2`) has the same specification of (1), but replacing the functions $\mu_b(\cdot, \cdot)$ and $h(\cdot, \cdot)$, by $\mu_L(\cdot, \cdot)$ and $h_L(\cdot, \cdot)$, respectively. These are computed using the 2D convolutional Gaussian Kernel of Rakshit et al. (2019), defined as $\hat{\lambda}(\mathbf{u}) = \frac{1}{c_L(\mathbf{u})} \sum_{i=1}^{n} \kappa(\mathbf{u} - \mathbf{x}_i), \quad \mathbf{u} \in L$, with κ a bivariate kernel function, that is, a probability density on $\mathbb{R}^2$.
`Model 2` can be extended by including external spatial covariates in $\mu_L(x, y)$. This last specification gives rise to `Model 3`, that is:

$$\lambda(t, x, y) = \mu_0 \mu_t(t) \mu_w(t) \mu_L(x, y, \boldsymbol{\beta}_{back}) + A \int_{-\infty}^{t_-} \int \int_{L} g(t-s) h_L(x-u, y-\nu) N(\mathrm{d}u \times \mathrm{d}\nu \times \mathrm{d}s), \tag{2}$$

where $\boldsymbol{\beta}_{back}$ denotes the parameters associated to the spatial covariates $Z(x, y)$ included in the model. As all the available covariates are continuous in space, they can be included linearly by basis functions.
Inference is carried out by the Estimation-Maximization (E-M) algorithm.

3 Model fitting to crime data and diagnostics

We analyse 2671 armed robberies in the city of Bucaramanga, Colombia, in 2018. We first fit `Model 1` and then our proposed extensions (`Model 2`) and (`Model 3`) to the data. Figure 1(a) displays the armed robberies for the entire city of Bucaramanga (red) and its downtown (blue) in 2018. In this study, we focus on analysing the latter subregion. We used all available 36 variables, including socio-economic factors, demographic aspects, environmental conditions, and geographical covariates. Figure 1(b) displays 6 of them. We estimated the relaxation coefficients $\hat{\mu}$ and $\hat{A}$ through a 40-loops iterative algorithm. We also computed the log-likelihood to assess the fit of the three space-time point process models. Table 1 reports the estimates of

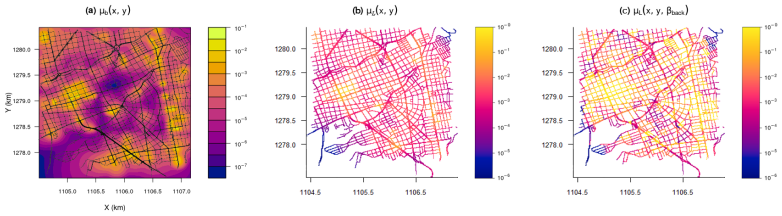


FIGURE 2. Spatial background rate: (a) `Model 1`, (b) `Model 2` and (c) `Model 3`.

the relaxation coefficients and the corresponding log-likelihood. $\hat{A}$ changes noticeably from specification (`Model 1`) to specifications (`Model 2`) and (`Model 3`): in the former, almost 64% of the crimes are triggered, while the latter models suggest that only 1.4% and 1.6% (respectively) of armed robberies are induced by previous crimes. The log-likelihood shows that our models fit better the armed robberies data. The spatial background rates of the fitted models come in Figure 2. Note that other diagnostic tools (e.g. transformed times and smoothed raw residuals) would further reinforce the model selection, but they are not shown here for reason of space.

4 Conclusions

In this paper, we analysed robbery crimes as events of a spatio-temporal point pattern living on a linear network structure. We first fitted a Euclidean planar model following and further proposed an extension of that model in order to take into account the linear network, through which we are also able to include the dependence on external covariates rightly constrained onto the spatial support of the road network, finding that our proposed models on the network achieve a much better fit when compared to the planar counterpart. Therefore, the current research paves the way for future developments in this promising direction, such as the inclusion of individual-related covariates into the triggering component.

References

Hawkes, A. G. (1971). *Introduction to the Theory of Point Processes. Volume II: General Theory and Structure,* Springer-Verlag, New York

Rakshit, S., Davies, T., Moradi, M. M., et al. (2019). Fast kernel smoothing of point patterns on a large network using two-dimensional convolution *International Statistical Review,* **87**(3), 531–556.

Zhuang, J. and Mateu, J. (2019). A semiparametric spatiotemporal hawkes-type point process model with periodic background for crime data *Journal of the Royal Statistical Society: Series A*. **182**(3), 919–942.

Bernoulli-exponential semiparametric regression model

Willian L. de Oliveira[1], María Durbán[2], Carlos A.R. Diniz[3]

[1] Department of Statistics, State University of Maringá, Brazil
[2] Department of Statistics, Carlos III University of Madrid, Spain
[3] Department of Statistics, Federal University of São Carlos, Brazil

E-mail for correspondence: wloliveira@uem.br

Abstract: A bivariate model for discrete and continuous responses is proposed in which joint distribution is constructed via the conditional approach. It is assumed that the discrete response follows the Bernoulli distribution and the continuous response, given the discrete outcome, follows the exponential distribution. Furthermore, the marginal means are related to the covariates by link functions using parametric and/or nonparametric predictors and a dependency structure between the responses is inserted into the model via the conditional mean. Estimation methods using P-Splines, diagnostic analysis and a simulation study are presented. Finally, this model is used in a real data set.

Keywords: Bivariate regression models; Conditional approach; P-Spline.

1 Introduction

It is common situations in which two responses associated with the same individual are simultaneously observed. In these cases the intrinsic relationship between the two variables should be considered in the analysis. In this context, bivariate models, which allocate a dependent structure between the two responses, should be taken into account.
Bivariate distributions can be built using different methods, including the mixing method, compounding method and via copula. Another direct way to built bivariate distributions is using the conditional approach, where the joint probability density function (pdf) is given by the product of a marginal pdf and a conditional pdf. In this approach, the variable associated with the marginal fdp is seen as the primary response while the conditioned variable is considered as the intermediate variable.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

Models for bivariate discrete and continuous outcomes, constructed by the conditional approach, are discussed in Fitzmaurice and Laird (1995), Yang et al. (2007) and more recently in de Oliveira et al. (2019), which proposed a general class of models for discrete and/or continuous responses, fitting the Bernoulli-exponential parametric model, a particular case of the class, to real data set. In this case, the predictor is formed only by linear terms, which is not always the most appropriate. Thus, in this paper, the Bernoulli-exponential model is extended, allowing parametric and/or nonparametric predictors. In this semiparametric context, the P-Spline technique is used, adapting to the bivariate context.

2 Bernoulli-exponential semiparametric model

It is assumed that $Y_i \sim \text{Bernoulli}(\mu_{iY})$ and $X_i|Y_i = y_i \sim \text{Exponencial}(\frac{1}{\alpha_i})$, where the dependence between the response variables is determined by the conditional mean $\alpha_i = E(X_i \mid Y_i = y_i)$. This mean is related to μ_{iX} (marginal mean of X_i), to Y_i (discrete response), to μ_{iY} (marginal mean of Y_i) and to γ (parameter included in the model, which may be a measure of association between the response variables), by a linear or nonlinear function, that is, $\alpha_i = h(y_i, \mu_{iY}, \mu_{iX}, \gamma)$. Further, covariates are available and are related to the marginal means by $g_1(\mu_{iY}) = \eta_{iY}$ and $g_2(\mu_{iX}) = \eta_{iX}$, where $g_1(\cdot)$ and $g_2(\cdot)$ are monotonic differentiable link functions,

$$\begin{aligned}
\eta_{iY} &= \beta_0 + \beta_1 z_{i1} + \cdots + \beta_{k-1} z_{ik-1} + s_1(z_{ik}) + s_2(z_{ik+1}) + \ldots \\
&\quad + s_{p-k}(z_{ip}) \text{ and} \\
\eta_{iX} &= \delta_0 + \delta_1 t_{i1} + \cdots + \delta_{d-1} t_{id-1} + u_1(t_{id}) + u_2(t_{id+1}) + \ldots \\
&\quad + u_{q-d}(t_{iq}),
\end{aligned}$$

with $z_{i1}, z_{i2}, \ldots, z_{ip}, t_{i1}, t_{i2}, \ldots, t_{iq}$, a set of covariates, $i = 1, 2, \ldots, n$; β_0, $\beta_1, \ldots, \beta_{k-1}, \delta_0, \delta_1, \ldots, \delta_{d-1}$ the unknown parameters and $s_1(.), s_2(.), s_3(.)$, $\ldots, s_{p-k}(.), u_1(.), u_2(.), \ldots, s_{q-d}(.)$ smooth functions.

2.1 Estimation method

When the responses are independent, each marginal model can be seen as a generalized additive model (GAM) and the P-Spline smoothers can be used to fit all smooth components simultaneously (Marx and Eilers, 1998). In this case, the GAM estimation is reduced to generalized linear regression context with a penalized version of the log-likelihood that attaching a penalty on B-Spline coefficients. Then, we extend the use of the P-Spline technique to the Bernoulli-exponential bivariate model in which both predictors are given in the context of generalized additive models.

Each smooth function is represented as a linear regression, using B-Spline bases. Without loss of generality, assuming that $\mathbf{B_Y}$ and $\mathbf{B_X}$ are non-singular bases and $\mathbf{a_Y}$ and $\mathbf{a_X}$ the vectors of regression coefficients associated with $\mathbf{B_Y}$ and $\mathbf{B_X}$, we have

$$\begin{aligned}
\eta_{\mathbf{Y}} &= \mathbf{1}\beta_0 + \mathbf{z_1}\beta_1 + \cdots + \mathbf{z_{k-1}}\beta_{k-1} + \mathbf{B_{1Y}a_{1Y}} + \cdots + \mathbf{B_{p-kY}a_{p-kY}} \\
&= \mathbf{B_Y a_Y}, \text{ and} \\
\eta_{\mathbf{X}} &= \mathbf{1}\delta_0 + \mathbf{t_1}\delta_1 + \cdots + \mathbf{t_{d-1}}\delta_{d-1} + \mathbf{B_{1X}a_{1X}} + \cdots + \mathbf{B_{q-dX}a_{q-dX}} \\
&= \mathbf{B_X a_X},
\end{aligned}$$

with $\eta_{\mathbf{Y}} = (\eta_{1Y}, \eta_{2Y}, \ldots, \eta_{nY})^\top$, $\eta_{\mathbf{X}} = (\eta_{1X}, \eta_{2X}, \ldots, \eta_{nX})^\top$, $\mathbf{B_Y} = (\mathbf{1} : \mathbf{z_1} : \mathbf{z_2} : \cdots : \mathbf{z_{k-1}} : \mathbf{B_{1Y}} : \mathbf{B_{2Y}} : \cdots : \mathbf{B_{p-kY}})$, $\mathbf{B_X} = (\mathbf{1} : \mathbf{t_1} : \mathbf{t_2} : \cdots : \mathbf{t_{d-1}} : \mathbf{B_{1X}} : \mathbf{B_{2X}} : \cdots : \mathbf{B_{q-dX}})$, $\mathbf{a_Y} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1}, \mathbf{a_{1Y}}^\top, \mathbf{a_{2Y}}^\top, \ldots, \mathbf{a_{p-kY}}^\top)^\top$ and $\mathbf{a_X} = (\delta_0, \delta_1, \delta_2, \ldots, \delta_{d-1}, \mathbf{a_{1X}}^\top, \mathbf{a_{2X}}^\top, \ldots, \mathbf{a_{q-dX}}^\top)^\top$. The penalized log-likelihood function of $\mathbf{a_Y}, \mathbf{a_X}$ and γ, may be written as

$$\begin{aligned}
\ell^*(\mathbf{a_Y}, \mathbf{a_X}, \gamma \mid \mathbf{x}, \mathbf{y}, \mathbf{Z}, \mathbf{T}) &= -\sum_{i=1}^{n} \log\alpha_i - \sum_{i=1}^{n} \frac{x_i}{\alpha_i} + \sum_{i=1}^{n} y_i \log\mu_{iY} \\
&\quad + \sum_{i=1}^{n} (1 - y_i)\log(1 - \mu_{iY}) \\
&\quad - \frac{1}{2}\mathbf{a_Y}^\top \mathbf{P_Y a_Y} - \frac{1}{2}\mathbf{a_X}^\top \mathbf{P_X a_X}, \qquad (1)
\end{aligned}$$

with $\mu_{\mathbf{Y}} = g_1^{-1}(\mathbf{B_Y a_Y})$ and $\mu_{\mathbf{X}} = g_2^{-1}(\mathbf{B_X a_X})$. The smoothing parameters $\lambda_{1Y}, \ldots, \lambda_{p-kY}$, $\lambda_{1X}, \ldots, \lambda_{q-dX}$ are chosen using generalized cross-validation (GCV) and are inserted in the penalty matrices

$$\begin{aligned}
\mathbf{P_Y} &= \text{block diagonal}(0, \ldots, 0, \lambda_{1Y}\mathbf{P_{1Y}}, \ldots, \lambda_{p-kY}\mathbf{P_{p-kY}}) \text{ and} \\
\mathbf{P_X} &= \text{block diagonal}(0, \ldots, 0, \lambda_{1X}\mathbf{P_{1X}}, \ldots, \lambda_{q-dX}\mathbf{P_{q-dX}}).
\end{aligned}$$

The maximum likelihood estimation is obtained by maximizing the Equation (1). A diagnostic analysis is presented considering the randomized quantile residual individually in each of the marginal models.

3 Simulation study

A simulation study is conducted in order to examine and compare the performance of three models:

- **M1**: Bernoulli-exponential model that assume independence between responses, that is, marginal models;

- **M2**: Bernoulli-exponential model with linear predictors;
- **M3**: Bernoulli-exponential model with semiparametric predictors.

We simulate data with samples sizes $n = 100, 200$ and 400 and for all n values, we generate $N = 100$ samples of n observations. The adopted predictors are given by

$$\eta_{iY} = 3.2 + 1.3 \times \text{Stay}_i + \sqrt{\text{Age}_i} \text{ and}$$
$$\eta_{iX} = 6.7 + 3 \times \text{Age}_i + 2 \times cos(\text{Stay}_i),$$

where Stay and Age are covariates included in the real dataset, described in the next section. We adopt a dependency structure

$$\alpha_i = \frac{0.4^{y_i}\mu_{iX}}{(1 + \mu_{iY}(-0.6))}.$$

The criterion to assess the efficiency of the P-Spline is the Mean Square Error (MSE). Our simulation results revealed that the **M3** is more flexible in fitting data with dependent responses and nonlinear predictors.

4 Application

A real data set containing information related to admissions of patients provided by managed care plans is analyzed by using the three Bernoulli-exponential models (**M1**, **M2** and **M3**). The data set is composed of information on 308 admissions. The *total cost of care* for each patient during hospitalization and the *use or not of the intensive care unit* are adopted as continuous and discrete response variables, respectively. The set of covariates includes *length of stay* (in days), *age*, *patient's status* upon arrival at the hospital and *requested medical specialty*, according to the patient's problem.
Patient's status is categorized as "Not Severe" and "Severe" and requested medical specialty is categorized as "CRD-S" (Circulatory, Respiratory, or Digestive system), "P" (pregnancy, childbirth, or genital organs), "Tumor" (Cancer) and "Other".
In Figure 1 are shown qqplots with envelope, considering the quantile residuals, for each of the marginal models, in **M1**, **M2** and **M3** models.

5 Conclusions

The proposed Bernoulli-exponential semiparametric model is more flexible since it encompasses a variety of different predictors. Using a simulation study in some predetermined scenarios we can note that the fitted model is satisfactory, especially for situations where the predictors are not linear.

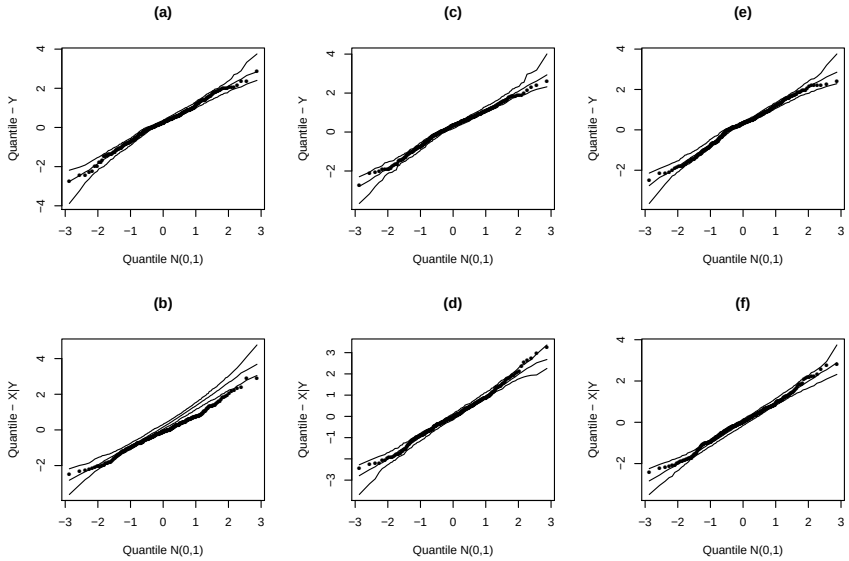


FIGURE 1. qqplots with envelope considering the quantile residuals. (a), (c), and (e): **M1**, **M2** and **M3** models, respectively - Y_i; (b), (d), and (f): **M1**, **M2** and **M3** models, respectively - $X_i \mid Yi = yi$.

References

de Oliveira, W. L., Diniz, C. A. R., and Durbán, M. (2019). A class of bivariate regression models for discrete and/or continuous responses. *Communications in Statistics-Simulation and Computation*, **58**, 2359 – 2383.

Fitzmaurice, G.M. and Laird, N.M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, **90**, 845 – 852.

Marx, B. D. and Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, **28**, 193 – 209.

Yang, Y., Kang, J., Mao,K. and Zhang, J. (2007). Regression models for mixed poisson and continuous longitudinal data. *Statistics in Medicine*, **26**, 3782 – 3800.

Sparse composite likelihood selection

Claudia Di Caterina[1], Davide Ferrari[1]

[1] Faculty of Economics and Management, University of Bozen-Bolzano, Italy

E-mail for correspondence: claudia.dicaterina@univr.it

Abstract: Composite likelihood has shown promise in settings with large number of parameters p due to its ability to break down complex models into simpler components, even when the full likelihood is intractable. However, there does not seem to exist agreement on how to construct composite functions that are computationally efficient and statistically sound when p is allowed to diverge. We present a flexible method to select sparse composite likelihoods via a criterion representing the statistical efficiency of the implied estimator and an L_1-penalty discouraging the inclusion of too many sub-likelihood terms. The theoretical properties of the proposed procedure are illustrated through simulation studies.

Keywords: Pseudo-likelihood inference; High-dimensional parameter; Sparsity-inducing penalization.

1 Introduction

Let Y be a $d \times 1$ random vector with distribution $f(y;\theta)$ indexed by the parameter $\theta \in \Theta \subseteq \Re^p$. Suppose the full d-variate $f(y;\theta)$ is hard to specify or compute but we can identify p distributions $f_j(y;\theta)$ $(j = 1,\ldots,p)$ defined on low-dimensional subsets of Y, such as marginals Y_j, pairs (Y_j, Y_k), or conditionals $Y_j|Y_k = y_k$ $(j \neq k)$. Given independent observations $Y^{(1)},\ldots,Y^{(n)}$, the composite likelihood (CL) estimator maximizes the corresponding log-likelihood function (Besag, 1975)

$$\ell(\theta; Y^{(1)},\ldots,Y^{(n)}) = \sum_{j=1}^{p} \ell_j(\theta; Y^{(1)},\ldots,Y^{(n)}), \tag{1}$$

where $\ell_j(\theta; Y^{(1)},\ldots,Y^{(n)}) = \sum_{i=1}^{n} \log f_j(Y^{(i)};\theta)$ is the sub-likelihood for the jth data subset. It is well known that the CL estimator has the same first-order properties as maximum likelihood (ML) (Varin et al., 2011).

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

Although the CL framework naturally suits problems where the parameter dimension p can diverge with the sample size n, how to select the sub-likelihood terms forming (1) in this setting (Lindsay et al., 2011) remains unclear. Such a selection is crucial since it determines both statistical properties and computing cost of the CL estimator (Cox and Reid, 2004; Lindsay et al., 2011; Huang et al., 2020); it is also related to model selection, with the two tasks coinciding when each sub-likelihood contains a distinct element of θ. Without any form of selection, the accuracy of CL estimators is found to deteriorate as the data dimension grows when the low-dimensional data subsets are sufficiently correlated (Cox and Reid, 2004).

2 Selecting sparse composite likelihood functions

Let $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ be sparse, i.e. with many zero elements, and p be allowed to grow with the sample size n. The marginal scores are defined by $u_j(\theta_j; y) = \partial \log f_j(y; \theta_j)/\partial\theta_j$ $(j = 1, \ldots, p)$, and $u(\theta; y) = \{u_1(\theta_1; y), \ldots, u_p(\theta_p; y)\}^{\mathrm{T}}$. We assume here that each sub-likelihood depends only on one single θ_j, but the approach is also valid if $\ell_j(\theta)$ depends on a finite number of parameter components. In order to reduce the model dimension by dropping all the zero elements of θ while estimating the rest, we use the sparse CL estimator $\hat{\theta}$ with jth component $\hat{\theta}_j = \tilde{\theta}_j I(\hat{w}_j \neq 0)$, where $\tilde{\theta}_j$ is the marginal estimator

$$\tilde{\theta}_j = \left\{ \theta_j : 0 = \sum_{i=1}^{n} u_j(\theta_j; Y^{(i)}) \right\} \quad (j = 1, \ldots, p)\,,$$

and $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_p)^{\mathrm{T}}$ is obtained by minimizing the penalized objective

$$\hat{d}_\lambda(w) = \frac{1}{2} w^{\mathrm{T}} \hat{C} w - w^{\mathrm{T}} \mathrm{diag}(\hat{C}) + \frac{\lambda}{n} \sum_{j=1}^{p} \frac{|w_j|}{\tilde{\theta}_j^2}, \tag{2}$$

for some constant $\lambda \geq 0$. Here $\mathrm{diag}(\hat{C})$ denotes the diagonal vector of any consistent estimator $\hat{C}$ of the $p \times p$ score covariance matrix $C(\theta) = \mathrm{var}\{u(\theta; Y)\} = E\{u(\theta; Y)u(\theta; Y)^{\mathrm{T}}\}$, for instance its empirical counterpart $\hat{C} = \sum_{i=1}^{n} u(\tilde{\theta}; Y^{(i)})u(\tilde{\theta}; Y^{(i)})^{\mathrm{T}}/n$.

Sparse sub-likelihood selection occurs through the minimization of the convex objective (2): the jth sub-likelihood $\ell_j(\theta)$ is included in the CL function if $\hat{w}_j \neq 0$, else $\ell_j(\theta)$ is dropped and the corresponding parameter estimate is set as $\hat{\theta}_j = 0$ $(j = 1, \ldots, p)$. The selected CL function is interpreted as one that maximizes statistical efficiency given a desired level of sparsity. Indeed, when $\lambda = 0$ the objective $\hat{d}_0(w)$ equals the finite-sample optimality criterion to find minimum variance estimators for unbiased estimating equations (Heyde, 2008, Ch. 2). The last term in (2) is a sparsity-inducing penalty discouraging overly complicated CL functions. The geometric properties of the L_1-penalty imply that several elements in $\hat{w}$ are exactly zero

TABLE 1. Estimated number of selected parameters $\hat{p}^*$, true positive probability (TPP%), true negative probability (TNP%) and false discovery probability (FDP%) for the coefficients of the p-variate probit regression $Y_j = I(Z_j \geq 0)$, given $Z \sim N_p(\mu, \Sigma)$ where $\mu_j = 0.1 + \theta_j x$ with $p = 100$ and $p^* = 25$ nonzero θ_js.

	$\{\Sigma\}_{jk} = 0$				$\{\Sigma\}_{jk} = 0.5$			
λ	$\hat{p}^*$	TPP%	TNP%	FDP%	$\hat{p}^*$	TPP%	TNP%	FDP%
1.170	44.482	99.4	73.8	43.7	26.484	92.4	95.5	11.4
2.107	35.487	99.4	85.8	29.5	22.766	86.6	98.5	4.4
3.796	28.839	99.3	94.6	13.5	19.956	78.9	99.7	0.9
6.840	25.565	99.0	98.9	3.0	17.354	69.2	99.9	0.1
12.323	24.597	98.1	99.9	0.2	15.117	60.4	100.0	0.0

for sufficiently large values of λ, which also induces sparsity in the estimator $\hat{\theta}$. The adaptive penalty (Zou, 2006) ensures consistent model selection as sub-likelihoods associated with $\tilde{\theta}_j$s closer to zero are more penalized. Yet the fact that this penalty acts on the coefficients w_js rather than on the parameters θ_js enables to separate the task of model selection from that of estimation. So, differently from usual penalized CL procedures (Xue et al., 2012; Gao and Carroll, 2017), the selected estimating equations stay unbiased and lead to consistent estimators conditionally on correct selection.

3 Simulation studies

The consistency of our strategy can be illustrated via Monte Carlo experiments. Due to space constraints, we report here partial results for sparse location estimation based on 2500 simulated samples of size $n = 250$.
Setting 1: p-variate normal model $Y \sim N_p(\theta, \Sigma)$ with $p = 100$. The mean vector θ has $p^* = 25$ nonzero elements and Σ is such that $\{\Sigma\}_{jj} = 1$ for all j and $\{\Sigma\}_{jk} \in \{0, 0.5\}$ $(j \neq k)$. *Setting 2:* p-variate probit regression with $p = 100$. The jth binary response $Y_j = I(Z_j \geq 0)$ $(j = 1, \ldots, p)$ is generated based on $Z \sim N_p(\mu, \Sigma)$ where $\mu_j = 0.1 + \theta_j x$, with nonzero $p^* = 25$ probit coefficients, Σ as in Setting 1, and x normal covariate.
Estimates of the true positive probability $\text{TPP} = \#\{j : \hat{\theta}_j \neq 0, \theta_j \neq 0\}/p^*$, true negative probability $\text{TNP} = \#\{j : \hat{\theta}_j = 0, \theta_j = 0\}/(p - p^*)$ and false discovery probability $\text{FDP} = \#\{j : \hat{\theta}_j \neq 0, \theta_j = 0\}/p^*$ are given in Table 1 for Setting 2, along with the average estimated parameters $\hat{p}^*$ at five values of λ. The sparse combination of CL scores exhibits remarkable model-selection properties in terms of type I error (FDP) and power (TPP), even if there exists correlation among scores. For varying $\hat{p}^*$, Figure 1 shows the efficiency of the sparse CL estimator compared to that of the unattainable oracle ML which estimates the nonzero parameters only and sets to zero the others. With uncorrelated scores the relative efficiency peaks at the true

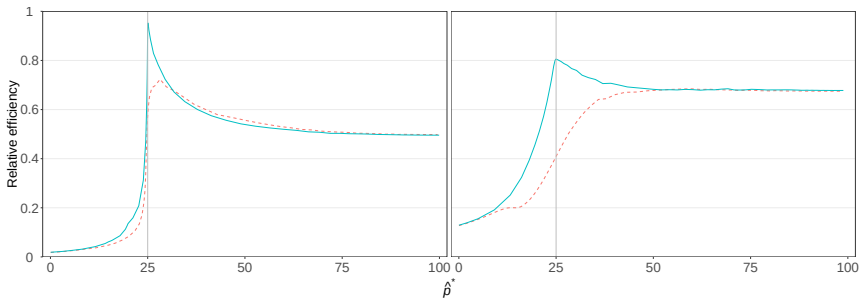


FIGURE 1. Estimated relative efficiency of the sparse CL estimator versus the average selected parameters $\hat{p}^*$. The trajectories for Setting 1 (left panel) and Setting 2 (right panel) correspond to uncorrelated (solid) and correlated (dashed) scores. The vertical line at $p^* = 25$ marks the true number of nonzero parameters.

$p^* = 25$, very close to one in Setting 1. If all $p = 100$ scores are correlated, as expected, estimation accuracy is hindered by a less reliable selection; yet efficiency remains high, with maximum reached after p^* in both settings.

References

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society, Series D*, **24**, 179 – 195.

Cox, D.R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, **91**, 729 – 737.

Gao, X. and Carroll, R.J. (2017). Data integration with high dimensionality. *Biometrika*, **104**, 251 – 272.

Heyde, C.C. (2008). *Quasi-likelihood And Its Application: A General Approach to Optimal Parameter Estimation*. Springer.

Huang, J., Ning, Y., Reid, N., and Chen, Y. (2020). On specification tests for composite likelihood inference. *Biometrika*, **107**, 907 – 917.

Lindsay, B.G., Yi, G. Y., and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, **21**, 71 – 105.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5 – 42.

Xue, L., Zou, H., Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse Ising models. *The Annals of Statistics*, **40**, 1403 – 1429.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of American Statistical Association*, **101**, 1418 – 1429.

Quirks with joint hierarchical models: Examples involving global relationships between sodium and potassium intake, GDP and healthy life expectancy

John Ferguson[1], Alberto Alvarez[1], Catriona Reddin[1], Robert Murphy[1], Martin O'Donnell[1]

[1] HRB Clinical Research Facility, NUI Galway, Galway Ireland

E-mail for correspondence: john.ferguson@nuigalway.ie

Abstract: We describe a joint hierarchical Bayesian model for regional and country-level per-capita potassium and sodium intake. The model is fit to data collected from 137 separate studies pertaining to 52 countries. Extending this joint Bayesian model to include a second level regression for country-level healthy life expectancy strongly affected some country and region level estimates for sodium and potassium intake. Accounting for spatial correlation in healthy life expectancy lessened this effect somewhat. These results raise the question of whether including such second level regressions (as a form of indirect evidence) is advised in such analyses when meta-analysis of country-level means or prevalences is of primary interest

Keywords: Hierarchical models; Joint models; STAN; Measurement error

1 Introduction

The nature of the causal relationships between sodium intake, potassium intake and downstream cardiovascular health are not completely determined. Recently, Messerli et al. (2021) found a significant, increasing relationship between country level sodium consumption and healthy life expectancy, adjusting for GDP and BMI. In this paper, we describe an alternative Bayesian joint-meta analysis model for potassium and sodium intake. This model was fit using data collected via literature review at NUI Galway, and encompassed 137 separate studies over 52 countries within 17 geographic

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

regions, selected based on population estimates for potassium intake. Information on sodium intake and estimates stratified by gender were also recorded when available, and were model imputed when not. Our primary hierarchical Bayesian model (see Section 2.1) focused on estimating global, region-level and country-level per capita intake for sodium and potassium. In a secondary analysis we extended this model to include a regression of healthy life expectancy on sodium, potassium and GDP. Here we contrast the country level sodium and potassium estimates from both approaches (excluding and including the second level regression). We also compare the Bayesian estimates for the relationships between sodium, potassium and healthy life expectancy under differing specifications for the joint model.

2 Models

2.1 Hierarchical model for per-capita sodium and potassium intake

Estimated sample means, and standard errors, of sodium and potassium were observed for N=198 observations, 126 of which were specific to males or females and 72 of which were not gender stratified. Note that there were only $S = 137$ studies, each study pertaining to a single country, with some studies contributing two observations if separate male and female estimates were given. We write this raw data as:

$$\begin{aligned} \hat{\theta_i^P} &\sim N(\theta_i^P, (SE_i^P)^2) \\ \hat{\theta_i^S} &\sim N(\theta_i^S, (SE_i^S)^2) \end{aligned} \tag{1}$$

for $i \in \{1 \dots 198\}$, θ_i^P and θ_i^S being the true potassium and sodium means for the sub-populations from which $\hat{\theta_i^P}$ and $\hat{\theta_i^S}$ were sampled. Standard errors SE_i^P and SE_i^S are assumed known. While sodium estimates: $\hat{\theta_i^S}$ were missing for 59 observations, means θ_i^S for these observations were naturally imputed in fitting the model. The sub-population means (θ_i^P, θ_i^S) depend both on the proportion of males in the subpopulation, P_i, and a binary indicator for whether potassium and sodium intakes were assessed using urinary measurements $D_i = 0$, or via dietary surveys, $D_i = 1$. Both these factors are assumed to have multiplicative effects on the measured outcome, denoted for potassium as $(1+\beta_G^P P_i)$ and $(1+\beta_D^P D_i)$, with similar notation for sodium:

$$\begin{aligned} \theta_i^P &= [\theta_{s(i)}^{P,study}(1 + \beta_G^P P_i) + \beta_{s(i),G}^P P_i](1 + \beta_D^P D_i) \\ \theta_i^S &= [\theta_{s(i)}^{S,study}(1 + \beta_G^S P_i) + \beta_{s(i),G}^S P_i](1 + \beta_D^S D_i) \end{aligned} \tag{2}$$

$\theta^{P,study}_{s(i)}$ and $\theta^{S,study}_{s(i)}$ are the study-population means for potassium and sodium in study $s(i) \in \{1 \dots 137\}$, assuming measurements were made using urinary assays and on a female-only population. However, note that equation (2) allows us to convert between expected study-level potassium and sodium levels if sodium and potassium were measured using dietary surveys as opposed to urinary assays, and for various mixtures of males and females. $\beta^{S}_{s(i),G}$ and $\beta^{P}_{s(i),G}$ are study-level adjustments to the overall multiplicative gender effect: $(1 + \beta^{P}_{G} P_i))$, with the respective adjustments for sodium and potassium having prior distributions: $\beta^{S}_{s(i),G} \sim N(0, \sigma^2_G)$ and $\beta^{P}_{s(i),G} \sim N(0, \tau^2_G)$. Potassium means are assumed to be linked using the following 3 level hierarchy, according to study within country, country within region and region, with independent sampling at each hierarchical level, conditional on hyper-parameters. Here, $c(s) \in \{1, ..., 52\}$ represents the country indicator for study s, and $r(c) \in \{1, ..., 15\}$ the region indicator for country c.

$$\begin{aligned} \theta^{P,study}_{s} &\sim N(\theta^{P,country}_{c(s)}, \tau^2_{study}) \\ \theta^{P,country}_{c} &\sim N(\theta^{P,region}_{r(c)}, \tau^2_{country}) \\ \theta^{P,region}_{r} &\sim N(\theta^{P}, \tau^2_{region}) \end{aligned} \tag{3}$$

Within a country, it is assumed that study level sodium and potassium mean values are linked via a linear model:

$$\theta^{S,study}_{s} \sim N(\theta^{country}_{c(s)} + \beta^{country}_{c(s)}(\theta^{P,study}_{s} - \theta^{P}_{c(s)}), \sigma^2_{study}) \tag{4}$$

with the country-level $\theta^{S,country}_{c}$ and region level $\theta^{S,region}_{r}$ means for sodium intake, and also country-level and region level sodium/potassium slopes $\beta^{country}_{c}$ and β^{region}_{r} being subject to a similar hierarchical structure:

$$\begin{aligned} \theta^{S,country}_{c} &\sim N(\theta^{S,region}_{r(c)}, \sigma^2_{country}) \\ \theta^{S,region}_{r} &\sim N(\theta^{S}, \sigma^2_{region}) \\ \beta^{country}_{c} &\sim N(\beta^{region}_{r}, \sigma^2_{\beta,country}) \\ \beta^{region}_{r} &\sim N(\beta, \sigma^2_{\beta,region}) \end{aligned} \tag{5}$$

with associated variance parameters at each level.
Global potassium and sodium means were estimated by a weighted average of the estimated region level means: $(\theta^{P,region}_{r}, \theta^{S,region}_{r})$ using population totals within each region as weights.

2.2 Joint model, including regression for Healthy Life Expectancy

In a secondary analysis the model described above was extended to include a regression for healthy life expectancy, Y_c, at the country level, with adjustment for country-level per capita GDP, G_c in 2019. Conditional on region-level effects for healthy life expectancy: ϵ_r, $r \in \{1, ..., 17\}$, we assume:

$$Y_c \sim N(\mu_c, \sigma^2_{HLE}) \tag{6}$$

where

$$\mu_c = \beta_0 + \beta_1 ln(G_c) + \beta_2 ln(G_c)^2 + \beta_3 \theta_c^{P,country} + \beta_4 \theta_c^{S,country} + \epsilon_{r(c)} \tag{7}$$

and

$$\epsilon_r \sim N(0, \phi^2)$$

The assumed non-linear relationship between Y_c and G_c was selected both on the basis of plots and theoretical considerations, and provided more successful adjustment for GDP confounding compared to the linear adjustment assumed by Messerli et al. The variance-parameter $\phi^2 > 0$ accounts for within-region clustering of values Y_c. Model fitting for the Hierarchical model described in Section 2.1 and the extended joint models described in Section 2.2 was performed via Hamiltonian Monte Carlo using RStan using 10,000 iterations (the first 5,000 as warm-up) of 8 parallel chains using semi-informative prior distributions for hyper-parameters.

3 Results and Conclusions

Sodium intake is known to be relatively high in Central Europe (Poland, Hungary and the Czech Republic). Using the model described in Section 2.1, the posterior-mean for estimated per capita sodium intake was 4.5 grams/day, with 95% credibility interval from (3.6g, 5.4g), see Table 1 and Figure 1; above the Global estimate of 3.8g/day (3.47g-4.15g). The estimate was lower: 4.2g/day (3.5, 5.2), when extending the model as described in Section 2.2. Informally, the model (effectively now a joint model for sodium, potassium and healthy life expectancy) pulls down the sodium estimates in Central Europe to amplify the correlation between sodium and healthy life expectancy. This effect is more pronounced if clustering is not allowed for in the second level regression ($\phi = 0$, ln(GDP), ln(GDP)^2), where the estimates fall to: 3.6g (3.2g-4.1g). These results raise the question of whether adding in such indirect information is sensible if the focus of the analysis is to estimate per-capita sodium and potassium intake.

TABLE 1. Comparison of estimates of potassium (β_3) and sodium (β_4) effects on healthy life expectancy and estimates for sodium intake in Central Europe ($\theta_{\text{CE}}^{S,region}$), under 4 differing model specifications:

	Hierarchical Model	Joint Model ($\phi > 0$, ln(GDP), $\ln(\text{GDP})^2$)	Joint Model ($\phi = 0$, ln(GDP), $\ln(\text{GDP})^2$)	Joint Model ($\phi = 0$, linear-GDP)
β_3	NA	2.7 (-2.2,7.9)	4.5 (0.8,9.0)	8.3 (4.2,13.0)
β_4	NA	3.8 (-1.3,9.6)	7.8 (3.8,14.3)	9.9 (5.9,14.6)
$\theta_{\text{CE}}^{S,region}$	4.5 (3.6, 5.4)	4.2 (3.5,5.2)	3.6 (3.2,4.1)	3.7 (3.3,4.2)

Appropriate inference regarding possible relationships between sodium, potassium and healthy life expectancy conditional on GDP requires consideration of clustering of life expectancy among neighbouring countries (Joint Model, $\phi > 0$, ln(GDP), $\ln(\text{GDP})^2$) and is indefinitive regarding whether such effects exist. Failing to take into account clustering ($\phi = 0$, ln(GDP), $\ln(\text{GDP})^2$), and inadequate adjustment for confounding by GDP ($\phi = 0$, linear-GDP) lead to stronger but possibly erroneous estimated effects.

References

Messerli, F. H (2021). Sodium intake, life expectancy, and all-cause mortality. *European heart journal*, **42**, 2013 – 2112

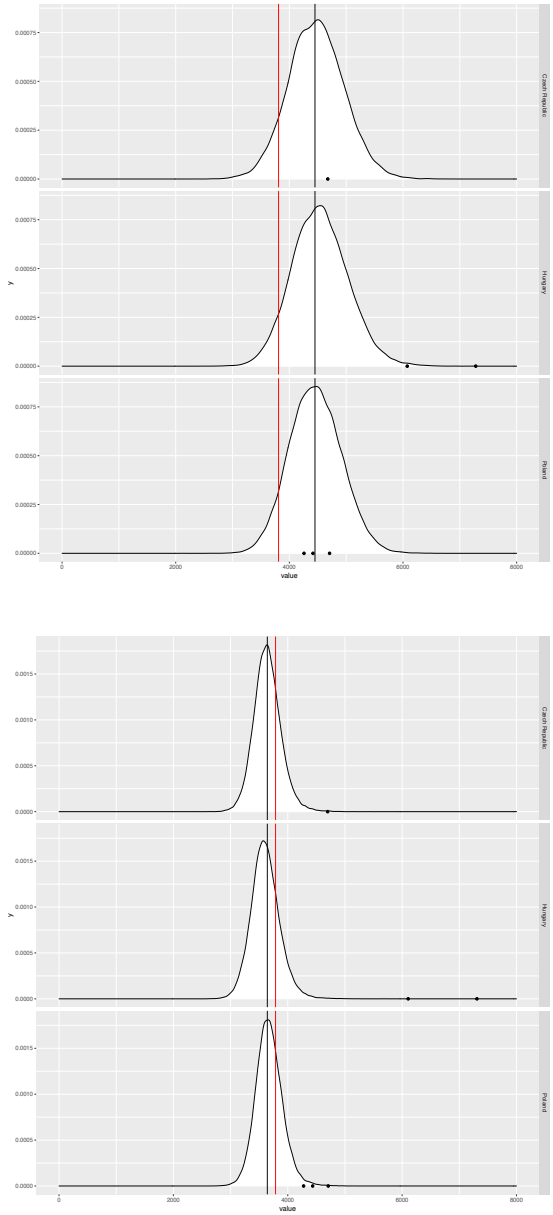


FIGURE 1. The top three plots show estimated posterior distributions for sodium intake (mg/day) in three Central European countries from the model described in Section 2.1. In each case, the red line represents estimated mean global sodium intake and the black line mean regional intake. Data from individual studies (calibrated to be gender balanced and measured using urinary assays) are shown as black points. Note the strong country-level clustering within region, which has shrunk sodium estimates for Hungary to the region average. The bottom 3 plots represent similar estimates but now from the joint model described in Section 2.2 where spatial clustering within region is not considered (ϕ=0). Posterior distributions are now much narrower and all three estimates have been shrunken aggressively. The reason this is happening is to improve the fit of equation (7), but arguably this gives undesirable estimates for sodium intake. Allowing spatial clustering in the second level regression ($\phi > 0$) somewhat lessens this effect.

Causes of the decrease of patent similarities from 1976 to 2021

Edoardo Filippi-Mazzola[1], Federica Bianchi[1], Ernst Wit[1]

[1] Institute of Computing, Università della Svizzera italiana, Lugano, Switzerland

E-mail for correspondence: edoardo.filippi-mazzola@usi.ch

Abstract: A citation network consists of patents citing prior art to which they are related. One way to study the relation between current patents and their antecedents is by analyzing the similarity between the textual elements of patents. Consistently across different ways of computing such measures, it was recently discovered that the similarity levels have been constantly decreasing since the mid 70s. Thanks to state of the art tools in Natural Language Processing, we propose a computationally efficient way to derive the similarity scores across patents citation pairs. Together with the use of General Additive Models, we analyzed the potential drivers for this downward trend. We found that the usage of non-linear models appropriately fits effects that are drastically influencing the similarity levels. With such corrections in place, the trend in similarity shows a different pattern than the one presented in previous studies.

Keywords: Citation networks, Neural Networks, BERT; Generalized Linear Models.

1 Introduction

Patent classes and classification systems have long been used by researchers to conduct socio-economical studies, concentrating the use of statistical methods on these classification systems as a reliable source of information. However, technological relatedness may not be directly related to sharing the same patent class. Despite the effort of analyzing and defining new measures of technological relatedness and closeness based on such classes, their usage can be problematic when patents need to be identified, compared, or matched with similar technologies. Younge and Kuhn (2015) proposed a new methodology for determining patent relatedness based on computing cosine similarities across pairs of citations. This is done by encoding

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

the textual description from such legal documents through a vector space model. Using more sophisticated methodology, involving the use of neural networks, Whalen et al. (2020) shows a constant decrease in the average similarity scores since the mid 70s. Kuhn et al. (2020) suggest that this may be the result of drastic changes that occurred in the citation generation process that has made the network of citations less informative and representative about the technological direction that patents are taking.
Instead, in this manuscript, we claim that this trend is affected by distinct phenomena that are characterizing the network. The implementation of models that apply corrections for such circumstances highlights a different pattern. With this work, we want to emphasize the use of Deep Learning techniques to compute similarity scores among the textual elements of patent citations as an appropriate measure of patent relatedness. Previous techniques have proven to be computationally intensive, especially for millions of patents. Instead, we propose a ready-to-use approach to compute similarity scores among patent citation pairs. We investigate through the use of Generalized Additive Models (GAMs) the potential causes for the downward trend in the patent similarity scores across years. Making use of distinct non-linear effects, we correct the similarity curve by properly reflecting a more informative trend.
This work is organized as follows. Section 2 presents the methodology used to define the similarity scores as well as the effects we used in the GAM. In Section 3 we discuss our results and in Section 4 we conclude with a short summary of our findings.

2 Methodology

2.1 Patent Similarity

Following the path suggested by Whalen et al. (2020), we focus on using a neural network approach to map the textual elements of patents into a multi-dimensional space despite the heavy computational requirements that comes with training a such a deep Machine Learning tool with so many inputs. One problem of the Whalen et al. (2020) approach is that they use entire technical descriptions as inputs. Instead, we agree with Choi et al. (2022) that patents abstracts contain the most useful and basic information regarding the patenting technology. Given the reduced textual elements inside the abstracts, we use a pre-trained BERT model to compute the matrix of embeddings. In this sense, we avoid the main this computational bottleneck. As such, we encoded approximately 7.5 Millions of patents into a 300-dimensional space. Through a scheme of lazy loading procedures, we managed to compute the patent similarity scores for almost 100 million citation pairs within minutes.

2.2 General Additive Model

Extending the linear modeling approach of Kuhn et al. (2020), we propose to model the vector of similarity scores through a Generalized Additive Model (GAM). With the combination of different linear and non-linear effects, we use distinct covariates to capture phenomena that have repercussions on the similarity levels.

From our exploratory analysis we note that the average temporal lag of citations is increasing over time. This suggests that textual similarity might be negatively affected by the presence of increasing temporal distances among citations. As such, we modelled through a smoothing term the temporal lag together with the publication date of the citing patent. Kuhn (2010) results suggest that legal changes brought an increase in the number of cited patents. To correct for such inflation we considered the amount of backward citations done by the citing patent via a smooth term. We also argue that if we need to consider the number of citations, we have to discriminate between owners types that provide the citation. As such, we distinguish between companies and private owners by adding three binary effects: *Same c.* (if the citing and the cited company coincide), *Cited c.* and *Citing c.* (if

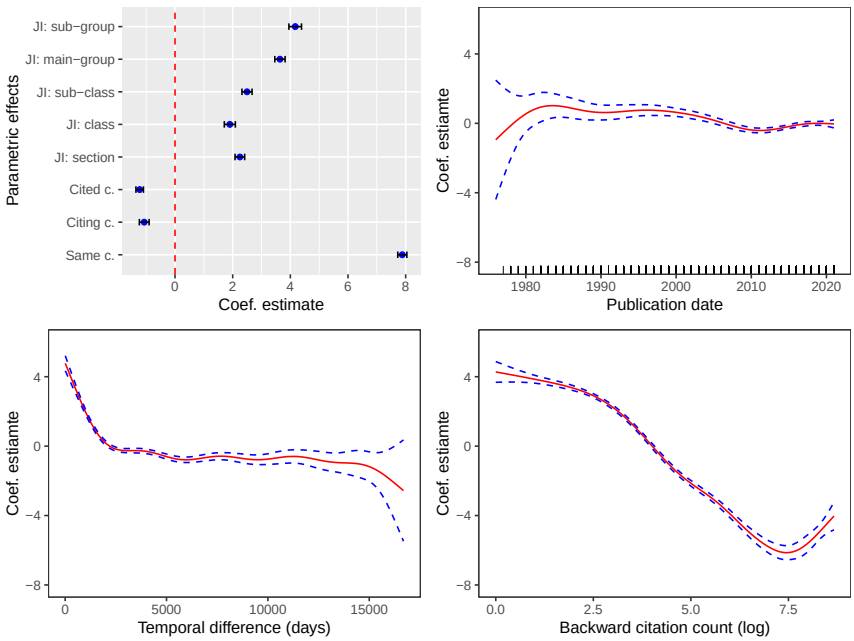


FIGURE 1. GAM estimated coefficients. The intercept estimated coef. is 41.06. *Top-left*: fixed parametric effects. *Top-right*: spline on publication date. *Down-left*: spline on temporal lag (days). *Down-right*: spline on backward citation count (log).

either the cited or the citing patent is owned by a company). To complete our analysis, we introduced effects that are directly related to the patenting technology. This was done by computing the *Jaccard index* for each hierarchical component of the International Patent Classification (IPC) scheme (section, class, sub-class, main group and sub-group) on each citing pair.

3 Results

A summarized version of the estimated coefficients is shown in Figure 1. The model was fitted with deviance explained of 18%, which is an important increase from previous studies. Interestingly, the combination of all the presented effects forces the publication date effect to assume an unusual behavior. Indeed, the spline in publication date shows that the similarity trend is not downward, but it alternates periods of increase and decrease.

4 Conclusion

With this work, we develop Deep Learning techniques to compute textual similarity scores between patents in a new and efficient way. Together with the use of GAMs, we modeled similarity values to fit those effects that have a major influence on the response. In light of our results, we show that contrary to previous studies, the downward trend in similarity scores is the consequence of a multitude of exogenous phenomena. If appropriate corrections are applied, the similarity levels show a different trend compared to other studies.

References

Choi, S., Lee, H., Park, E., and Choi, S. (2022). *Deep learning for patent landscaping using transformer and graph embedding.* Technological Forecasting and Social Change, 175:121413.

Kuhn, J., Younge, K., and Marco, A. (2020). *Patent citations reexamined.* The RAND Journal of Economics, 51(1):109–132.

Kuhn, J. M. (2010). *Information overload at the u.s. patent and trademark office: Reframing the duty of disclosure in patent law as a search and filter problem.* page 52. Oxford: Clarendon Press.

Whalen, R., Lungeanu, A., DeChurch, L., and Contractor, N. (2020). *Patent similarity data and innovation metrics.* Journal of Empirical Legal Studies, 17(3):615–639.

Younge, K. A. and Kuhn, J. M. (2015). *Patent-to-patent similarity: A vector space model.* SSRN Electronic Journal.

Probabilistic prediction: aims and solutions

Giovanni Fonseca[1], Federica Giummolè[2], Paolo Vidoni[1]

[1] Dept. of Economics and Statistics, University of Udine, Udine, Italy
[2] Dept. of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy

E-mail for correspondence: giovanni.fonseca@uniud.it

Abstract: The goodness of a predictive distribution depends on the aim of the prediction. This presentation intends to shed light on properties of predictive distributions in use nowadays. We also propose a new predictive distribution that may be useful to obtain calibrated predictions for the probabilities of a future random variable of interest. This predictive distribution can be easily computed by a simple bootstrap procedure. In order to compare the different predictive distributions, some simulation studies are also presented.

Keywords: Bootstrap, Calibration, Prediction.

1 Introduction

Let us define the notation and the general assumptions that we will use in the sequel. Suppose that $\{Y_i\}_{i\geq 1}$ is a sequence of continuous random variables with probability distribution depending on an unknown d-dimensional parameter $\theta \in \Theta \subseteq \mathbf{R}^d$, $d \geq 1$; $Y = (Y_1, \cdots, Y_n)$, $n > 1$, is observable, while $Z = Y_{n+1}$ is a future or not yet available observation. For simplicity, we consider the case of Y and Z being independent random variables and we indicate with $G(z;\theta)$ and $Q(\alpha;\theta)$ the distribution function and the quantile function of Z, respectively. Given the observed sample $y = (y_1, \ldots, y_n)$, we look for a predictive distribution $\hat{G}(z;y)$, with corresponding quantile function $\hat{Q}(\alpha;y)$, that fullfills some good requirements for prediction.
There are different desirable properties that a predictive distribution should possess. Here we consider only two of the most important:

(A) calibrated quantile function: $E_Y[G\{\hat{Q}(\alpha;Y);\theta\}] = \alpha$, $\forall \alpha \in (0,1)$

(B) calibrated distribution function: $E_Y[Q\{\hat{G}(z;Y);\theta\}] = z$, $\forall z \in \mathbf{R}$.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

Unfortunately, these properties cannot be satisfied at the same time. They regard different aspects of a predictive distribution and depend on the target of the prediction itself. A quantile function is calibrated if, in mean, it coincides with the inverse of the true distribution function. This last property can also be expressed in terms of coverage probabilities, since

$$E_Y[G\{\hat{Q}(\alpha;Y);\theta\}] = P_{Z,Y}\{Z \leq \hat{Q}(\alpha;Y)\}.$$

Similarly, a predictive distribution function is calibrated if, in mean, it coincides with the inverse of the true quantile function. While from a theoretical point of view the knowledge of the distribution function coincides with that of the quantile function for continuous random variables, this is not true when we talk about predictive distributions. Thus, a predictive distribution may be good for estimating quantiles but not as good for estimating probabilities and the converse is also true.
In the sequel we always consider the maximum likelihood estimator (mle) $\hat{\theta} = \hat{\theta}(Y)$ for θ, or an asymptotically equivalent alternative. The estimative predictive distribution and quantile functions, $G(z;\hat{\theta})$ and $Q(\alpha;\hat{\theta})$ respectively, usually satisfy properties (A) and (B) with an error term of order $O(n^{-1})$, as the sample size $n \to +\infty$, see e.g. Barndorff-Nielsen and Cox (1996). It is well known that this error term could be substantial, in particular for small sample sizes.

2 Calibrated quantile functions

Modern literature has largely focused on the problem of prediction limits, that is the problem of finding a predictive distribution which quantiles satisfy property (A) with a high approximation. This requirement is usually met when a pivotal quantity for prediction is available. Unfortunately in many situations of practical interest, a pivot is not known. Furthermore, even in the case of the normal distribution, sometimes the unknown parameters are estimated using ad hoc estimators whose exact distribution is unknown. As a consequence, the distribution of a quantity such as $(Z - \hat{\mu})/\hat{\sigma}$ is not known. Thus, it becomes of interest in the applications to find alternative approximate solutions.
Here we quickly recall the procedure used in Fonseca et al. (2014), since we will follow the same steps in the next section. The starting point is the coverage probability associated to the estimative quantile function $Q(\alpha;\hat{\theta})$:

$$P_{Y,Z}\{Z \leq Q(\alpha;\hat{\theta});\theta\} = E_Y[G\{Q(\alpha;\hat{\theta});\theta\}] = C(\alpha,\theta).$$

Although an explicit expression of this coverage probability is rarely available, it is well-known that it does not match the target value α. Fonseca et al. (2014) noticed that the function $G_c(z;\hat{\theta},\theta) = C\{G(z;\hat{\theta}),\theta\}$, obtained by substituting α with $G(z;\hat{\theta})$ in $C(\alpha,\theta)$, is a proper predictive distribution

function, whose associated quantile function is calibrated, giving coverage probability equal to the target nominal value α, for all $\alpha \in (0,1)$. A suitable parametric bootstrap estimator for $G_c(z;\hat{\theta},\theta)$ may be readily defined as

$$G_c^{boot}(z;\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} G\{Q(\alpha;\hat{\theta}^b);\hat{\theta}\}|_{\alpha=G(z;\hat{\theta})},$$

where $\hat{\theta}^b$, $b = 1,\ldots,B$, are estimates obtained with B bootstrap samples from $G(z;\hat{\theta})$. The corresponding α-quantile defines, for each $\alpha \in (0,1)$, a prediction limit having coverage probability equal to the target α, with an error term which depends on the efficiency of the bootstrap simulation procedure.

3 Calibrated distribution functions

In this section we address the dual problem, looking for predictive distributions that satisfy property (B). We use exactly the same ideas proposed by Fonseca et al. (2014) and recalled in the previous section, applied to the distribution function instead of the quantile function. The result is a new predictive distribution that may be useful for predicting probabilities for the interest variable Z, instead of quantiles.
The estimative distribution function is not well calibrated in the sense of property (B). Infact, the mean of quantiles of level equal to $G(z;\hat{\theta})$ is

$$E_Y[Q\{G(z;\hat{\theta});\theta\}] = A(z,\theta)$$

and it does not match the target value z. Instead, the function

$$Q_c(\alpha;\hat{\theta},\theta) = A\{Q(\alpha;\hat{\theta}),\theta\}, \tag{1}$$

obtained by substituting z with $Q(\alpha;\hat{\theta})$ in $A(z,\theta)$, is a proper predictive quantile function whose distribution function $G_c(z;\hat{\theta},\theta) = G\{A^{-1}(z,\theta);\hat{\theta}\}$ satisfies property (B) for every $z \in \mathbf{R}$. Indeed,

$$\begin{aligned} E_Y[Q\{G_c(z;\hat{\theta},\theta);\theta\}] &= E_Y[Q\{G(A^{-1}(z,\theta);\hat{\theta});\theta\}] \\ &= A\{A^{-1}(z,\theta),\theta\} = z. \end{aligned}$$

The predictive quantile function (1) and the corresponding calibrated predictive distribution are not useful in practice, since they depend on the unknown parameter θ. However, a suitable parametric bootstrap estimator for $Q_c(\alpha;\hat{\theta},\theta)$ may be readily defined. Let y^b, $b = 1,\ldots,B$, be parametric bootstrap samples generated from the estimative distribution of the data and let $\hat{\theta}^b$, $b = 1,\ldots,B$, be the corresponding estimates. We can thus write

$$Q_c^{boot}(\alpha;\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} Q\{G(z;\hat{\theta}^b);\hat{\theta}\}|_{z=Q(\alpha;\hat{\theta})}.$$

The corresponding distribution function allows to predict the target probability $P(Z \leq z_0)$, for each $z_0 \in \mathbf{R}$, with an error term which depends on the efficiency of the bootstrap simulation procedure. Indeed, the estimate is the value α_0 such that $Q_c^{boot}(\alpha_0; \hat{\theta}) = z_0$.

4 The normal distribution

Let $Y_1, \ldots, Y_n, Z$ be independent and normally distributed with mean μ and standard deviation σ, both unknown. In this context the pivotal quantity $T = \sqrt{n/(n+1)}(Z - \bar{Y})/S$ is useful for prediction, with $\bar{Y}$ and S the sample mean and sample standard deviation, respectively. Its distribution is Student t with $n-1$ degrees of freedom. The quantile function $\sqrt{(n+1)/n}\, Q_t\,(\alpha; n-1)\, S + \bar{Y}$, satisfies property (A). Hence, in this case, the calibrated quantile function presented in section 2 replicates the distribution obtained from the pivot. However, as shown in the following simulation study, the pivotal distribution is not the best choice for prediction of probabilities.

The following tables show the results of Monte Carlo simulations based on $M = 10000$ replications and $B = 500$ bootstrap replications for the computation of the calibrated distributions. The sample size is $n = 10, 25$ and the true parameter values are $\mu = 0$ and $\sigma = 1$. We have compared the estimative distribution with the mle, the predictive distribution obtained from the pivotal quantity and the two bootstrap calibrated predictive distributions on the basis of the corresponding coverage probability for $\alpha = 0.5, 0.9, 0.95, 0.99, 0.999$ (Table 1) and the mean quantiles of levels $\hat{G}(z; y)$ for $z = 0, 1.5, 2, 2.5, 3.5$ (Table 2). The best performances are written in bold face, clearly showing how the aim of the prediction should influence on the choice of the predictive distribution.

TABLE 1. Coverage probabilities. Standard errors smaller than 0.001.

	Target	estim.	pivotal	qu. calib.	pr. calib.
n=10	$\alpha = 0.5$	**0.500**	**0.500**	**0.500**	**0.500**
	$\alpha = 0.9$	0.861	**0.900**	0.899	0.892
	$\alpha = 0.95$	0.914	**0.950**	0.949	0.939
	$\alpha = 0.99$	0.967	**0.990**	**0.990**	0.981
	$\alpha = 0.999$	0.989	**0.999**	**0.999**	0.995
n=25	$\alpha = 0.5$	**0.500**	**0.500**	**0.500**	**0.500**
	$\alpha = 0.9$	0.885	**0.900**	**0.900**	0.897
	$\alpha = 0.95$	0.936	**0.950**	**0.950**	0.946
	$\alpha = 0.99$	0.983	**0.990**	**0.990**	0.987
	$\alpha = 0.999$	0.997	**0.999**	**0.999**	0.998

TABLE 2. Mean quantiles of level $\hat{G}(z; y)$. Standard errors smaller than 0.001

	Target	estim.	pivotal	qu. calib.	pr. calib.
n=10	$z = 0$	-0.001	-0.001	**0.000**	**0.000**
	$z = 1.5$	1.734	1.411	1.411	**1.504**
	$z = 2$	2.312	1.803	1.804	**2.004**
	$z = 2.5$	2.889	2.151	2.153	**2.505**
	$z = 3.5$	4.044	2.732	2.741	**3.498**
n=25	$z = 0$	**0.000**	**0.000**	**0.000**	**0.000**
	$z = 1.5$	1.581	1.465	1.465	**1.500**
	$z = 2$	2.108	1.920	1.920	**2.000**
	$z = 2.5$	2.635	2.350	2.350	**2.500**
	$z = 3.5$	3.689	3.130	3.133	**3.500**

References

Barndorff-Nielsen, O.E. and Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli*, **2**, 319 – 340.

Fonseca, G., Giummolè F. and Vidoni, P. (2014). Calibrating predictive distributions. *Journal of Statistical Computation and Simulation*, **84**, 373 – 383.

Model Selection for Predicting Readmissions with Health Insurance Data

Alexander Gerharz[1], Carmen Ruff[2], Andreas Groll[1], Andreas D. Meid[2]

[1] Department of Statistics, TU Dortmund University, Dortmund, Germany
[2] Department of Clinical Pharmacology and Pharmacoepidemiology, Heidelberg University Hospital, Germany

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

Abstract: Over the last three decades the number of admissions to German hospitals per year has increased rapidly. About a third of these admitted patients have to be readmitted within the same year. Within the healthcare system in Germany, health insurances collect lots of data submitted by physicians (outpatient sector) and hospitals (inpatient sector), among others. In our research project, we predict the probability that a patient has to be readmitted within the next 90 days after an admission by only using these routine data of a health insurance company. We compare classical statistical (regularization) methods and machine learning methods for this task.

Keywords: Boosting; Variable Selection; Benchmarking; Machine Learning.

1 Introduction

Since 1991 the number of admissions to a hospital in Germany has increased from roughly 14.6 million to 19.4 million in 2019. We find that roughly a third of the patients are admitted more than once within a year. Previous studies have already concluded that a lot of these readmissions might be avoided with better care or appropriate discharges from the hospital (see, e.g., Yam et al., 2010). As readmissions represent a huge burden on the healthcare systems, there are lots of studies about their prediction (see, e.g., Artetxe et al., 2018). In our research project, we use health insurance data to model the probability that a patient has to be readmitted within the next 90 days. As in the data preprocessing a large number of variables has

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

been retrieved we compare classical statistical approaches, regularization methods and machine learning techniques in this benchmark analysis.

2 Data

The data used for this study was provided by the AOK Baden-Württemberg, a large German health insurance company. Following data processing, continuously insured people aged 65 years or older were included from 2011 to 2016. The provided data contains information about basic demographics, inpatient diagnosis, and the outpatient medication recieved before and after a stay in hospital (see Ruff et al., 2021). To determine cases for the prediction of readmissions, we chose all cases as index admissions in which the main diagnosis can be assigned to one of the following diseases (see Gerharz et al., 2022): acute myocardial infarction (AMI; 18,641 cases), heart failure (HF; 78,065 cases), a composite of stroke, transient ischemic attack, or atrial fibrillation (S/AF; 81,225 cases), chronic obstructive pulmonary disease (COPD; 27,725 cases), type 2 diabetes mellitus (DM; 23,491 cases) and osteoporosis (OS; 6,101 cases). For each of these diseases a data set is constructed containing the index admissions for the respective disease. As a readmission we then define every following admission within 90 days that also belongs to the respective disease.

The variables contain information about the age and gender of the patient, the number of previous hospitalizations within the last 365 days, the month of the index hospitalization, information regarding the discharge, the main diagnosis, secondary outpatient inpatient diagnoses, medication as disclosed to the health insurance company, comorbidities and STOPP/START (screening tool of older persons' potentially inappropriate prescriptions/screening tool to alert doctors to the right treatment) criteria (Meid et al., 2018). We also consider every recieved medication from 90 days before the index admission up to until 10 days after the index admission. The Elixhauser comorbidities and the STOPP/START criteria are built with all the information from the diagnoses and the medication (Elixhauser et al., 1998). All the occuring secondary diagnoses, medication, comorbidities and STOPP/START criteria are added to the data set for the respective disease if their prevalence exceeds 1%. Altogether, this results in 200-300 additional binary variables per disease.

3 Methods

As benchmark model we decide to use a simple GLM containing just the basic information about patients. This contains basic demographics and the information about the discharge of the patients. For the whole dataset, we chose to compare a simple GLM, classical regularization approaches

(LASSO, RIDGE, relaxed-LASSO), classical machine learning techniques (kNN, CART, RF, NEUR_NET) and classical boosting methods (GLM-BOOST, XGBOOST).

Based on the idea of the relaxed-LASSO (sometimes also named post-LASSO), where variables are selected based on the non-zero coefficient estimates of the LASSO regularization and afterwards a full GLM is fitted on these variables, we also put together a similar approach combining random forests and a base model. First, we use the variable importance of a random forest to rank our variables. This can be done either with the GINI-Impurity-Decrease or with the Permutation variable importance. Afterwards, we fit a base model on the most important variables. As base model, we choose a GLM or a CART as an alternative. The number of most important variables used is considered as a hyperparameter and can be tuned. This results in four methods, which we abbreviate with GINI_GLM, GINI_CART, PERM_GLM and PERM_CART.

A hyperparameter tuning step is conducted for the methods where necessary. The performance is evaluated by computing the AUC in a 10-fold-cross-validation for each of the six data sets individually.

4 Results

TABLE 1. Ranges of median AUC performance for all methods except baseline.

	AMI	HF	S/AF	COPD	DM	OS
Range	0.60-0.63	0.61-0.64	0.65-0.69	0.64-0.69	0.67-0.69	0.51-0.56

Previous literature has shown that it is very hard to predict readmissions. Additionally, in this study no clinical data from conducted tests was provided. Nevertheless, the methods show decent performances (see Table 1). The only disease that is not very well predicted by all the methods is osteoporosis. As it had just a very small number of index admissions it was by far the smallest data set and no method achieved a good performance.

In Table 2, the ranks of the median performances of the methods for each disease are shown. As the performance for osteoporosis was very different from the others, in this table an average rank for all diseases and an average rank for all diseases except osteoporosis is displayed for each method. Here, it is shown that especially methods which use a variable selection step perform very well.

TABLE 2. Ranks of the median performance (AUC) of the methods for each disease.

	AMI	HF	S/AF	COPD	DM	OS	$\varnothing Rank$	$\varnothing Rank_{noOS}$
Baseline	15	15	15	15	15	9	14.00	15.0
GLM	9	2	7	5	6	3	5.33	5.8
LASSO	2	1	1	3	4	6	2.83	2.2
RIDGE	7	7	3	7	2	1	4.50	5.2
relaxed-LASSO	8	4	6	8	10	2	6.33	7.2
kNN	13	11	14	14	14	15	13.50	13.2
CART	11	14	13	12	12	13	12.50	12.4
RF	10	10	10	9	9	10	9.67	9.6
NEUR_NET	4	6	9	6	7	7	6.50	6.4
GLMBOOST	5	8	4	4	3	11	5.83	4.8
XGBOOST	1	9	8	10	8	8	7.33	7.2
GINI_GLM	6	5	5	2	5	5	4.67	4.6
GINI_CART	14	12	12	13	13	14	13.00	12.8
PERM_GLM	3	3	2	1	1	4	**2.33**	**2.0**
PERM_CART	12	13	11	11	11	12	11.67	11.6

The best performance was achieved by using the permutation variable importance of a random forest to rank all of the variables and then use hyperparameter tuning to evaluate how many variables should be included in the successive GLM starting from the most important one. The variable selection by using the GINI-Impurity-Decrease variable importance and a subsequent GLM also ranks pretty good, but slightly worse than the same procedure with the permutation variable importance. Also, it is a huge advantage that these models are easy to interpret.

The next best method is the LASSO, which is part of the classical regularization methods and also includes variable selection. It is especially interesting that this method performed best for the two diseases with the largest amounts of cases. RIDGE also ranks pretty good.

From the group of machine learning and boosting methods both boosting methods and the NEUR_NET rank better than the other methods. GLMBOOST and XGBOOST contain intrinsic variable selection in the sense that only the best variables are used in each step of the model building process. Especially, the GLMBOOST ranks very good and should be considered as possible method for these modeling tasks, while the XGBOOST has really excelled for just one of the diseases. Even though the CART and RF also include intrinsic variable selection, these methods do not rank well in this comparison.

Principally, the research field of neural nets is very broad. For this project,

a simple neural net with one hidden layer and a fixed amount of nodes was used. The number of epochs needed for training was determined in a hyperparameter tuning step. In this project, the NEUR_NET already achieves a decent performance, but an even better performance might be achieved with a different kind of architecture of the net.

On should notice that the CART, RF, XGBOOST and the variable-importance-based variable selection with successive CART are all tree-based methods and that all of them perform not very good. This is an indicator that tree-based methods might not work well for this modeling task compared with the other methods (see Figure 1). The only method, which in most of the cases performs even worse is the k-nearest-neighbors method.

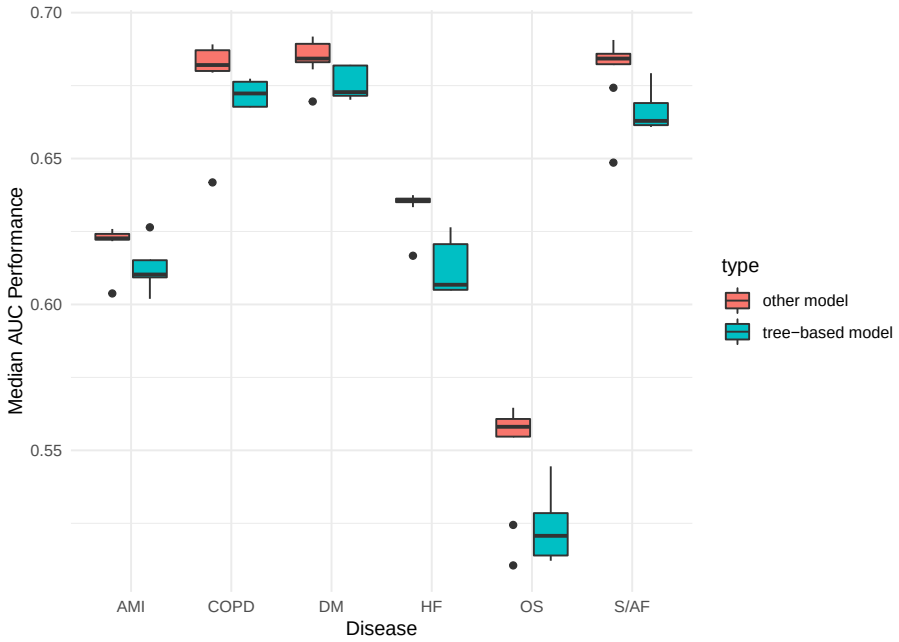


FIGURE 1. Comparison of the median AUC performances between the tree-based methods and the other methods.

5 Conclusion

In this benchmark study, the performance of different methods, which estimated the probability for readmission for six different diseases on health insurance data, was investigated. The data contained just a few metric and

categorical variables, but a huge amount of additional binary variables. For this binary classification task, especially methods considering variable selection performed extremely well. Particularly, a simple GLM following the variable selection using the variable importance of a random forest, the LASSO and the GLMBOOST methods performed really well and should be considered for these kind of tasks.

Acknowledgments: This work was supported by the German Innovation Funds according to § 92a (2) Volume V of the Social Insurance Code (§ 92a Abs. 2, SGB V - Fünftes Buch Sozialgesetzbuch), grant number: 01VSF18019. The funding body did not play any role in the design of the study, the collection, analyses, and interpretation of data, or the writing of the manuscript. Andreas D. Meid is funded by the Physician-Scientist Programme of the Medical Faculty of Heidelberg University.

References

Artetxe, A., Beristain, A., and Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer Methods and Programs in Biomedicine*, **164**, 49 – 64.

Elixhauser, A., Steiner, C., Harris, D.R., and Coffey, R.M. (1998). Comorbidity measures for use with administrative data. *Med Care*, **36 (1)**, 8 – 27.

Gerharz, A., Ruff, C., Wirbka, L., Stoll, F., Haefeli, W.E., Groll, A., and Meid, A.D. (2022). *Predicting Hospital Readmissions from Health Insurance Claims Data: A Modeling Study Targeting Potentially Inappropriate Prescribing.* Methods Inf Med.

Meid, A.D., Groll, A., and Heider, D., Mächler, S., Adler, J.-B., Günster, C., König, H.-H., and Haefeli, W.E. (2018). Prediction of drug-related risks using clinical context information in longitudinal claims data. *Value Health*, **21(12)**, 1390 – 1398.

Ruff, C., Gerharz, A., Groll, A., Stoll, F., Wirbka, L., Haefeli, W.E., Meid, A.D. (2021). Disease-dependent variations in the timing and causes of readmissions in Germany: A claims data analysis for six different conditions. *PLoS One*, **16(4)**, e0250298.

Yam, C.H., Wong, E.L., Chan, F.W., Wong, F.Y., Leung, M.C., and Yeoh, E.K. (2010). Measuring and preventing potentially avoidable hospital readmissions: a review of the literature. *Hong Kong Med J*, **16**, 383 – 389.

Estimating periodicity in disease dynamics

Niamh Graham[1], Natalia Bochkina [1], Damian Mole [1]

[1] University of Edinburgh, UK

E-mail for correspondence: n.e.graham@sms.ed.ac.uk

Abstract: Periodicity in disease progression has not been widely explored, although there has been some investigation into the periodicity of CRP, a blood plasma protein which is a biomarker for inflammation, Dorraki (2018). Here we aim to model periodicity in the disease progression of an inflammatory disease, acute pancreatitis (AP). Two factors, termed protective and damaging factors, characterise the disease progression and are modelled by a complex valued Gaussian process with Markovian structure. We develop appropriate inference for this approach showing that the first two principal components in the AP data have the correct interpretation conjectured by Bart et al. (2005) and have a clear medical interpretation. Periodicity of disease progression is identified, as well as optimal medical intervention times and their uncertainty.

Keywords: Statistical disease modelling; Gaussian processes; Missing data analysis

1 Introduction to model

Bart et al. (2005) formulated a model to describe the progression of chronic glomerulonephritis. The model incorporates both disease progression and the organism's response to the disease. Here, disease progression is characterised by the interaction of damaging (pathogenetic) and protective (sanogenetic) factors, which are in opposition throughout the disease course and result in the following oscillating disease progression function,

$$S(t) = \sigma^2 e^{-\eta t} \cos \tau t. \tag{1}$$

The parameters of the function, η and τ, correspond to the severity and periodicity of the disease respectively and the estimation of these parameters allow the identification of points in the disease progression where medical intervention will be most effective.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

In their further work, Bart et al. hypothesised that these factors could be identified from clinical data using principal component analysis and proposed a parameter estimation framework based on the method of moments. Following Bart et al. (2005), we consider the following model. First, using principal component analysis (PCA) we identify two factors, U_t and V_t, which have the interpretations of damaging and protective factors corresponding to the disease progression and the body's response to the disease. Secondly, k regularly spaced observations of these principal components are modelled as a Gaussian Markov chain. In vector representation, $x_t = \begin{pmatrix} U_t \\ V_t \end{pmatrix}$ is modelled as

$$\begin{aligned} x_{t+1} &= e^{-\eta}\mathbf{A}x_t + \sqrt{1-e^{-2\eta}}\,\mathbf{B}\xi_t, \quad t = 2,\ldots,k, && (2) \\ x_t,\, \xi_t &\sim N_2(0, 0.5\sigma^2 I_2) \text{ independently } t = 1,\ldots,k, && (3) \end{aligned}$$

where

$$\mathbf{A} = \begin{pmatrix} \cos\tau & -\sin\tau \\ \sin\tau & \cos\tau \end{pmatrix},\ \mathbf{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

This model can be written equivalently in the complex form:

$$X_t = U_t + iV_t \sim N_C(0,\sigma^2), \quad E(X_s X^{\star}_{s+t}) = \sigma^2 e^{-\eta|t| - i\tau t}, \quad E(X_s X_{s+t}) = 0, \tag{4}$$

where $X^{\star}_{s+t}$ is a complex conjugate of X_{s+t}.

2 Main results

2.1 Simulated data

Using the Markov relationship from Equation 2, a complex data set has been simulated to assess the suitability of the parameter estimators and confidence regions used in Bart et al., (2005).
We derive alternative maximum likelihood estimators (MLEs) for the model and investigate their consistency and asymptotic distribution. We find that the confidence regions presented in Bart et al., (2005) are not appropriate since in the limit of the noise tending to 0, the confidence region converges to an ellipse rather than to a true value, and they do not include an open set around the estimate. We therefore construct at Wald-type asymptotic $(1-\alpha)100\%$ confidence region for (θ, τ) of the following form

$$\frac{(1+\hat{\theta}^2)}{(1-\hat{\theta}^2)^2}(\hat{\theta}-\theta)^2 + \frac{\hat{\theta^2}}{(1-\hat{\theta}^2)}(\hat{\tau}-\tau)^2 \leq \frac{\chi^2_2(\alpha)}{2m(k-1)}, \tag{5}$$

where $\theta = e^{-\eta}$.
These confidence regions exhibit more appropriate behaviour, tending to the parameter estimate in the asymptotic limit. The behaviour of the Wald-type confidence regions are shown in Figure 1.

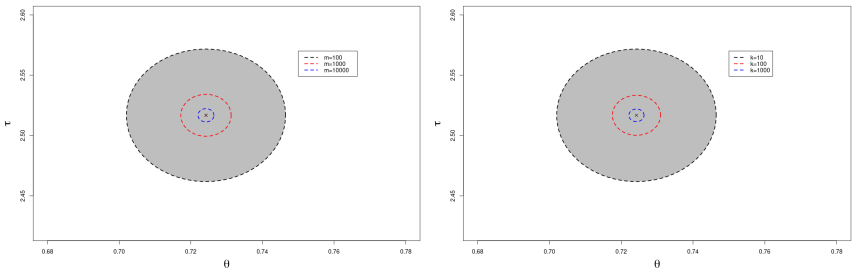


FIGURE 1. The Wald confidence region for η and τ with increasing individuals (left) and increasing time points (right)

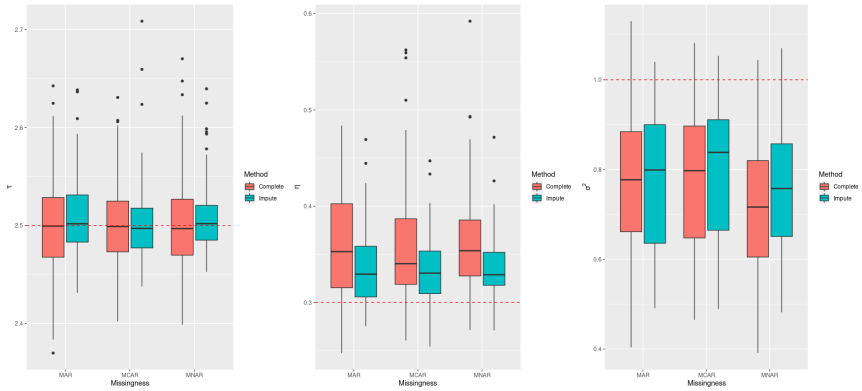


FIGURE 2. Box plots comparing the effect of complete case analysis with imputation by PCA on the parameter estimates

Since the real data that the model is applied to contains missing data, two missing data handling methods, complete case analysis and imputation by PCA, were compared. Complete case analysis, where any observation with any missing values is removed from the data set and the analysis is applied only on the remaining data usually results in biased estimates unless the missingness is missing completely at random (MCAR). Imputation by PCA is an imputation technique developed by Josse et al. (2016). It involves iteratively carrying out PCA on the data with imputed values projected onto the principal components until convergence is met. Figure 2 shows the results of parameter estimation using both of these methods. It shows that the τ estimate is robust to all types of missing data and can be well estimated using both complete case analysis and imputation by PCA. For the η and σ^2 parameters, the estimates are less biased when using imputation by PCA and the variance around the η parameter is reduced.

2.2 Application to clinical data

The method has been applied to acute pancreatitis (AP) clinical trial data. The data set consists of regularly structured sampling observation times (0h, 3h, 6h, 12h, 24h, 48h, 72h, 168h), which are common over all patients. They are arranged such that subsets of the sampling times, which are regularly spaced, can be used separately for the parameter estimation. The results from using the 0h, 12h and 24h time points have been used initially as they contain the least amount of missing data.
The first two principal components correspond well with the hypothesis from Bart et al. (2005). The first principal component has strong correlation with variables such as APACHE which is a score of severity in AP and age, and the second principal component has strong correlation with immune cells. These can be interpreted as damaging and protective factors in the model.
With the number of patients $m = 28$ and number of time steps $k = 3$, estimates for τ, η and σ^2 can be obtained, where η is the survival rate and τ determines the cycle of the disease. Using the likelihood based approach as described in previously, the following parameter estimates are obtained:

- $\hat{\eta} = 0.23$
- $\hat{\tau} = 6.16$
- $\hat{\sigma^2} = 8.97$

The periodicity of this process was estimated as 12.25 hours with a confidence interval of [11.4,13.2] hours.
By adapting the likelihood function, more of these time points from the data frame can be used to get a more reliable estimate of the parameters.

References

Bart, A. G. et al. (2005). Modeling disease dynamics and survivor functions by sanogenesis curves. *Journal of statistical planning and inference* , 132(1-2), **33-51**.

Bondarenko B.B. et al. (1980). Mathematical analysis of the evolution of chronic glomerulonephritis. *Medicine*, **213–225**.

Coventry, Brendon J., et al. (2009). CRP identifies homeostatic immune oscillations in cancer patients: a potential treatment targeting tool?. *Journal of translational medicine* **1-8**

Dorraki, Mohsen, et al. (2018). On detection of periodicity in C-reactive protein (CRP) levels. *Scientific reports* **1-7**.

Josse, Julie and Husson, François (2016). missMDA: a package for handling missing values in multivariate data analysis *Journal of Statistical Software* **1–31**

Identification of Possible COVID-19 Pandemic Impact on Scottish National Therapeutic Indicators

Alba Halliday[1,2]

[1] Public Health Scotland, Scotland
[2] University of Glasgow, Scotland

E-mail for correspondence: 2714134h@student.gla.ac.uk

Abstract: National Theraputic Indicators (NTIs) are a prescribing tool to aid improvement in patient care, prescribing safety and spending efficiency. This analysis investigates whether any NTIs for primary care in Scotland have been influenced by the presence of the COVID-19 pandemic in order to aid monitoring and targeted interventions where necessary. To achieve this a piecewise linear regression using a Multivariate Adaptive Regression Spline (MARS) modelling approach is used to identify significant changes in the NTIs occuring at the start of 2020.

Keywords: Prescribing; NTI; Scotland; Piecewise; MARS.

1 Introduction

The National Therapeutic Indicators (NTIs) for Scotland were initally developed in 2012 to identify differences between general practitioners (GPs) preformance quality, as discussed in MacBride-Stewart, S. et al (2019). Originally 12 indicators were developed to aid the maintenance and improvement of patient safety and spending efficiency in primary care. These have now been extended to include 40 indicators which has allowed more areas of concern to be monitored. For example, one indicator added in the 2018/19 primary care NTI report was 'MHRA Warning (valproate in women of childbearing age %)'. This indicator measures the percentage of female patients aged 13-45 who are prescribed the drug valproate for treatment of epilepsy or bipolar disorder out of all females prescribed it. Valproate is known to cause an increase in the risk of birth defects and

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

development disabilities when taken during pregnancy, but there is still evidence that woman taking the drug are not being made sufficiently aware of this risk, as alerted by the UK Government Medicines and Healthcare products Regulatory Agency (2017). NTIs are open data published on the Public Health Scotland (PHS) website dashboard. Quantifying the direction, magnitude and time of changes in trends of NTIs could determine whether past interventions have been effective. Identifying NTIs that have a stable behaviour will hopefully result in removing indicators where improvements have been satisfactory, making room for higher priority reviews of prescribing performance. Comparrison between each of the 14 Scottish NHS health boards could also help to determine whether those with larger populations were skewing the overall trend for Scotland. Furthermore, it will allow for differences in priorities and performance of prescribing improvement between health boards to be investigated.

2 Methodology

2.1 Piecewise Linear Regression

Piecewise linear regression is where the linear predictor of a model is made up of several straight line segments defined by basis functions. Where two separate segments meet is known as a knot, and these can either be joined or disjoined depending on whether the two lines are continuous or not. The underlying functions of alternative approaches, such as smooth spline regression, are not usually available. This makes them less suitable for this context as it would be difficult to quantify patterns in the data. Additionally, a piecewise linear regression can offer quick execution which is beneficial since a large number of data sets have to be modelled if the data is updated. Finally, this modelling framework offers the flexibility to account for non-linearities and seasonality.
A similar methodology was used by MacBride-Stewart, S. et al (2017) to investigate three NTIs associated with high risk prescribing of NSAIDs. A disjoint piecewise linear model was fitted with three segments; one for a known intervention period of twelve months, one for the time before this, and one for the time after. The impact of the intervention on prescribing and whether this was continued after the intervention period ended could then be evaluated. However, this approach is reliant on expert knowledge about the implementation of prescribing policies and when they start to impact prescribing. The affect of new policies can sometimes take a while to be reflected in the data due to the delay between publication and the awareness or action of those who prescribe the related drugs. On the other hand, sometimes changes start before the intervention as practioners are likely to be aware of possible issues and start to change their conduct accordingly before official advice can be amended. Therefore, personal assumptions may induce bias into the model. Moreover, other factors can

alter the trend in an NTI such as drug classifications, data collection, population demographics and disease outbreaks which aren't acknowledged in this previous framework.

2.2 Multivariate Adaptive Regression Splines

The Multivariate Adaptive Regression Splines (MARS) algorithm was introduced in Friedman, J.H. (1991) which was inspired by recursive partioning but allowed for continuous models with multivariate interactions. This was then developed into the R package "earth" by Milborrow, S. (2011). It allows for the estimation of the location and number of basis functions of each explanatory variable as well as any interactions between them. However, since the only available explanatory variable is time in this case, Equations 1 and 2 outline the univariate piecewise regression that is fitted by MARS. The given NTI measure y is modelled using a Normal distribution with a constant variance σ^2. The time t represents the yearly quarter the NTI was measured over. The location of the knots are given by c_k where k is the total number of knots. The model coefficients are $\beta_1, ..., \beta_k$ which correspond to the given knot's basis function, and β_0 is the intercept term of the model. A log link for the linear predictor of the mean μ is used since all NTI measures are strictly positive.

$$y \sim \text{Normal}(\mu, \sigma^2) \tag{1}$$

$$log(\mu) = \beta_0 + \beta_1 \text{max}(0, c_1 - t) + ... + \beta_k \text{max}(0, c_k - t) \tag{2}$$

Since the number of knots in the linear predictor is unknown beforehand this is a non-parametric modelling technique. Even though the class of functions constructing the linear predictor has been predetermined to be a series of linear functions, the parameters defining the functions are decided by the data. MARS achieves this by preforming a Forward Pass, which adds pairs of hinge basis functions to the initial intercept only linear predictor using least squares regression. A pair of hinge basis functions have the form $\beta_{m_1} max(0, c_m - t) + \beta_{m_2} max(0, t - c_m)$. This algorithm terminates once the coefficient of determination value (R^2) of the resulting model increases by less than 0.001. The Forward Pass is followed by a Backward Pass that "prunes" any insignificant basis functions using Generalised Cross Validation (GCV) which approximates leave-one-out validation and penalises for the number of parameters to prevent over-fitting. All terms in the model after the Forward Pass are removed one-by-one determined by which results in the greatest reduction in the RSS. The GCV is calculated for each sub-model created and is then used to pick the final model.

3 Results

From fitting the model to all NTIs it was evident that a change in trend at the beginning of 2020, which coincides with the start of the COVID-19 pan-

demic, was present in 5 main prescribing areas. Further investigations were carried out to try to hypothesize the cause of the change at the Scottish level. For example, the total number of antibiotics prescriptions dropped from an estimated decreasing monthly rate of -0.007 items per 1000 list size per day to a monthly decreasing rate of -0.019 items per 1000 list size per day, where list size refers to the number of registered patients. At the same time the percentage of patients with more than 4 antibiotic presriptions a year, out of all people prescribed an antibiotic, in Scotland increased. Hence, it could be speculated that the number of patients seeking antibiotics for minor infections was decreasing and reducing the total patients on antibiotics whilst the number of patients needing multiple courses of antibiotics for more serve infections was constant or possibly increasing. This may be due to access to health care services being transitioned to remote consultations and hospital outpatients being mostly closed except for emergency cases. In addition, strong media messaging about the strain the NHS was under and personal anxietes about becoming infected may have reduced the number of people seeking care for conditions besides COVID-19.

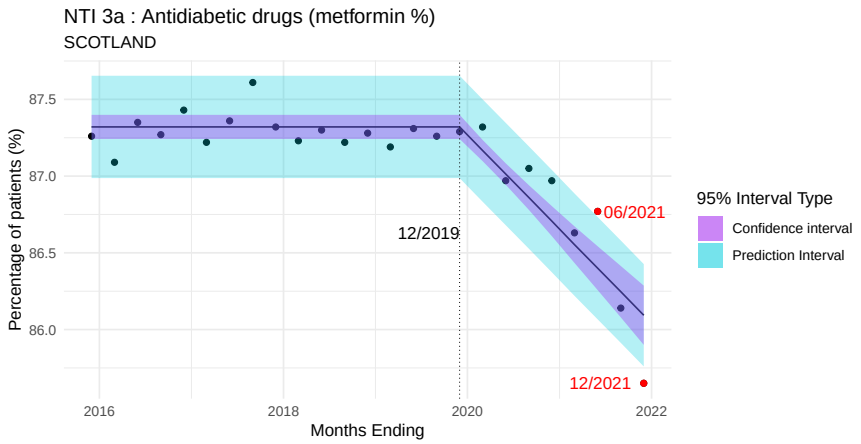


FIGURE 1. Quartely percentages of patients receiving metformin out of all patients prescribed anti-diabetic drugs in Scotland. The solid black line indicates the estimated mean from the piecewise linear regression with the corresponding 95% confidence and prediction intervals. Red data points and dates indicate possible outliers, the vertical dotted line and date indicate the model knot location.

Similar reasoning could be behind the impact on the prescribing of the anti-diabetic drug metformin, shown as a percentage of all people prescribed an anti-diabetic drug in Figure 1. Evidently, there is a Scottish wide downward trend in the percentage of patients being prescribed metformin from 2020 onwards after a relatively stable rate. According to the National Thera-

peutic Indicators 2018 report published by the Scottish Government and NHS Scotland, metformin is the recommended first prescription for type-2 diabetes due to it's proven survival advantage. As reported by Carr, M. J. et al (2021) this reflects the fall in type-2 diabetes diagnoses in the UK when compared to historical trends resulting in an estimated 60 000 new diagnoses being missed or delayed between March and December 2020. They also note that during the same time period mortality rates in the UK for people with type-2 diabetes increased by approximately 0.13. Similarly, a global survey conducted by Chudasama, Y. V. et al (2020) found that healthcare professionals thought that the reduction in health care resources caused by the pandemic impacted diabetes the most out of all areas of chronic disease. At the start of the pandemic many NHS services experienced an increased burden due to high levels of COVID-19 patients, shortages in personal protective equipment (PPE) as well as staff absences due to COVID-19 infections.

4 Summary

This analysis has used a MARS approach to investigate the COVID-19 impacts on piecewise linear regression models of the NTIs for primary care in Scotland. Whilst this approach was appropriate for the majority of the indicators some may have benefitted from a disjoint piecewise regression to account for more extreme jumps between trends in the data. Therefore, future work may involve developing a framework that can determine whether a discontinuity is present in the data during the model fitting process. However, this approach has allowed for the easy identification of regions and time points where significant changes in NTI prescribing data has occurred. The monitoring of affected NTIs can help asses the repercussion of the COVID-19 pandemic by evaluating if prescribing trends return to their previous state as the pandemic becomes less influential on health care delivery, social distance restrictions and the burden on NHS services. Finally, a dashboard to display the results of the modelling carried out in this analysis, which includes models where extreme outliers have been removed and yealry seasonal cycles have been taken into account, has been developed. This tool can be updated as new data becomes available for internal use at PHS. It is hoped this will aid analysts in carrying out investigation of a similar manner and make it a less time consuming task despite the large amount of data.

Acknowledgments: Special Thanks to Gavin MacColl, Raul Barrocal-Martin, Rita Nogueira and Stuart McTaggart at Public Health Scotland for supervising this project.

References

Carr, M. J., Wright, A. K., Leelarathna, L., Thabit, H., Milne, N., Kanumilli, N., Ashcroft D. M. and Rutter M. K. (2021). Impact of COVID-19 on diagnoses, monitoring, and mortality in people with type 2 diabetes in the UK. *Lancet Diabetes Endocrinol*, **9** (7), 413 – 415.

Chudasama, Y. V., Gillies, C. L., Zaccardi, F., Coles, B., Davies, M. J., Seidu, S. and Khunti, K. (2020). Impact of COVID-19 on routine care for chronic diseases: A global survey of views from healthcare professionals. *Diabetes & metabolic syndrome*, **14** (5), 965 – 967.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *Ann. Statist.*, **58** (19), 1 – 67.

MacBride-Stewart, S., Guthrie, B., Marwick, C. and Hurding, S. (2019). National Therapeutic Indicators in Scotland and Financial Incentives *International Journal of Population Data Science*, **4** (3).

MacBride-Stewart, S., Marwick, C., Houston, N., Watt, I., Patton, A. and Guthrie, B. (2017). Evaluation of a complex intervention to improve primary care prescribing: a phase IV segmented regression interrupted time series analysis. *Br J Gen Pract*, **67** (658), 352 – 360.

Medicines and Healthcare products Regulatory Agency (2017). Valproate and developmental disorders: new alert asking for patient review and further consideration of risk minimisation measures. Available at: *https://www.gov.uk/drug-safety-update/valproate-and-developmental-disorders-new-alert-asking-for-patient-review-and-further-consideration-of-risk-minimisation-measures* [Accessed: 23rd May 2022]

Milborrow, S. Derived from mda:mars by Hastie, T. and Tibshirani, R. (2011). earth: Multivariate Adaptive Regression Splines. *R Package*.

A Bayesian hierarchical model for improving exercise rehabilitation in mechanically ventilated ICU patients.

Luke Hardcastle[1], Samuel Livingstone[1], Claire Black[2], Federico Ricciardi[3], Gianluca Baio[1]

[1] Department of Statistical Science, University College London, London, UK
[2] University College London Hospitals NHS Foundation Trust, London, UK
[3] Owlstone Medical, Cambridge, UK

E-mail for correspondence: Luke.Hardcastle.20@ucl.ac.uk

Abstract: Patients who are mechanically ventilated in the intensive care unit (ICU) participate in exercise as a component of their rehabilitation to ameliorate the long-term impact of critical illness on their physical function. The effective implementation of these programmes is hindered, however, by the lack of a scientific method for quantifying an individual patient's exercise intensity level in real time, which results in a broad one-size-fits-all approach to rehabilitation and suboptimal patient outcomes. In this work we have developed a Bayesian hierarchical model with temporally correlated latent Gaussian processes to predict $\dot{V}O_2$, a physiological measure of exercise intensity, using readily available physiological data. For practical use by clinicians $\dot{V}O_2$ was classified into exercise intensity categories. Internal validation using leave-one-patient-out cross-validation was conducted based on these classifications, and the role of probabilistic statements describing the classification uncertainty was investigated.

Keywords: Bayesian Hierarchical Model, INLA, Exercise Rehabilitation

1 Introduction

Patients who are mechanically ventilated in the intensive care unit (ICU) as a result of critical illness are often left with a range of impairments, due to the pathological effects of critical illness and its treatments on nerve, muscle, cardiac and respiratory function, Guarneri et al. (2008).
Rehabilitation, while the patient is still receiving mechanical ventilation in the ICU, involves progressing patients through various simple exercises,

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

and is considered the best way to ameliorate the impact of critical illness and its associated treatments on physical function.
The current approach to rehabilitation is a one-size-fits-all approach based on the assumption that the metabolic cost of individual rehabilitation activities does not differ across patients. Recent work has shown this assumption to be invalid, however, suggesting that individuals with different physiological profiles receive similar, and often sub-optimal, exercise programmes.
To address these issues, a scientific method for quantifying the exercise intensity level of an individual patient in real time is required.
Exercise load during rehabilitation can be quantified by measuring a patient's rate of oxygen consumption ($\dot{V}O_2$), but these measurements are not available in many intensive care units.
The primary contribution of this work is the development and internal validation of a first of its kind prediction model of $\dot{V}O_2$ for mechanically ventilated intensive care patients, providing clinicians with real-time predictions of patients' levels of absolute exercise intensity, allowing tailored exercise rehabilitation plans to be implemented.

2 Data and exploratory analysis

The data are from the observational study conducted by Black et al. (2020), which took measurements as mechanically ventilated ICU patients participated in various rehabilitation activities. Measurements were recorded on a breath by breath basis, resulting in high frequency data consisting of 74,332 measurements from 37 patients and 103 rehabilitation sessions. They are hierarchical in nature with repeated measurements within sessions and often multiple sessions per patient. The data contained measurements of $\dot{V}O_2$, physiological covariates typically available in ICU - tidal volume (V_T), respiratory rate (RR) and partial pressure of end tidal CO_2 ($P_{ET}CO_2$) which have known relationships with $\dot{V}O_2$ - and patients' baseline characteristics. The primary findings of the exploratory analysis were strong linear relationships between $\dot{V}O_2$ and V_T when placed on the log-log scale, as well as interactions between $\log(V_T)$ and both $\log(RR)$ and $\log(P_{ET}CO_2)$.

3 Model and inference

Model development was undertaken within the Bayesian paradigm. In practice predictions of $\dot{V}O_2$ would likely be presented to clinicians in the form of classifications; the Bayesian approach allows for uncertainty around $\dot{V}O_2$ predictions to be propagated through the model, and used to quantify uncertainty about classifications. Future work is needed to determine these classification categories, however, and in practice different categories may be used in different contexts. This modelling approach allows us to adapt the model to any classification task that may arise.

In order to account for the hierarchical and temporal nature of the data we have developed a Bayesian hierarchical model for $\dot{V}O_2$ with temporally correlated latent Gaussian processes. We indicate the value of $\log(\dot{V}O_2)$ taken at time t for the i^{th} patient's j^{th} rehabilitation session as y_{ijt} and assume that

$$y_{ijt} \mid \mu_{ijt}, \tau \sim \text{Normal}(\mu_{ijt}, \tau^{-1}),$$

where μ_{ijt} is a linear predictor consisting of the physiological covariates and baseline characteristics identified in the exploratory analysis, a session level varying intercept term, a patient level varying coefficient for the effect of $\log(V_T)$, and a temporal error term characterised using an Ornstein-Uhlenbeck process, Øksendal (2003). The model was fitted using the Integrated Nested Laplace Approximation (INLA), with the `R-INLA` package, Rue et al. (2017).

4 Results

The model was fitted to both the raw data and data smoothed using a three-value rolling average in order to limit the impact of spikes in $\dot{V}O_2$ values induced by patients coughing. Posteriors for the varying intercept and varying coefficient terms revealed a high level of between patient and between session heterogeneity.

To assess the predictive performance of the model we used cross-validation, leaving observations out at the patient level to account for similarity between observations within patients and prevent leakage. This was separately performed for the models using raw and smoothed data.

To assess the model as it would be used in practice, $\dot{V}O_2$ was classified into rest, low, medium and high categories of exercise intensity. Accuracy for these classifications was 60.0% and 61.1% for the raw and smoothed data models respectively, however the raw data model had far higher accuracy for observations classified as high compared to the smooth data model (74% vs 56%). This is of particular clinical relevance as it indicates when a patient's exercise load needs to be reduced.

We note that for almost all rehabilitation sessions the model does a remarkable job of matching the shape of the $\dot{V}O_2$ curve over time (Figure 1). As the plot for session 109 indicates, however, for some sessions it is unable to quantify the scale of the curve, resulting in inaccurate classifications.

A key strength of the model is that it returns probabilistic statements about classifications, providing important information to clinicians. Using these statements and a suitable decision rule, we found that the model correctly directed the clinician to lower the exercise load 88.0% and 79.1% of the time for the raw and smoothed data models respectively, when the true value of $\dot{V}O_2$ indicated a high level of exercise intensity.

The model presented here is a first of its kind prediction model for exercise intensity in ventilated intensive care patients, using Bayesian hierarchical

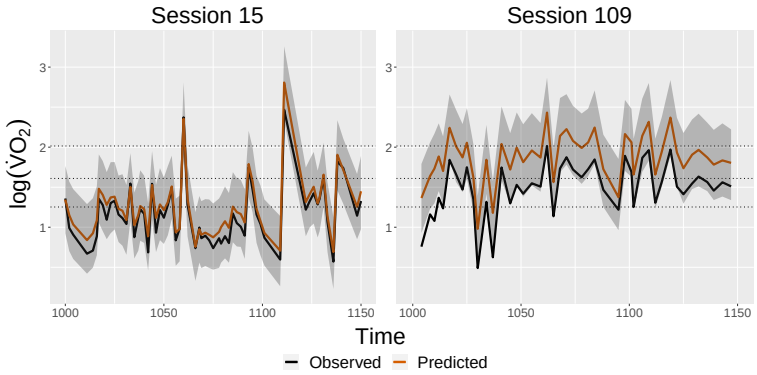


FIGURE 1. Compares predicted (with 95% credible intervals) and observed values of $\dot{V}O_2$ for two example sessions, one where the model performs well (session 15) and one where it performs poorly (session 109).

modelling and covariates that are readily available in the majority of intensive care settings. If externally validated this model could be the first step towards patients receiving personalised exercise rehabilitation regimes and significant improvements in post-ICU outcomes.

Acknowledgments: LH acknowledges support from EPSRC grant number EP/W523835/1 during part of this work. The authors would like to thank the UCL CHIMERA group, and in particular Prof. Christina Pagel and Prof. Rebecca Shipley for invaluable comments and feedback.

References

Black, C., Grocott, M.P.W., and Singer, M. (2020) The oxygen cost of rehabilitation interventions in mechanically ventilated patients: an observational study. *Physiotherapy (United Kingdom)*, **107**, 169 – 175

Guarneri, B., Bertolini, G., and Latronico, N. (2008) Long-term outcome in patients with critical illness myopathy or neuropathy: the Italian multicentre CRIMYNE study. *J Neurol.Neurosurg.Psychiatry*, **7**, 838 – 841

Øksendal, B.K. (2003). *Stochastic differential equations : an introduction with applications*. New York: Springer.

Rue, H., Riebler, A.,Sørbye S.H., Illian, J.B., Simpson, D., and Lindgren, F.K. (2017) Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application*, **4**, 395 – 421

Experimental Results and Comparisons of Semantically Enriched Query Alternatives in Information Retrieval Models

Sargon Hasso [1], Kenan M Matawie[2]

[1] Loyola University Chicago, UL, Northbrook, IL USA
[2] Western Sydney University, Australia

E-mail for correspondence: k.matawie@westernsydney.edu.au

Abstract: In this paper we examined how a semantically enriched query alternatives improve the score and rank of the search results in Information Retrieval Systems. This enrichment is statistically analysed and presented using TREC data and utilizing Solr full-text search platform. The improvement are measured using scoring functions such as BM25. Mean Average Precision (mAP) measure was used to compare different configurations of search engine using a scheme we designed for this purpose.

Keywords: Relevance; IR models; Semantic Enrichment; BM25; Repeated Measures; Post-Hoc test.

1 Introduction

The Information retrieval (IR) process consists of indexing corpus data that are analysed and indexed, and searching during which a user submits a query. The search engine retrieves documents where each term in the query is matched against the terms stored for each document in the collection. The search results, i.e. the retrieved documents, are ranked and returned in descending order. Traditionally, relevance process is based on keyword searching for all the terms in the query with little or no modifications. Techniques exist to augment the user's query with additional terms to maximize the relevance of search results. Search engines support semantic-based searching using language-based synonym, domain-specific, and custom thesauri. In this paper, we will discuss our findings using many techniques of semantic-based searching along with different ways to configure the search engine in the attempt answer this question: what factors

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

influence the relevance of retrieved documents and how to validate the differences statistically.
In section 2 we survey some of the related research. In section 3, we briefly explain the scoring, ranking, and the evaluation model. In section 4, we describe a methodology we used in this research. Section 5 describes our experiment. An analysis and evaluation of the test results are discussed in section 6. We provide summary in section 7.

2 Related Work

Zobel and Moffat (1998) had investigated the similarity measure, by decomposing it into eight orthogonal factors. By re-configuring this similarity measure in different ways and running document search against each configuration, they were able to observe the effects on the ranking of the retrieved results. Our experimental research focused on observing the effect of changing the search engine configurations on the ranking of the retrieved results. Buscaldi et al. (2014) uses search and similarity measure based on ontology. The goal of this research is to improve either recall or precision, the technique does require prepossessing content and modifying or changing the scoring function.

3 Scoring, Ranking and Evaluation Models

Lucene-based Solr Engine (Apache Software Foundation, 2021) was used to evaluate statistical information models with different indexing configurations. An efficient implementation of Lucene scoring evolved from the underlying information retrieval models to rank the relevance of matched documents to user's query (Apache Software Foundation, 2022). Most of the models are based on the maximum likelihood estimate of the relative counts. The best Match family (BM25) by Jones et al. (2000) will be the only one presented here as implemented by Lucene:

$$f(q,d) = \sum_{w \in q \bigcap d} c(w,q) \frac{(k+1)c(w,d)}{c(w,d) + k(1 - b + b\frac{|d|}{avdl})} log \frac{M+1}{df(w)} \tag{1}$$

where $b \in [0,1]$ is part of the *normalizer* term $1 - b + b\frac{|d|}{avdl}$; $avdl$ denotes average document length.
To evaluate the precision, Mean Precision Average (mPA) evaluation model is used.

$$mPA = \frac{1}{n} \sum_{k=1}^{n} AP_k \tag{2}$$

Where AP_k is the Average Precision for class k and n is the number of classes.

4 Query Expansion

Query expansion is a technique to improve relevancy of the result sets returned by search engine. The goal is to augment the original query submitted by users. In this research we used the following query expansion techniques:

Language-based Synonym Expansion : we augment the query at runtime with synonyms based on keywords in the original query. We used two types of synonyms: domain-specific, NASA Thesaurus [NASA, 2010] and generic English language-based thesaurus, [WordNet, 2010].

Content-based Synonym Expansion : we extract related concepts that tend to co-occur nearby in the corpus. We used word embeddings, also known as word2vec model, that allows us to find similar words that occur in the same context thus displaying their semantic affinity [Teofili, 2019].

Semantic Knowledge Graph (SKG) is a graph data structure that creates relationships between entities, e.g. terms, phrases, or extracted concepts built from a corpus of data, Grainger et al. (2016). as a way to find and rank terms that best match a query.

5 Experiments

We have developed a scheme by which a query is transformed into an alternative form, in addition to its original form, and combined with other factors to produce different configurations as input to Solr search engine and examined the results. The scheme can be illustrated as follows:
$q - c - f$, where q refers to query type, and can have any of these values:
$q = \{t, c, b, w\}$ the numbers refer to unmodified, concepts, concepts+booster, and similar concepts using word2vec, respectively.
c refers to how we configured Solr before each indexing phase.
$f = \{title\ only, title + content\}$ these are the fields that are used to search on. For 'title', we instruct Solr to search in the title of the document, and for 'title+content', we instruct Solr to search in both of these fields.

Solr search engine configurations Table will be inserted here showing all possible Solr configurations we ran our queries against. Figure 1 shows a scatter plot of all the test runs shwoing the Mean Average Precision (mAP) measure.

6 Analysis and Evaluation

The results of using unmodified, boosted, concept, and word2vec query types combined with solr configuration and target field search gave us results which can hardly be generalized to determine what combination of

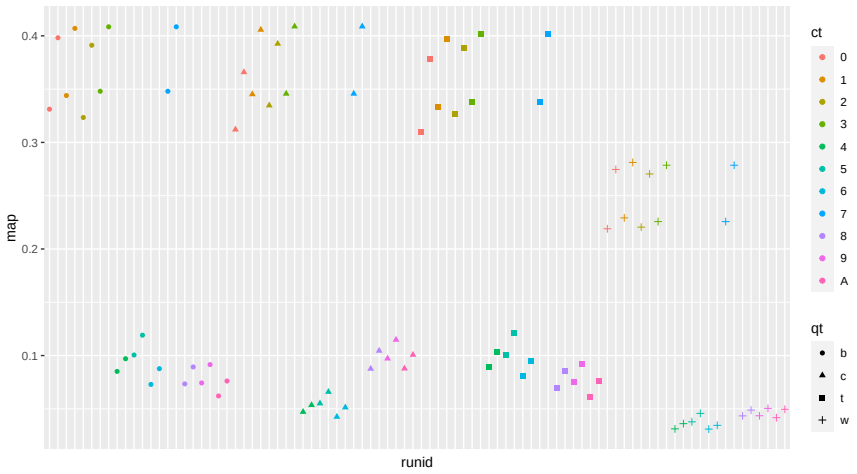


FIGURE 1. Summary results of all test runs

these configurations work best. It certainly showed what are the best performers relative to this limited set of corpus data.
We used Two-Way repeated measures analysis to test the main and interaction effects between the three factors with 11, 4 and 2 levels, respectively (per our configuration scheme as described in section 4). All the results were highly significant confirming different results for different alternative semantic enrichment. To further investigate and identify the pattern of the differences we used Student-Newman-Keuls (SNK) Post-Hoc test on 74 pairwise comparisons. Post-Hoc results showed (as expected) a specific enrichment approach is more suitable than others for this data, however, in some cases more investigation and analysis are required to further confirm or refute the results obtained. For example, the NASA thesaurus still requires further refinement to be useful as a domain-specific thesaurus. Furthermore, the size of the TREC aviation corpus data we used for this experiment is not large enough to generate good word2vec models that give us a better content-based synonym configuration that influences the relevance score.

7 Conclusion

We have developed an experimental methodology that comprises a testing platform using Solr search engine to generate and evaluate the test results TREC's mAP evaluation measure. The Solr search engine was configured and queries were expanded using different techniques. We have observed difference in the results obtained. The Statistical analysis supported these results with highly significant differences($p < 0.01$).

References

Apache Software Foundation (2021) *The Apache Lucene*. `http://lucene.apache`. Accessed Feb 2022.

Apache Software Foundation (2022) *Class TFIDSimilarity*. `https://lucene.apache.org/core/9_1_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html`. Accessed March 2022.

Buscaldi, D., Bessagnet, M., Royer, A., and Sallaberry, C. (2014) *Using the Semantics of Texts for Information Retrieval: A Concept- and Domain Relation-Based Approach*. Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 257-266

Grainger, T., AlJadda, K., Korayem. M., and Smith, A. (2016) The Semantic Knowledge Graph: A compact, auto-generated model for real-time traversal and ranking of any relationship within a domain.

Jones, S. K., Walker, S. and Robertson, S. E. (2000) A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. In *Information Processing and Management*, pages 779—840.

NASA (2012) *NASA Thesaurus: Hierarchical Listing with Definitions.* Vol 1, NASA.

Teofili, T. (2019) *Deep Learning for Search*. Manning Publications Co.

Wordnet (2010) *About WordNet. WordNet.* Princeton University

Zobel, J. and Moffat, A. (1998) *Exploring the Similarity Space*. SIGIR Forum , Vol. 32, No. 1 ACM: New York, NY, USA p. 18–34

Regularization and Model Selection for Item-on-Item Regression

Aisouda Hoshiyar[1], Jan Gertheiss[1]

[1] School of Economics and Social Sciences, Helmut Schmidt University, Hamburg, Germany

E-mail for correspondence: aisouda.hoshiyar@hsu-hh.de

Abstract: We propose a novel strategy for model selection of ordinally scaled predictors in the cumulative logit model. The original group lasso is expanded by use of a difference penalty on neighboring dummy coefficients, thus taking into account the ordinal structure. We apply stability selection for error control and for choosing a proper amount of regularization for structure estimation. We consider a survey on consumers' perception and acceptance for sustainable use of boar tainted meat consisting of Likert-type items.

Keywords: Cumulative Logit; Group Lasso; Likert-Scale; Proprtional Odds Model; Stability Selection

1 Introduction

We investigate a survey concerning consumers' perception and acceptance of boiled sausages from strongly boar tainted meat, conducted by the Department of Animal Sciences, University of Göttingen (Meier-Dinkel et al., 2016). In order to be able to process highly tainted meat sustainably, we aim to detect the relevant factors for the overall liking of the 120 consumers regarding boiled sausages with a proportion of 50% tainted boar meat. The six ordinal predictors considered (*expected liking, appear, odour, flavour, texture, aftertaste*) and the ordinal response (*overall liking*) are measured on a 9-point scale (1 = dislike extremely to 9 = like extremely). Now, the effect of the ordinal covariates on the ordinal factor *overall liking* is to be investigated, which also requires a strategy for variable selection.
In the cumulative logit model, the link between observed variable Y and latent variable u is defined by the threshold mechanism: $Y_i = r \Longleftrightarrow \theta_{r-1} < u_i \leq \theta_r, r = 1, ..., c$, where $-\infty < \theta_0 < \theta_1 < ... < \theta_c = \infty$ are the ordered

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

thresholds. The latent variable is modeled as $u_i = -\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, i = 1, ..., n$, where $\boldsymbol{x}_i$ are observed ordinal covariates, $\boldsymbol{\beta}$ is the parameter vector and ϵ_i is an error variable with logistic distribution function. Now for variable selection, the logistic group lasso estimator $\boldsymbol{\beta}_\lambda$ (Meier et al., 2008; Yuan and Lin, 2006) can be extended to the class of cumulative logit models and is given by the minimizer of the function $l_\lambda(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} J_j(\boldsymbol{\beta}_j)$, where $l(\boldsymbol{\beta})$ is the log-likelihood of the cumulative logistic distribution. In order to take into account the ordinal structure of the predictors, we modify the usual L_2-norm by the first-order difference penalty functions

$$J_j(\boldsymbol{\beta}_j) = \sqrt{\left\{ \sum_{l=1}^{k_j} k_j(\beta_{jk} - \beta_{j,k-1})^2 \right\}},$$

with k_j being the number of levels (for each variable) and $\beta_{j0} = 0 \forall j$, as introduced by Gertheiss et al. (2011). To enhance the cumulative group lasso, we apply stability selection as suggested by Meinshausen and Bühlmann (2010), a promising subsampling strategy in combination with high dimensional variable selection. In general, instead of selecting/fitting one model, the data are pertubed/subsampled many times and we choose those variables that occur in a large fraction of runs. For every variable $\boldsymbol{x}_j, j = 1, ..., p$, the estimated probability $\hat{\pi}_j^\lambda$ of being in the stable selection set corresponds to the frequency of being chosen over all subsamples. Or, in other words, we keep variables with a high selection probability $\hat{\pi}_j^\lambda \geq \pi_{\mathrm{thr}}$ and neglect those with low selection probabilitiy. The cutoff value can therefore be seen as a tuning parameter and a typical choice is $\pi_{\mathrm{thr}} \in (0.6, 0.9)$.

2 Numerical Experiment

Before applying the proposed method to the boar taint data, we carry out a simulation study to investigate the properties of the proposed ordinal-on-ordinal selection with stability selection (OSSS) approach. In order to fit OSSS, samples of size $\lfloor n/2 \rfloor$ are drawn without replacement in each iteration. For comparison, we fit the model also without stability selection, which we call here ordinal rank selection (ORS). For ORS we assign ranks to variables depending on their importance, determined from the coefficient path. We generate $p = 20$ ordinally scaled variables, including 8 noise variables as follows: $x_1, ..., x_4$ are non-monotone, $x_5, ..., x_8$ are monotone but non-linear and the effect of $x_9, x_{10}, x_{13}, x_{14}$ is linear across categories. The remaining eight predictors are irrelevant, i.e., with effects being zero. Factor levels are randomly drawn from $\{1, ..., 5\}$, meaning that each covariate has the same number of levels. The (true) effects for some covariates are shown in Figure 1 (left). Using those predictors, we construct the ordinal (logistic) response of different sample sizes ($n = 200, 500, 1000$). To evaluate the selective performance, we construct the Receiver Operating Characteristic

(ROC) by/via varying selection thresholds and calculate the Area Under this Curve (AUC) in each iteration. Figure 1 (right) shows the performance in terms of AUC as obtained with a selection of different λ values and $n = 500$. The grey solid line corresponds to an AUC = 0.5, which would result from pure guessing. Comparing results of the different scenarios, it can be stated that, stability selection has the potential to markably improve selection results. It is seen that, when λ is varied within a reasonable range, the stable selection sets are quite insensitive to the choice of λ.

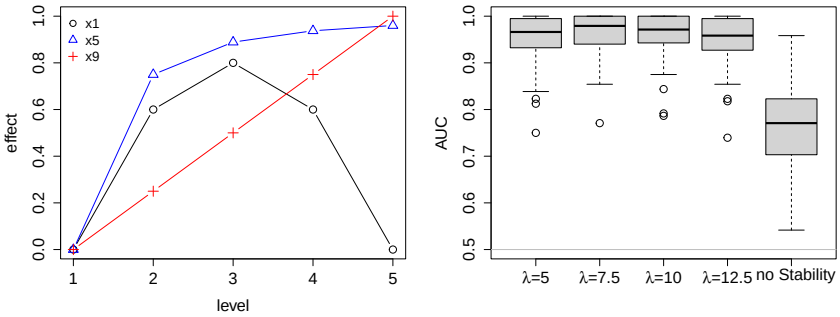


FIGURE 1. True effects of influential predictors x_1, x_5, x_9 (left). Boxplot of AUC for stability selection ($\lambda \in \{5, 7.5, 10, 12.5\}$) and without stability selection (right); calculations based on 100 simulated data sets.

3 Case Study: Sustainable Use of Boar Tainted Meat

Figure 2(a)–(d) shows the resulting estimates for relevant covariates and different values of tuning parameter λ when applying the proposed method to the boar taint data. It is seen that for smaller λ (light gray), the estimates are more wiggly and become more and more smoothed out/shrunken as λ increases. In general, the overall liking increases with the liking level of the covariates. The coefficient path as a function of λ can be found in Figure 2(e) and Figure 2(f) shows the stability path, indicating the order of relevance of the predictors according to stability selection. The probability to be selected within resampling is highest for *flavour*, *texture* and *appearance* (in descreasing order). It can be concluded that the stability path is potentially very useful for improved variable selection.
In summary, our preliminary results suggest that item-on-item regression with stability selection works well in the cumulative logit model with ordinally scaled predictors and ordinal group lasso penalty. Current research focusses on estimating item-on-item graphical models with stability selection in high dimensional settings.

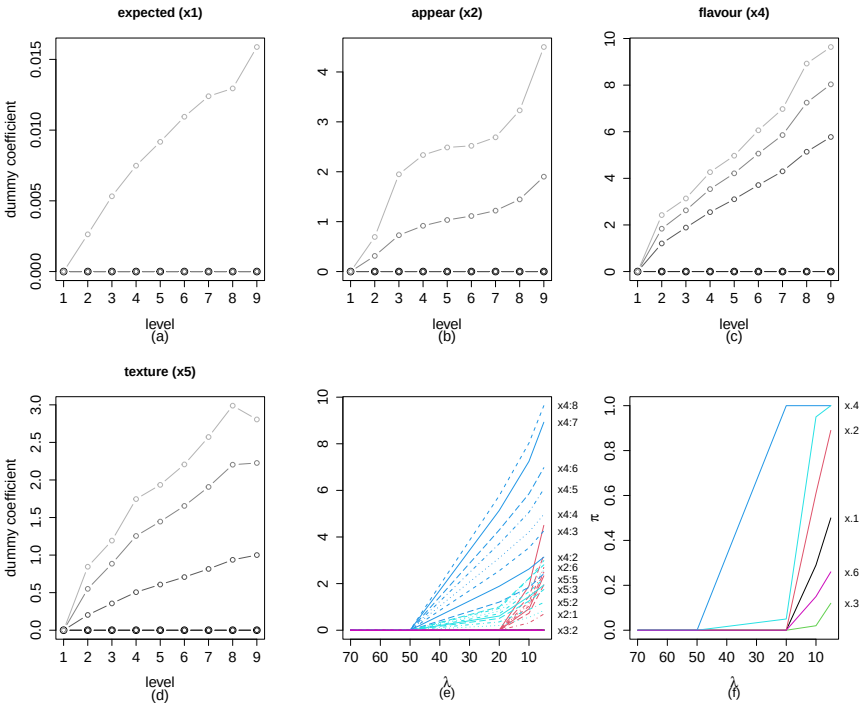


FIGURE 2. Cumulative group lasso estimates of dummy coefficients as functions of class labels ($\lambda \in \{70, 50, 20, 10, 5\}$) (a)–(d); Paths of cumulative group lasso estimates of dummy coefficients as functions of penalty parameter λ (e); Stability path of the cumulative group lasso (f).

References

Gertheiss, J., Hogger, S., Oberhauser, C., and Tutz, G. (2011). Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. *Journal of the Royal Statistical Society, Series C*, **60**, 377 – 395.

Meier, L., 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.

Meier-Dinkel, L., Gertheiss, J., Schnäckel, W. and Mörlein, D. (2016). Consumers' perception and acceptance of boiled and fermented sausages from strongly boar tainted meat. *Meat science*, **118**, 34 – 42.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, **72**, 417 – 473.

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, **68**, 49 – 67.

Directional test for the comparison of moderate-dimensional normal mean vectors

Caizhu Huang[1] and Nicola Sartori[1]

[1] Department of Statistical Sciences, University of Padova, Italy

E-mail for correspondence: caizhu.huang@phd.unipd.it

Abstract: Testing the equality of mean vectors of g independent groups of moderate-dimensional normal random variables with different unknown covariance matrices is a challenging problem. When $g = 2$ an approximation solution is given by the famous Behrens-Fisher test. For moderate-dimensional data where the dimension p of the observation vector is moderately large relative to the group sample sizes n_i, $i = 1, \ldots, g$, standard inferential approaches can be misleading. We present here a directional test and a modification of log-likelihood ratio test. The empirical evidence shows that the directional test improves over the alternative solutions and gives accurate inference, at least when $p = o(n_i^{\kappa})$ for some $\kappa \in (0.5, 1)$.

Keywords: Behrens-Fisher; Directional test; Higher-order asymptotics; Saddlepoint approximation.

1 Introduction

Mccormack et al. (2019) showed that when testing a specific value of the mean of a multivariate normal random variable the directional test (Fraser et al. 2016) coincides with the exact Hotelling's T^2. Exactness is seen to extend also when testing the equality of g means of independent groups of vectors with identical unknown covariance matrix, in the sense that the directional p-value is exactly uniformly distributed provided that $n_i \geq p + 1 + g$. Here we consider the more general case in which the g groups may have different unknown covariance matrices.
Directional inference on a vector parameter of interest is developed by Davison et al (2014) and Fraser et al. (2016) using saddlepoint approximations and one-dimensional numerical integration. Its accuracy is related to the accuracy of the saddlepoint approximation. Empirical results of Davison et

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

al (2014) and Fraser et al. (2016) showed that the directional test is extremely accurate even in regimes where p, however lower than n_i, increases with n_i.
Huang et al. (2021) showed that the directional test is indeed exact in some high-dimensional normal hypothesis testing problems. Here instead we consider a setting when exactness does not hold. Extended simulation studies show that the directional test has a higher accuracy than that of standard log-likelihood ratio and higher-order modifications proposed by Skovgaard (2001).

2 Comparison of normal mean vectors

Suppose we have independent $Y_{ij} \sim N_p(\mu_i, \Lambda_i^{-1})$, $i = 1, \ldots, g$ $(g \geq 2)$ and $j = 1, \ldots, n_i$, with unknown means μ_i and unknown positive definite covariance matrices Λ_i^{-1}. We are interested in testing the hypothesis

$$H_0 : \mu_1 = \cdots = \mu_g. \tag{1}$$

Let $\mathrm{tr}(A)$ denote the trace operator of a matrix A and $\mathrm{vech}(A)$ transform a matrix A into a vector by eliminating all supradiagonal elements of A. The log-likelihood for the parameter $\theta = \{\mu_1^{\mathrm{T}}, \ldots, \mu_g^{\mathrm{T}}, \mathrm{vech}(\Lambda_1)^{\mathrm{T}}, \ldots, \mathrm{vech}(\Lambda_g^{\mathrm{T}})\}^{\mathrm{T}}$ is

$$\ell(\theta) = \sum_{i=1}^{g} \frac{n_i}{2} \log|\Lambda_i| - \frac{n_i}{2}\mu_i^{\mathrm{T}}\Lambda_i\mu_i - \frac{1}{2}\mathrm{tr}(\Lambda_i y_i^{\mathrm{T}} y_i) + \frac{n_i}{2}\bar{y}_i^{\mathrm{T}}\Lambda_i\mu_i + \frac{n_i}{2}\mu_i^{\mathrm{T}}\Lambda_i\bar{y}_i,$$

where $y_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$. The maximum likelihood estimates are $\hat{\mu}_i = \bar{y}_i = 1_{n_i}^{\mathrm{T}} y_i / n_i$ and $\hat{\Lambda}_i^{-1} = y_i^{\mathrm{T}} y_i / n_i - \bar{y}_i \bar{y}_i^{\mathrm{T}}$, $i = 1, \ldots, g$, where 1_{n_i} is a n_i-dimensional vector of ones. The constrained maximum likelihood estimate under H_0 are denoted by $\tilde{\mu}_i = \tilde{\mu}$ and $\tilde{\Lambda}_i^{-1}$, where $\tilde{\Lambda}_i^{-1} = \hat{\Lambda}_i^{-1} + (\bar{y}_i - \tilde{\mu})(\bar{y}_i - \tilde{\mu})^{\mathrm{T}}$. We compute the constrained maximum likelihood estimate $\tilde{\mu}$ numerically by maximization of the profile log-likelihood $\ell_P(\mu) = -\sum_{i=1}^{g}(n_i/2)\log|\hat{\Lambda}_i^{-1} + (\bar{y}_i - \mu)(\bar{y}_i - \mu)^{\mathrm{T}}|$.
To develop the directional test under the null hypothesis (1), following Fraser et al. (2016) we consider the parameterization (ψ, λ) with the parameter of interest $\psi = (\mu_2^{\mathrm{T}} - \mu_1^{\mathrm{T}}, \ldots, \mu_g^{\mathrm{T}} - \mu_1^{\mathrm{T}})^{\mathrm{T}}$ and the nuisance parameter $\lambda = \{\mu_1^{\mathrm{T}}, \mathrm{vech}(\Lambda_1)^{\mathrm{T}}, \ldots, \mathrm{vech}(\Lambda_g)^{\mathrm{T}}\}^{\mathrm{T}}$. This parameterization places nonlinear constraints on canonical parameter φ. Thus, the tilted log-likelihood is $\ell(\varphi; t) = \sum_{i=1}^{g} \ell_i(\varphi_i; t)$ with $\varphi_i = \{\mu_i^T \Lambda_i, \mathrm{vech}(\Lambda_i)^T\}^T$ and the i-th group's contribution

$$\begin{aligned}\ell_i(\varphi_i; t) = &\frac{n_i}{2}\log|\Lambda_i| - \frac{n_i}{2}\mu_i^{\mathrm{T}}\Lambda_i\mu_i - \frac{1}{2}\mathrm{tr}\ \big(\Lambda_i \left[y_i^{\mathrm{T}} y_i + (1-t)\left\{n_i\tilde{\mu}(\bar{y}_i - \tilde{\mu})^{\mathrm{T}}\right.\right. \\ &\left.\left. + n_i(\bar{y}_i - \tilde{\mu})\tilde{\mu}^{\mathrm{T}}\right\}\right]) + n_i\left\{t\bar{y}_i + (1-t)\tilde{\mu}\right\}^{\mathrm{T}}\Lambda_i\mu_i.\end{aligned}$$

The maximum likelihood estimate along the line $s(t) = (1-t)s_\psi$ is $\hat{\mu}_i(t) = t\bar{y}_i + (1-t)\tilde{\mu}$ and $\hat{\Lambda}_i^{-1}(t) = \tilde{\Lambda}_i^{-1} - t^2(\bar{y}_i - \tilde{\mu})(\bar{y}_i - \tilde{\mu})^{\mathrm{T}}$, where the expected value s_ψ of the corresponding sufficient statistics s under H_0 has components $[n_i\tilde{\mu}^{\mathrm{T}} - n_i\bar{y}_i^{\mathrm{T}}, \frac{n_i}{2}\mathrm{vech}\{\tilde{\mu}(\bar{y}_i - \tilde{\mu})^{\mathrm{T}} + (\bar{y}_i - \tilde{\mu})\tilde{\mu}^{\mathrm{T}}\}^{\mathrm{T}}]$. The maximum likelihood estimate $\hat{\Lambda}_i^{-1}(t)$ exists if $t_{sup} = \min\limits_{1\le i\le g}\left[\{(\bar{y}_i - \tilde{\mu})^{\mathrm{T}}\tilde{\Lambda}_i(\bar{y}_i - \tilde{\mu})\}^{-1/2}\right]$ with $(\bar{y}_i - \tilde{\mu})^{\mathrm{T}}\tilde{\Lambda}_i(\bar{y}_i - \tilde{\mu}) \neq 0$. After some algebra, the directional p-value can be computed as

$$p(\psi) = \frac{\int_1^{t_{sup}} t^{d-1}h\{s(t);\psi\}dt}{\int_0^{t_{sup}} t^{d-1}h\{s(t);\psi\}dt},$$

with $h\{s(t);\psi\} = \prod_{i=1}^g |\hat{\Lambda}_i^{-1}(t)|^{\frac{n_i-p-2}{2}}|\tilde{C}_1(t)|^{1/2}$, and where $\tilde{C}_1(t) = \sum_{i=1}^k n_i\ \tilde{\Lambda}_i\left[I_p - t^2\left\{(p+1)I_p - \mathrm{tr}\ (\hat{\Lambda}_i^{-1}\tilde{\Lambda}_i)I_p - \hat{\Lambda}_i^{-1}\tilde{\Lambda}_i\right\}\right]$.
For comparison, we consider here the log-likelihood ratio test and its modifications proposed by Skovgaard (2001). The log-likelihood ratio test takes the form

$$W = \sum_{i=1}^g n_i\left(\log|\tilde{\Lambda}_i^{-1}| - \log|\hat{\Lambda}_i^{-1}|\right).$$

Under H_0, the statistic W has approximate chi-square distribution with degrees of freedom $d = p(k-1)$, when p is fixed.
Skovgaard (2001) proposes two modifications of W designed to maintain high accuracy in the tails of the distribution

$$W^* = W\left(1 - \frac{1}{W}\log\gamma\right)^2 \quad \text{and} \quad W^{**} = W - 2\log\gamma.$$

Both W^* and W^{**} have approximately a chi-square distribution with d degrees of freedom under H_0. In this specific case, the expression of the correction factor γ simplifies to

$$\gamma = \frac{\left\{\sum_{i=1}^g n_i(\tilde{\mu} - \bar{y}_i)^{\mathrm{T}}\tilde{\Lambda}_i(\tilde{\mu} - \bar{y}_i)\right\}^{d/2}}{W^{d/2-1}\sum_{i=1}^g n_i(\tilde{\mu} - \bar{y}_i)^{\mathrm{T}}\hat{\Lambda}_i(\tilde{\mu} - \bar{y}_i)}\left\{\prod_{i=1}^g\frac{|\hat{\Lambda}_i|}{|\tilde{\Lambda}_i|}\right\}^{\frac{p+2}{2}}\left\{\frac{|\tilde{C}|}{|\sum_{i=1}^g n_i\tilde{\Lambda}_i|},\right\}^{1/2},$$

where $\tilde{C}_1 = \sum_{i=1}^g n_i\tilde{\Lambda}_i\left\{\mathrm{tr}\ \left(\hat{\Lambda}_i^{-1}\tilde{\Lambda}_i - I_p\right)I_p + \hat{\Lambda}_i^{-1}\tilde{\Lambda}_i\right\}$.
When $g = 2$, this is a test for the multivariate Behrens-Fisher problem. Therefore, we also compare the directional test with the procedure proposed by Nel and Van der Merwe (1986) since it was shown to have near optimal power while maintaining reasonable type I error. The test by Nel and Van der Merwe (1986) is computed using the formula

$$T^{*2} = (\bar{y}_1 - \bar{y}_2)^{\mathrm{T}}\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1}(\bar{y}_1 - \bar{y}_2)$$

TABLE 1. Empirical type I error for DT, LRT, Sko1, Sko2 and/or BF with $n_i = 100$ at the nominal level $\alpha = 0.05$, with $p = n^\kappa$.

Σ_i	κ	$g = 2$					$g = 30$			
		DT	BF	LRT	Sko1	Sko2	DT	LRT	Sko1	Sko2
(I)	5/12	0.051	0.051	0.058	0.051	0.051	0.048	0.118	0.049	0.048
	6/12	0.050	0.050	0.059	0.050	0.050	0.048	0.188	0.050	0.048
	7/12	0.050	0.050	0.069	0.050	0.050	0.048	0.382	0.051	0.048
	8/12	0.048	0.048	0.079	0.049	0.048	0.049	0.754	0.060	0.052
	9/12	0.053	0.052	0.110	0.054	0.053	0.050	0.993	0.087	0.061
	10/12	0.046	0.045	0.180	0.050	0.047	0.054	1.000	0.223	0.099
(II)	5/12	0.051	0.053	0.059	0.051	0.051	0.049	0.117	0.049	0.048
	6/12	0.050	0.053	0.064	0.051	0.050	0.049	0.187	0.050	0.049
	7/12	0.050	0.058	0.074	0.051	0.051	0.048	0.382	0.051	0.049
	8/12	0.049	0.066	0.093	0.051	0.051	0.048	0.754	0.059	0.050
	9/12	0.052	0.081	0.144	0.057	0.054	0.050	0.993	0.086	0.059
	10/12	0.050	0.107	0.262	0.063	0.059	0.052	1.000	0.225	0.099
(III)	5/12	0.050	0.054	0.059	0.050	0.050	0.049	0.118	0.049	0.048
	6/12	0.053	0.058	0.065	0.053	0.053	0.048	0.187	0.049	0.048
	7/12	0.048	0.062	0.076	0.050	0.049	0.049	0.383	0.052	0.050
	8/12	0.049	0.075	0.104	0.051	0.049	0.049	0.755	0.060	0.051
	9/12	0.050	0.104	0.171	0.057	0.054	0.050	0.993	0.085	0.059
	10/12	0.049	0.154	0.345	0.071	0.062	0.053	1.000	0.225	0.099

where $S_i = \frac{n_i}{n_i - 1}\hat{\Lambda}_i^{-1}$, $i = 1, 2$. The statistic $\frac{\nu - p + 1}{p\nu} T^{*2}$ under H_0 has approximate F-distribution with degrees of freedom $(p, \nu - p + 1)$, i.e. $F(p, \nu - p + 1)$, where ν in the degrees of freedom is

$$\nu = \frac{\text{tr}\left\{\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)\right\} + \left\{\text{tr}\left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)\right\}^2}{\frac{\text{tr}\left\{\left(\frac{S_1}{n_1}\right)\left(\frac{S_1}{n_1}\right)\right\} + \left\{\text{tr}\left(\frac{S_1}{n_1}\right)\right\}^2}{n_1 - 1} + \frac{\text{tr}\left\{\left(\frac{S_2}{n_2}\right)\left(\frac{S_2}{n_2}\right)\right\} + \left\{\text{tr}\left(\frac{S_2}{n_2}\right)\right\}^2}{n_2 - 1}}.$$

See Rencher (1998, Section 3.9) for more details.

3 Simulation studies

The performance of the directional test for hypothesis (1) has been evaluated via Monte Carlo simulations based on 10,000 replications. The directional test (DT) is compared with the chi-square approximation for the log-likelihood ratio test (LRT) and two modifications of it proposed by Skovgaard (2001) (Sko1 and Sko2). When $g = 2$, we also considered the performance of the approximation of the Behrens-Fisher test T^{*2} (BF). The approaches are evaluated in terms of type I error and power.
Suppose that the data matrix y_i generated from a multivariate normal distribution $N_p(0_p, \Sigma_i), i = 1, \ldots, g$. Under the null hypothesis H_0, we set

TABLE 2. Empirical type I error for DT, BF, LRT, Sko1 and Sko2 with $n_i = 1000$ at the nominal level $\alpha = 0.05$. $\Sigma_i = \rho_i^{|j-l|}$.

κ	$\rho_1 = 0.1$ and $\rho_2 = 0.2$					$\rho_1 = 0.1$ and $\rho_2 = 0.9$				
	DT	BF	LRT	Sko1	Sko2	DT	BF	LRT	Sko1	Sko2
5/12	0.048	0.048	0.050	0.048	0.048	0.052	0.054	0.055	0.052	0.052
6/12	0.045	0.045	0.050	0.045	0.045	0.050	0.053	0.056	0.050	0.050
7/12	0.050	0.050	0.061	0.050	0.050	0.051	0.058	0.065	0.051	0.051
8/12	0.049	0.049	0.071	0.050	0.049	0.995	0.064	0.996	0.995	0.995
9/12	0.071	0.050	0.160	0.072	0.072	1.000	0.086	1.000	1.000	1.000
10/12	0.174	0.050	0.582	0.184	0.178	1.000	0.187	1.000	1.000	1.000

up $\mu_1 = \cdots = \mu_g = 0_p$ and consider three structure of the covariance matrices, i.e. (I) $\Sigma_i = I_p$, where I_p denotes a $p \times p$ identity matrix; (II) $\Sigma_i = (\sigma_{jl})_{p\times p} = \rho_i^{|j-l|}$ with ρ_i is a sequence from 0.1 to 0.9 with length g and increment $(0.9-0.1)/(g-1)$. In particular, when $g = 2$, $\Sigma_1 = (0.1)^{|j-l|}$ and $\Sigma_2 = (0.9)^{|j-l|}$; (III) $\Sigma_i = (1-\rho_i)I_p + \rho_i 1_p 1_p^{\mathrm{T}}$.

Under the local alternative hypothesis H_1, we set up $\mu_1 = 0_p$, $\mu_2 = \cdots = \mu_g = \delta/\sqrt{n_i p}1_p$ and $\Sigma_1 = I_p$, $\Sigma_2 = \cdots = \Sigma_g = \Sigma_1 + \delta/\sqrt{n_i p}I_p$. We consider the sample size $n_i = 100$ and the dimension $p = \lceil n^\kappa \rceil$ with $\kappa = (5/12, 6/12, 7/12, 8/12, 9/12, 10/12)$. Here we consider two choices of the number of groups $g = 2, 30$. Additional simulation studies with larger sample sizes have also been considered.

Table 1 reports the empirical type I error at the nominal level $\alpha = 0.05$ under the null hypothesis. The directional p-value is more accurate than its competitors in terms of the empirical type I error in this small sample setting. At the same time, BF is not stable in the setting with increasing correlation. In addition, Sko1 and Sko2 perform well when κ or g is small, while LRT is not accurate even for small κ and g.

With larger sample sizes the results in Table 2 for $g = 2$ indicate that when increasing the sample size all tests will not be valid when p is larger than n^κ, for some points of κ, although a larger κ seems to apply for the directional test. The performance in terms of type I error also depends on the true underlying structure of covariance matrices.

We also investigate the performance in terms of local power of the tests. We report the results of corrected power which is based on the corrected type I error, i.e. the 5% quantile of empirical p-value under H_0. Figure 1 shows that all tests have similar local power with the setting $g = 2$. However, when increasing the number of groups, DT, Sko1 and Sko2 are more powerful than that of LRT.

Acknowledgments: The author Caizhu Huang thanks Guangzhou Uni-

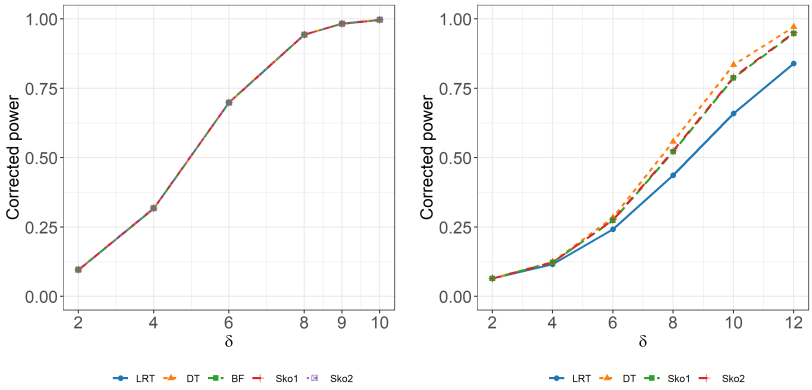


FIGURE 1. Empirical power functions of tests for hypothesis (1) under local alternative, for different values of δ and fixed $\kappa = 7/12$. The left and right panels correspond to the result with $g = 2$ and $g = 30$, respectively.

versity - University of Padova joint PhD program.

References

Davison, A. C., Fraser, D. A. S., Reid, N. and Sartori, N. (2014). Accurate directional inference for vector parameters in linear exponential families. *J. Amer. Statist. Assoc.*, **109**, 302 – 314.

Fraser, D. A. S., Reid, N. and Sartori, N. (2016). Accurate directional inference for vector parameters. *Biometrika*, **103**, 625 – 639.

Huang, C., Di Caterina, C. and Sartori, N. (2021). Directional testing for high-dimensional multivariate normal distributions. *arXiv:2107.09418*.

Mccormack, A., Reid, N., Sartori, N. and Theivendran, S. A. (2019). A directional look at F-tests. *Canad. J. Statist.*, **47**, 619 – 627.

Nel, D. and Van der Merwe, C. (1986). A solution to the multivariate Behrens-Fisher problem. *Comm. Statist. Theory Methods*, **15**, 3719 – 3735.

Rencher, A. C. (1998). *Multivariate statistical inference and applications*. New York: Wiley.

Skovgaard, I. (2001). Likelihood asymptotics. *Scand. J. Stat.*, **28**, 3 – 32.

On a genetically modified mode jumping MCMC approach for multivariate fractional polynomials

Aliaksandr Hubin[1], Riccardo De Bin[1]

[1] University of Oslo, Norway

E-mail for correspondence: aliaksah@math.uio.no

Abstract: In this work, we suggest a framework to fit fractional polynomials based on the Bayesian Generalized Nonlinear Models (BGNLM, Hubin et al, 2021). A version of the Genetically Modified Mode Jumping Markov Chain Monte Carlo (GMJMCMC) algorithm (Hubin et al, 2020) is adopted. Preliminary simulation runs show promising results in terms of identifying the data generation mechanism: The suggested approach uniformly outperforms the existing Bayesian fractional polynomial framework of Sabanés Bové and Held (2011) both in terms of Power and false discovery rate (FDR). Also, the performance is on par (somewhat better) with that of frequentist fractional polynomials of Royston and Altman (1994). Still, the results indicate that further work on the priors is likely to improve the performance even further.

Keywords: Bayesian model selection; MCMC; Fractional Polynomials.

1 Model and inference

Fractional polynomials (FP) were introduced by Royston and Altman (1994) for nonlinear regression modelling. A transformation of each covariate, from a set of possible functions, which includes the identity ($\mathbf{F}_0 = \{x\}$), 7 simple functions ($\mathbf{F}_1 = \{x^{-2}, x^{-1}, x^{-0.5}, \log x, x^{0.5}, x^2, x^3\}$), and 8 interactions based functions ($\mathbf{F}_2 = \{x^{-2}\log x,\ x^{-1}\log x, x^{-0.5}\log x, \log x \log x, x^{0.5}\log x,\ x\log x,\ x^2\log x, x^3\log x\}$), is performed and the transformed variables are added to the linear model. Let D be indexes of the set $\{\mathbf{F}_0 \bigcup \mathbf{F}_1 \bigcup \mathbf{F}_2\}$. Furthermore, let a function $\rho_k(x) = \{\mathbf{F}_0 \bigcup \mathbf{F}_1 \bigcup \mathbf{F}_2\}_k$, $k \in \mathcal{D}$ of the input covariate x be called *a polynomial term* (PT). In contrast to existing methods, such as Bayesian fractional polynomials (BFP) of

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

Sabanés Bové and Held (2011), we study fractional polynomial regression in the context of BGNLM (Hubin et al, 2021),

$$Y \sim \mathfrak{f}(y|\mu(\boldsymbol{X});\phi) \quad (1)$$

$$h(\mu(\boldsymbol{X})) = \alpha + \sum_{j=1}^{m}\sum_{k\in\mathcal{D}} \gamma_{jk}\beta_{jk}\rho_k(x_j), \quad (2)$$

where $\mathfrak{f}$ denotes the parametric distribution of Y belonging to the exponential family with mean $\mu(\boldsymbol{X})$ and dispersion parameter ϕ. The function h is a link function, α and $\beta_{jk}, j \in \{1,...,m\}, k \in \mathcal{D}$ are unknown parameters, and γ_{jk} is the indicator variable which specifies whether the PT $\rho_k(x_j)$ is included in the model. The vector $M = \{\gamma_{jk}, j \in \{1,...,m\}, k \in \mathcal{D}\}$ fully characterizes a model in terms of which PTs are included. We define the prior for M by

$$P(M) \propto \mathbb{I}(|M| \leq q) \prod_{j=1}^{m}\prod_{k\in\mathcal{D}} a_k^{\gamma_{jk}}. \quad (3)$$

Here, $|M| = \sum_{j=1}^{m}\sum_{k\in\mathcal{D}}\gamma_{jk}$ is the number of PTs in the model and q is the maximum number of PTs allowed per model. The factors $a_k^{\gamma_{jk}}, 0 < a_k < 1$ are prior penalties of including individual PTs. We then consider standard Jeffreys priors (Jeffreys, 1946) as in Hubin et al (2021). For inference, we adopt the GMJMCMC algorithm from Hubin et al (2021). As FPs are a specific case of BGNLM with only *modification* transformations allowed, the algorithm is simplified by setting $P_c = 0$ and $P_p = 0$.

2 Simulation results

For evaluating the performance of the suggested approach, a modification of the ART study (Royston and Sauerbrei, 2008) is proposed: We change the original data generative model by including an $x_1^{0.5} + x_1$ effect instead of $x_1^{-0.2}$ and by modifying the effect of x_3 to $x_3^{-0.5} + x_3^{-0.5} * \log(x_3 + \varepsilon)$, making the problem more challenging. The true model is

$$y = x_1^{0.5} + x_1 + x_3^{-0.5} + x_3^{-0.5} * \log(x_3 + \varepsilon) + x_{4a} + x_5^{-1} + \log(x_6 + \varepsilon) + x_8 + x_{10} + \epsilon,$$

where x_{4a} denotes the second level of x_4 and $\varepsilon = 0.00001$ is a small positive number required for numerical stability. Also, we assume $\epsilon \sim N(0, \sigma^2)$ and run 9 different scenarios with $\sigma^2 \in \{6.25e-1, 1.25e-1, 2.50e-2, 5.00e-3, 1.00e-3, 2.00e-4, 1.00e-4, 2.00e-5, 1.00e-7\}$ allowing to quantify the performance under different signal to noise ratios. For the model priors, we used $q = 20$ and $p = 20$. Further, a_k was chosen such that $\log a_k = -\log n$ for $\rho_k \in \mathbf{F}_0$, $\log a_k = -(1 + \log 2)\log n$ for $\rho_k \in \mathbf{F}_1$, and $\log a_k = -(1 + \log 4)\log n$ for $\rho_k \in \mathbf{F}_2$.

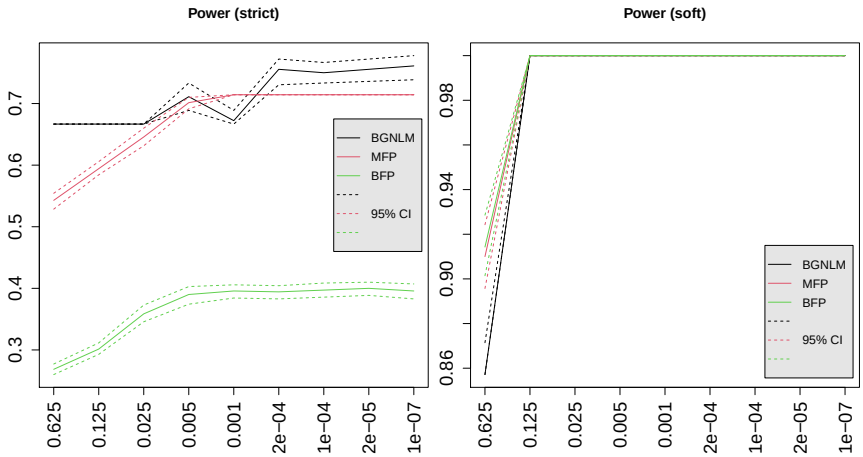


FIGURE 1. Left panel: Overall Power estimates for the data generative PTs for BGNLM (black), MFP (red), and BFP (green). Right panel: Overall Power estimates for detecting any function of the data generative covariates by BGNLM (black), MFP (red), and BFP (green).

BGNLM was fitted by GMJMCMC using the `EMJMCMC` package available at ***http://aliaksah.github.io/EMJMCMC2016/***. The simulations for each σ^2 were run on 32 parallel threads 100 times. Each thread was run for $20,000$ iterations with a mutation rate of 250 and the last mutation at iteration $15,000$. The population size of GMJMCMC algorithm was set to 20. For detection of PTs, the median probability rule was used (Barbieri and Berger, 2004).

For comparison, the frequentist version of multivariate fractional polynomials (MPF), was run using the R package `mfp` by Heinze et al (2021). We allowed for fractional polynomials of maximal order 2, and used a significance level $\alpha = 0.05$. Further, BFP of Sabanés Bové and Held (2011) was run using the R package `bfp` by Sabanés Bové et al (2022) with flat priors (hyperparameter for hyper-g prior equal to 4) and sampling to explore the posterior model space. Also, in this case, the maximal allowed order of the fractional polynomials was set to 2. For BFP, the median probability model (Barbieri and Berger, 2004) was also used for the detection of PTs.

Then, for each case for the three compared methods (BGNLM, MFP, BFP), 100 repetitions of 20 simulations were sampled with replacement to bootstrap the medians and 95% confidence intervals of the evaluation metrics. These metrics include Power (overall PTs, for individual PTs - *strict*, and for individual covariates - *soft*), FDR (w.r.t. individual PTs - *strict* and w.r.t. individual covariates - *soft*). Also for the BGNLM approach, the best found marginal posteriors were evaluated.

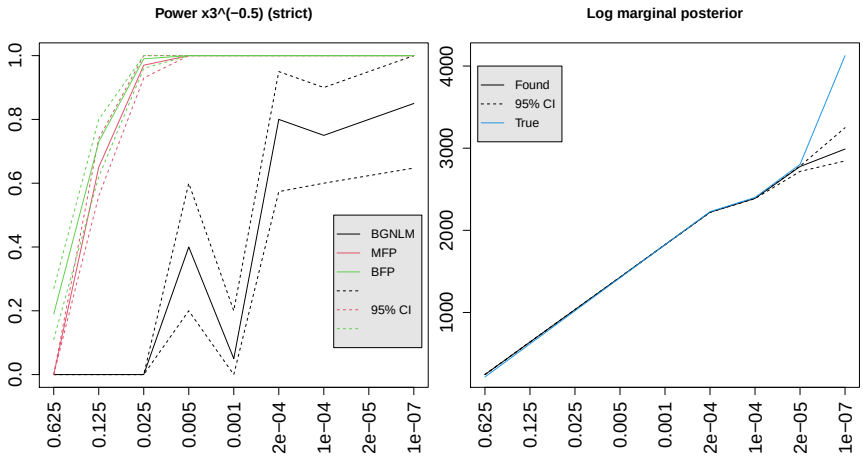


FIGURE 2. Left panel: Individual power estimates for $x_3^{-0.5}$ detected by BGNLM (black), MFP (red), and BFP (green). Right panel: Best log marginal posteriors found with GMJMCMC (black) and those of a data generative model (blue).

In Figure 1, we see that the overall Power of detecting both the data generative TPs and any functions of the data generative covariates grow as we increase the signal to noise ratio for all three compared methods. For the *strict* definition of Power, BGNLM and MFP are uniformly outperforming BFP. And for both the strong and the weak signal, BGNLM outperforms significantly MFP. For the medium signal, they are on par except for the settings of $\sigma^2 = 0.001$, for which MFP performs better. Overall, in 6 of 9 settings, BGNLM is significantly better, in 2 BGNLM and MFP are on par with each other, and in 1 MFP is significantly better than BGNLM. For the *soft* definition of Power counting detection of any function of the true covariates as a true positive, we quickly (at $\sigma^2 = 0.125$) reach 1 for all the three methods. For $\sigma^2 = 0.625$, MFP and BFP with a median Power of above 90% outperform significantly BGNLM with a median Power of just above 85%. For BGNLM, the reason for both the *strict* and *soft* definitions of Power to grow is the ability to recover $x_3^{-0.5}$, which is depicted in Figure 2. This particular PT is significantly better detected by MFP and BFP than by BGNLM. At the same time, the term $x_3^{-0.5} * \log(x_3 + \varepsilon)$ is never detected by any of the methods. And the term $x_1^{0.5}$ is never detected by BGNLM (all other data generative PTs are *always* recovered by BGNLM). MFP and BFP have their limited Power due to not always detecting $x_1^{0.5}$ and some other data-generative PTs. Remarkably, for $x_3^{-0.5} * \log(x_3 + \varepsilon)$ and $x_1^{0.5}$ the detection is very challenging since other similar terms are present in the data generative process: For $x_3^{-0.5} * \log(x_3 + \varepsilon)$, we have $x_3^{-0.5}$ with a correlation of 0.9379. For $x_1^{0.5}$, we have x_1 with a correlation of 0.9978.

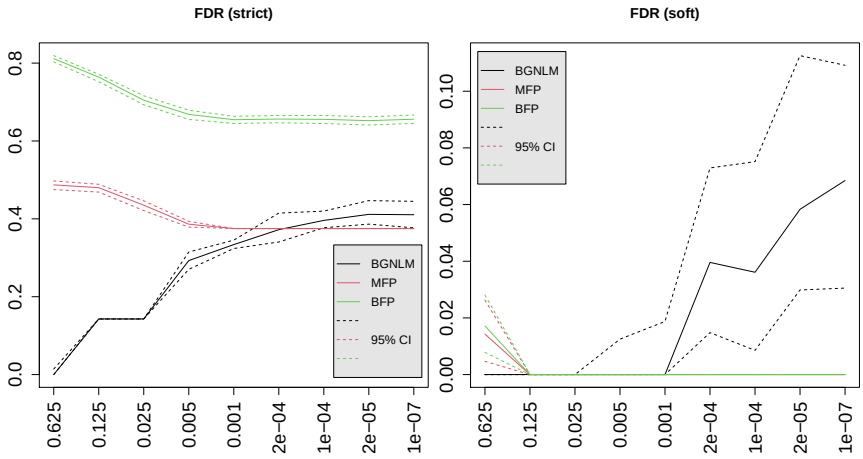


FIGURE 3. Left panel: overall FDR estimates by BGNLM (black), MFP (red), and BFP (green) when only allowing for data generative PTs to be counted as true positives. Right panel: overall FDR estimates by BGNLM (black), MFP (red), and BFP (green) when allowing any function of the data generative covariates to be counted as true positives.

For both the *strict* and the *soft* definition of FDR depicted in Figure 3, we see that it decreases with the increased signal for MFP and BFP, but increases for BGNLM. The latter is a bit unexpected and it likely happens due to multiplicities and a huge number of highly correlated PT to the data generative ones. As the signal to noise ratio increases, the likelihood part of the posterior becomes more important paying off for potentially including these terms. For the log marginal posteriors depicted in the right panel of Figure 2, we still see that for all signal to noise ratios except the highest one, models similar to or better than the data generative one were discovered. These two facts indicate that the prior inclusion probability should decrease as the signal to noise ratio increases in BGNLM. Alternatively, a hyper g-prior for the parameters from BFP should be studied within the context of BGNLM. Having said that, BGNLM is still uniformly outperforming BFP in terms of the *strict* definition of FDR. In this sense, BGNLM is also significantly better than MFP for 5 settings out of 9 corresponding to those with the strongest noise. For the three largest signal to noise ratios, however, MFP is doing better. And for $\sigma^2 = 2e-04$, BGNLM and MFP perform similarly. In terms of the *soft* definition of FDR, we see that BFP and MFP only detect wrong PTs of the true covariates for all of the noise levels below 0.625. At the same time, BGNLM starts to recover a few TPs of the wrong covariates for the strongest signal to noise ratios indicating once again that the model and/or parameter priors should be more conservative.

3 Discussion

In this paper, we studied how BGNLM fitted by GMJMCMC introduced in Hubin et al (2021) can deal with fractional polynomials. The approach was then compared to the existing implementation of MFP by Heinze et al (2022) and the implementation of BFP by Sabanés Bové et al (2022). The simulation study shows promising results of BGNLM. It uniformly outperforms BFP in terms of a strict definition of Power and FDR. Also, the performance is on par with that of MFP. At the same time, we see a strong indication that further work on the priors is likely to improve the performance even further.

References

Barbieri, M.M. and Berger, J.O. (2004). Optimal predictive model selection. *The Annals of Statistics*, **32**, 870–897.

Heinze, G., Ambler, G., and Benner, A. (2022). MFP: Multivariable Fractional Polynomials. *https://CRAN.R-project.org/package=mfp*, 1-10.

Hubin, A., Storvik, G., and Frommlet, F. (2021). Flexible Bayesian nonlinear model configuration. *Journal of Artificial Intelligence Research*, **72**, 901–942.

Hubin, A., Storvik, G., and Frommlet, F. (2020). A novel algorithmic approach to Bayesian logic regression (with discussion). *Bayesian Analysis*, **15**, 263–333.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A*, **186**, 453–461.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **43**, 429–453.

Royston, P. and Sauerbrei, W. (2008). *Multivariable Model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley: Chichester.

Sabanés Bové, D. and Held, L. (2011). Bayesian fractional polynomials. *Statistics and Computing*, **21**, 309–324.

Sabanés Bové, D., Held, L., Gravestock, I., Davies, R., Moshier, S., Ambler, G., and Benner, A. (2022). BFP: Bayesian Fractional Polynomials. *https://CRAN.R-project.org/package=bfp*, 1-30.

Joint Species Spatial Modelling of Deer Count Data: A Simulation Study

Aoife K Hurley[1], James Sweeney[1]

[1] University of Limerick, Ireland

E-mail for correspondence: Aoife.Hurley@ul.ie

Abstract: When estimating the population sizes of numerous species, modelling them jointly can be advantageous. It allows for the inclusion of correlation between the species and the borrowing of information. We base this simulation study off data for deer count data in the Republic of Ireland which has spatial data on two different levels: counts at a 10 km^2 grid and recorded deaths at a county level. We propose using a Conditional Autoregressive approach with correlations and spatial precision parameters to differ between the species.

Keywords: Joint Species Modelling; Spatial Statistics; Spatial Modelling

1 Introduction

In the Republic of Ireland, there are three main species of deer: Fallow (*Dama dama*), Red (*Cervus elaphus*), and Sika (*Cervus nippon*). Both Fallow and Sika deer have been introduced to the Republic of Ireland, with an on-going debate as to whether Red deer are native or introduced.
Gaining an accurate estimation of the population of each species would undoubtedly aid in efforts when investigating diseases, disease transfer in livestock from wildlife and forest health. Previous studies have investigated the distributions of the different species across the island of Ireland, but have not investigate the impacts of the differing land types or correlation that exists between the species. For this problem we have data on the three species at two distinct spatial levels. Firstly, we have the number of each species sighted at a $10km^2$ grid across the Republic of Ireland. For each species, we also have the number of recorded deaths at a county level.
Before we implement our models on the data, we perform a simulation study. In this study we investigate the use of Conditional Autoregressive

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

models, and show that it can be used for such a problem and the associated benefits.

2 Simulation Study

For this study, we investigate the use of Conditional Autoregressive models for correlated multi-species data. We run our simulations for one, two and three species scenarios, but show only the modelling for the three species case. The first level of data is counts at a gridded level, similar to the 10km^2 grid for the sightings of deer. The second is at an accumulated or county level, similar to the county structure for the culled data.
For the three species scenario:

$$y_{ji} \sim \text{Pois}(\lambda_{ji})$$

$$\log(\lambda_{ji}) = X_i^{\text{T}}\beta_j + S_{ji}$$

$$S \sim MVN(\mathbf{0}, [\phi \otimes (D - \alpha A)]^{-1})$$

$$\phi = C \begin{pmatrix} \tau_1(1-\rho_{23}^2) & (\rho_{13}\rho_{23}-\rho_{12})\sqrt{\tau_1\tau_2} & (\rho_{12}\rho_{23}-\rho_{13})\sqrt{\tau_1\tau_3} \\ (\rho_{13}\rho_{23}-\rho_{12})\sqrt{\tau_1\tau_2} & \tau_2(1-\rho_{13}^2) & (\rho_{13}\rho_{12}-\rho_{23}\sqrt{\tau_2\tau_3} \\ (\rho_{12}\rho_{23}-\rho_{13})\sqrt{\tau_1\tau_3} & (\rho_{12}\rho_{13}-\rho_{23})\sqrt{\tau_2\tau_3} & \tau_3(1-\rho_{12}^2) \end{pmatrix}$$

$$C = \frac{1}{1-\rho_{12}^2-\rho_{13}^2-\rho_{23}^2+2\rho_{12}\rho_{13}\rho_{23}}$$

$$\beta_{ji} \sim \mathcal{N}(0, 10)$$

$$\alpha \sim \text{Unif}(0, 1)$$

$$\tau_j \sim \Gamma(2, 0.5)$$

$$\rho \sim \text{Unif}(-1, 1)$$

where j indicates species, X_i^{T} are the associated percentages of land coverings for square i, α is the spatial dependence parameter, ρ_{12} is the correlation between species 1 and 2, ρ_{13} is the correlation between species 1 and 3 and ρ_{23} is the correlation between species 2 and 3. We also allow τ, the spatially varying precision parameter, to vary per species.
For the recorded deaths at county k:

$$z_{jk} \sim \text{Pois}(\gamma_{jk})$$

$$\gamma_{jk} = \kappa_j \times \sum_{i \text{ in } k} \lambda_i$$

$$\kappa_j \sim \text{Unif}(0, 1)$$

where κ is a death rate, which we can differ across species but remains constant across the counties.

3 Results

We investigate the models in two phases, the first using the grid level data only and then including the accumulated county level data.
We simulated a 100×4 land cover matrix, X, on a 10×10 grid and group the squares into 5 counties. The models were run using the *rstan* package, for 5,000 iterations with a warm-up iteration period of 2,500 across 4 chains. This left 10,000 post warm-up draws for estimation.

3.1 Grid Level Data Only

For the model only including grid level data, we investigate the convergence of the parameters τ, ρ, α and β. Figure 1 illustrates the 95% credible intervals and true values for the underlying parameters of the model.

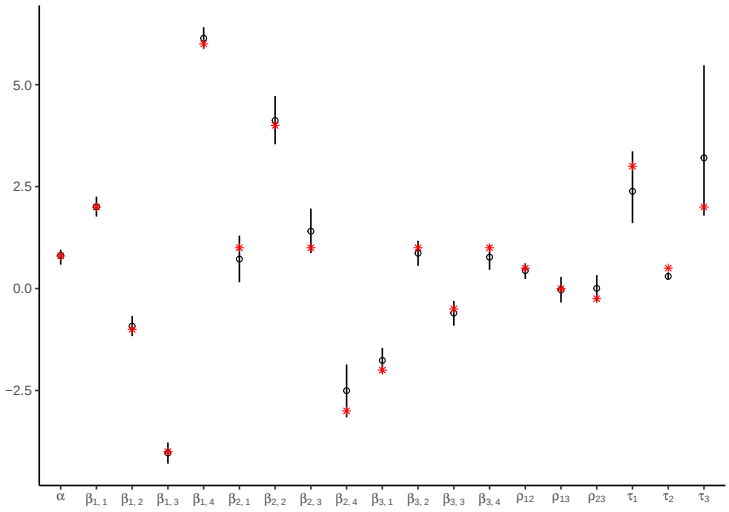


FIGURE 1. 95% Credible Intervals for parameters α, β, ρ and τ. The true values of these parameters is shown by the red star.

In Figure 1 $\beta_{j,n}$ represents the n^{th} beta coefficient for species j. It is clear some intervals are wider than others, particularly for τ_3, the spatially varying precision parameter for species 3. For these intervals to be narrower, the model may possibly need to be run for more iterations.

3.2 County and Grid Level Data

When we include the county level recorded deaths, we add the additional cull percentage parameter κ. In this study, we keep κ constant within

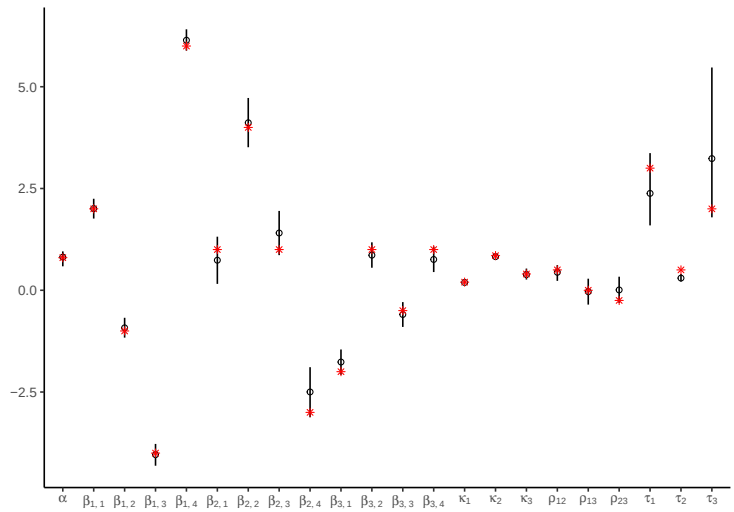


FIGURE 2. 95% Credible Intervals for parameters α, β, κ, ρ and τ. The true values of these parameters is shown by the red star.

species. Similar to Figure 1, Figure 2 shows the 95% credible intervals for the parameters associated with the full model.
Similar to the model that uses only grid level data, we notice the wider interval for τ_3 than any other parameter.

4 Conclusions and Future Work

By jointly modelling the species, we can account for between species correlation, borrow information as well as outputting interpretable coefficients for land cover per species. Computationally, these models take approximately 70 minutes to run on the 10×10 grid. Applying this to the real data will take significantly longer due to the increase in both the number of squares (approximately 8 times larger) and counties (actual data has 26 counties).
Developing these models further, we plan to investigate death rates, κ, that also vary by county. This may reflect reality more appropriately than a constant κ per species as legislation on hunting in the Republic of Ireland can change from county to county. We would also like to investigate the impact of allowing α, the spatial dependence parameter, vary per species. However, this would increase the computational complexity of the models. One of the final items we would like to examine is to include a detection probability. The addition of a detection probability might correspond well to reality, but may not be supported by the actual data. The final step is

to then apply our models used above to the deer data for the Republic of Ireland.

Acknowledgments: This work has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI 18/CRT/6049.

References

Carden, R.F., Carlin, C.M., Marnell, F., McElholm, D., Hetherington, J., and Gammell, M.P. (2011). Distribution and range expansion of deer in Ireland. *Mammal Review*, **4**, 313-325.

Gelfand, A.E., and Vounatsou P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis *Biostatistics*, **4(1)**, 11-15.

Jack, E., Lee, D., and Dean, N. (2019). Estimating the changing nature of Scotland's health inequalities by using a multivariate spatiotemporal model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182(3)**, 1061-1080.

Jin, X., Carlin, B. P., and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, **61(4)**, 950-961.

Comparison of models of the unemployment duration in the Czech Republic

Ivana Malá[1], Adam Čabla[1]

[1] Prague University of Economics and Business, Czech Republic

E-mail for correspondence: malai@vse.cz

Abstract: The duration of the unemployment spell provides information on the situation in the labour market as well as the more frequently used unemployment rate. There are two available data sources on the duration of unemployment in the Czech Republic. In the text, we model the distribution of the duration of the unemployment spell based on data from the Labour Force Sample Survey provided by the Czech Statistical Office (aggregated and individual data) and aggregated data from the database of registered unemployed people published by the Ministry of Labour and Social Affairs. Two parametric lognormal distribution is used as a model for the period from 2000 to 2019; methods of survival analysis and maximum likelihood estimates of parameters are used for individual data; the minimum chi-squared method is applied for aggregated data. We compare the time series of estimated parameters and median duration, discuss differences concerning different definitions of the being unemployed state, the exact meaning of the analysed variable and the impact of incomplete data on results.

Keywords: unemployment; aggregated data; incomplete data.

1 Introduction

The unemployment rate, the frequently used indicator of the labour market situation, and the median length of unemployment are basic measures but quite different in their reaction to the labour market. Unemployment time decreases when many employees lose their jobs, and, although people are not finding work, the unemployed with short periods of unemployment prevails (the unemployment rate is rising). In the period of a recession or crisis, people cannot find a job, and the duration of the unemployment spell will increase.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Available datasets and methods

The source of data on available job applicants, who are registered by the labour offices, is the Ministry of Labour and Social Affairs (referred to as MoLSA data). All registered unemployed people are included with the exact date of the end of previous job (exact unemployment spell duration is known), but the database is limited to registered people.
The Labour Force Sample Survey (LFSS; Eurostat 2022) is harmonised through the European Union and is performed by the Czech Statistical Office as a quarterly rotating panel (sampled households are included in the survey for one year). It covers households from approx. 23 thousand dwellings on the territory of the whole Czech Republic (about 0.6% of all permanently occupied dwellings), with the sample size higher than 42 thousand respondents aged over 15.
We apply two parameter lognormal distribution to model the distribution of the unemployment spell duration (or time to reemployment). The state "being unemployed" definition differs for MoLSA and LFSS datasets; moreover, for aggregated data, we model unemployment duration while in the case of individual data we can model actually time to reemployment using methods of survival analysis for time-to-event data. The datasets describe the same phenomenon "unemployment" but uses different definitions.
For our modelling, we have three datasets (duration in months):

- aggregated data from labour offices; referred to as MoLSA

  frequencies of the unemployed in intervals 0–3, 3–6, 6-9, 9–12, 12–24, 24+.

  The analysed variable is defined as the duration of unemployment spell. The parameters are estimated using minimum chi squared method. We obtain two estimated parameters of variance of T if we use a gender covariate.

- aggregated data from LFSS; referred to as LFSSa

  frequencies in intervals 0–1, 1–3, 3–6, 6–9, 9–12, 12–24, 24–18, 48+.

  The analysed variable is defined as the duration of unemployment spell. The parameters are estimated using minimum chi squared method.

- individual data from LFSS; referred to as LFSSb

  We include all respondents suffering at least once during the surveyed period from unemployment. We obtain right censored value if an individual stays unemployed, interval censored datum if an individual finds a job. No exact values are included. The survey is conducted quarterly, we use panel type of data and also information on the week of interviews to update intervals. Number of unemployed varies

in particular quarters from 600 to 2,800 with the percentage of finding a job from 8 % to 34 %.

The model (T is time to reemployment in month)

$$lnT = \mu + \sigma\epsilon$$

is applied for all data and the model

$$lnT = \mu + \alpha\, x_{[gender=Female]} + \sigma\epsilon$$

is used to differentiate between gender. Random error ϵ is distributed as standard normal resulting to T distributed as lognormal. Parameters μ and σ are parameters of lognormal distribution, in case of using the gender covariate of the baseline distribution for men. We obtain only one estimate of variance parameter. To estimate the survival model, R package *survival* was applied (Therneau, 2022).
Unlike some analyses (for example, Čabla and Malá, 2017), we considered all unemployed people, not only those with a period of unemployment shorter than two years.
The graphical presentation of smoothed quarterly time series (78 time points, Q1 2000 - Q2 2019) of estimated medians and both estimated parameters (expected and standard deviation of the logarithm of the analysed duration) is used to compare results from all datasets.

3 Results

In our model, we obtain highly skewed estimated distributions because of the relatively large ratio of unemployment spells longer than two years for aggregated data or unemployment spells longer than 4 or 10 years in individual data. It results in larger characteristics of a location in our study and even longer times based on the individual data. This data describes the tails using respondents with a very long spell. We do not estimate the proportion of those, who will never reemploy.
We refer to Figure 1 for smoothed time series of estimated parameters.The development of estimated parameters is similar for all datasets, showing higher values and more sensible time dependence for individual data.
From the parametric point of view, the decrease during an economic boom is driven mainly by a decrease of location parameter μ and not by the parameter σ, which means that skewness and relative variability of the distribution remains relatively stable.
The gender gap is visible in terms of higher parameter (and so the median) for women than for men and higher σ (and so skewness and relative variability) for men in MoLSA dataset with no clear persistent difference in the LFSS dataset. Gender is not supposed to be highly significant in comparing

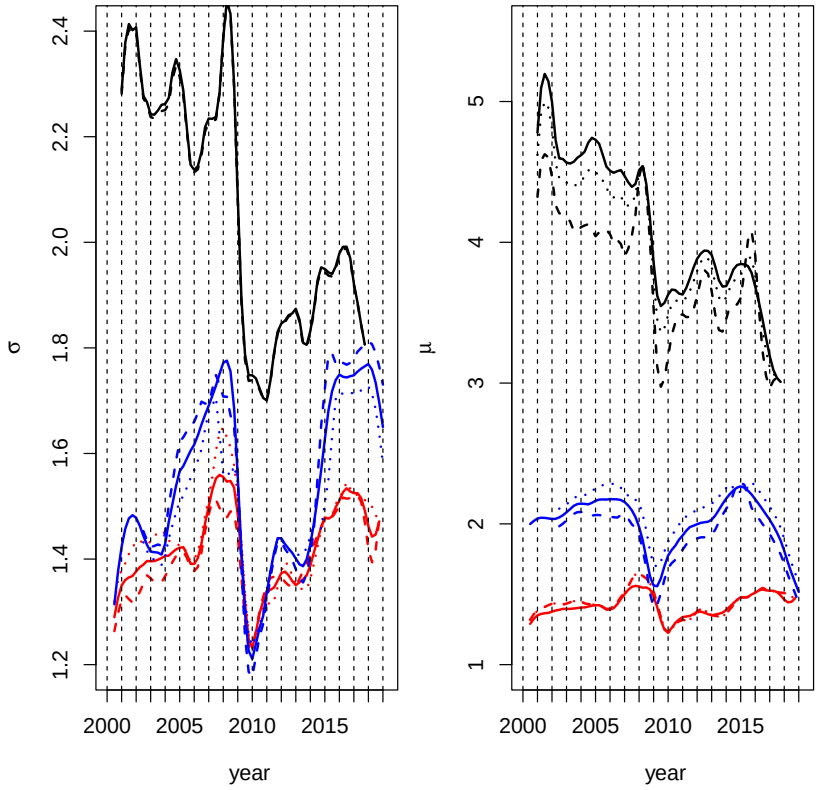


FIGURE 1. Estimated parameters from analysed data sources. Censored LFSSb data in black, LFSSa data in red, MoLSA data in blue. Female dotted, male respondents dashed curve.

labour market position but we show differences in estimated parameters. If gender is included as a covariate into the model, no visible difference in variance of baseline distribution is found.

We choose three distributions: pre-crisis Q2 2008, peak crisis Q1 2010 and economic boom Q4 2018. The estimated densities are shown in Figure 2 . For MoLSA data there is a clear shift rightwards during the crisis and then back leftwards with a much more pronounced mode for the last economic boom. This again suggests that much more unemployment spell is in the lower part of the distribution than in pre-crisis levels.For LFSSb data we can observe somewhat similar development, alas with increasing value of the mode.

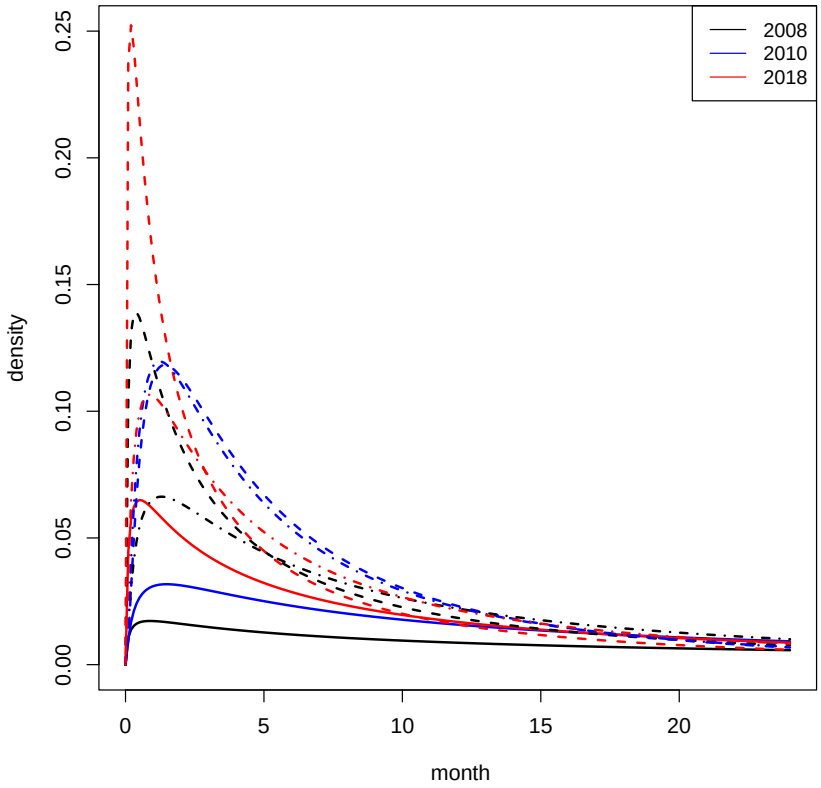


FIGURE 2. Estimated densities. LFSSb solid line, LFSSa dashed dotted line, MoLSA dashed line.

4 Conclusion

All datasets (in Malá and Čabla there is also included MoLSA data on applicants eligible for unemployment benefits) can be used to analyse the same phenomenon but are of different types from the point of view of origin, individual or aggregated form as well as there are two definitions of the unemployed. The unemployment rate is usually applied to describe the phenomenon and we try to describe different views to the same situation on the labour market and the shift between time series of the unemployment rate and median unemployment spell length.

We show that the mean and median unemployment duration decreases at the beginning of the economic crisis, which is caused by the inflow of newly unemployed, then slowly rises even at the beginning of economic recovery and boom, as the layoffs are limited. Probably the persons with lower unemployment spell are preferred for reemployment. If the economic

boom is long and stable enough, the mean and median unemployment duration starts to decrease, which can be viewed as a sign of reemployment of persons with long unemployment spells. From the aggregated data, we can model only the time of unemployment (as observed times of reemployment are not observed); in case of incomplete individual values, the time to reemployment is analysed and modelled. We show similar development of duration modelled based on different data and the impact of data to the estimated parameters, usually, it is not reasonable to compare absolute values of estimated characteristics or parameters.

Acknowledgments: This paper is supported by the long term institutional support of research activities by the Faculty of Informatics and Statistics, Prague University of Economics and Business. Data access in the SafeCentre of the Czech Statistical Office is gratefully acknowledged.

References

Čabla A. and Malá I. (2017) Modelling of Unemployment Duration in the Czech Republic. *Prague economic papers*, **26**, 438 – 449.

Malá I. and Čabla A. Modelling of the unemployment duration in the Czech Republic based on aggregated complete and individual censored data. *Ekonomický časopis.* (accepted).

Therneau, T. M. (2022) A Package for Survival Analysis in R, R package version 3.3-1, url = https://CRAN.R-project.org/package=survival.

LFSS survey, Eurostat (2022) Labour Force Sample Survey, url = https://ec.europa.eu/eurostat/cros/content/labour-force-sample-survey-lfss-en.

Modelling the population dynamics of the Blackspot Seabream (Pagellus Bogaraveo) on the Portuguese coast

Rui Martins[1,2], Lisete Sousa[1,2], Iúri Correia[2], Inês Farias[3], Ivone Figueiredo[3]

[1] Faculdade de Ciências da Universidade de Lisboa (FCUL), Portugal
[2] Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), Portugal
[3] Instituto Português do Mar e da Atmosfera (IPMA), Portugal

E-mail for correspondence: rmmartins@fc.ul.pt

Abstract: We present in this work the model for a problem brought to us by the Portuguese Institute for Sea and Atmosphere (IPMA): to estimate the stock structure and the main biological parameters of the Blackspot Seabream (Pagellus Bogaraveo) on the Portuguese coast (ICES – International Council for the Exploration of the Sea – Subarea 9 – Atlantic Iberian waters) for management purposes.
Population dynamics modelling is accomplished through a Bayesian state-space model based on length-classes. Thus no age estimates are required. This has the advantage of structuring the population in terms of a quantity that is directly observable, requiring no indirect estimation of age distributions according to length which can be notoriously difficult. The main biological subprocesses considered are growth, survival, mortality and recruitment. In the case of blackspot seabream there is another important process to account for – sex change.

Keywords: bayesian; fisheries; nimble; state-space; stock-assessment.

1 Pagellus Bogaraveo

Blackspot seabream (*Pagellus Bogaraveo*) is a fish with an high commercial interest. Distributes from the south of Norway to Cape Blanc, in the Mediterranean Sea, and in the Azores, Madeira, and Canary Archipelagos occurring from the continental shelf to 700m deep and on seamounts.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

Breeding occurs in shallow waters where the juveniles tend to remain. Subsequently there is an ontogenetic migration towards deeper waters. In Atlantic waters the main spawning period occurs during the first quarter with a peak of spawning from March to April. It is a protandric hermaphrodite species; most individuals are first functional males and then develop into functional females. In the Northeast Atlantic its stock structure is unclear and the level of mixing in the population from Gulf of Cadiz with those at the occidental Iberian coast is unknown. Further genetic studies showed a restricted gene flow among the populations located in the Azores and those on the Portuguese continental slope and Madeira (Farias *et al.* 2019). Given this uncertainty, and for management purposes, Portugal mainland area is managed jointly with Spain subareas – ICES Division 9.
If proven that the Portuguese mainland stock of *Pagellus Bogaraveo* is isolated from the remaining ones that might have some consequences in terms of its management by Portuguese authorities.
The data available consists of total reported catches in numbers for 6 years (2014 – 2019). The numbers were aggregated into length-classes.

2 State–Space Model

Let us consider two time-series running in parallel: (i) the state vector, that is unobserved, with value N_t at time t and (ii) the observation process with value N_t^O at time t, that is observed and is a function of the state process, $t = 1, 2, \ldots, T$. Both state and observation process might be vectors.
Here the state process describes the true, but unobservable, population demographics as it changes over time and its components are the numbers of fish abundance per each of the m length-class (Mäntyniemi *et al.* 2015). So, we denote the state of the population at t by $N_t = (N_{t,1}, \ldots, N_{t,m})$. The observation process has components that correspond to the numbers of caught fish per length-class, i.e. $N_t^O = (N_{t,1}^O, \ldots, N_{t,m}^O)$. The probabilistic structure for the state-space model (SSM) can thus be written as a set of three probability density/mass functions:

$$g_0(N_0 \mid \eta); \quad \text{Initial state}, \tag{1}$$

$$g_t(N_t \mid N_{t-1}, \eta); \quad \text{State process}, \tag{2}$$

$$f_t(N_t^O \mid N_t, \psi); \quad \text{Observation process}, \tag{3}$$

where $\theta \equiv (\eta, \psi)$ is the parameters vector. In our Bayesian context, the inference objectives include generating a sample from the posterior distribution for the states and unknown parameters conditional on the entire observation time series, $N_{1:T}^O = \left(N_1^O, N_2^O, \ldots, N_T^O\right)$. We consider a similar notation for the state vectors, $N_{1:T} = (N_1, N_2, \ldots, N_T)$. Denoting by $\pi(\theta)$ the prior distribution of the parameters, the joint posterior distribution can

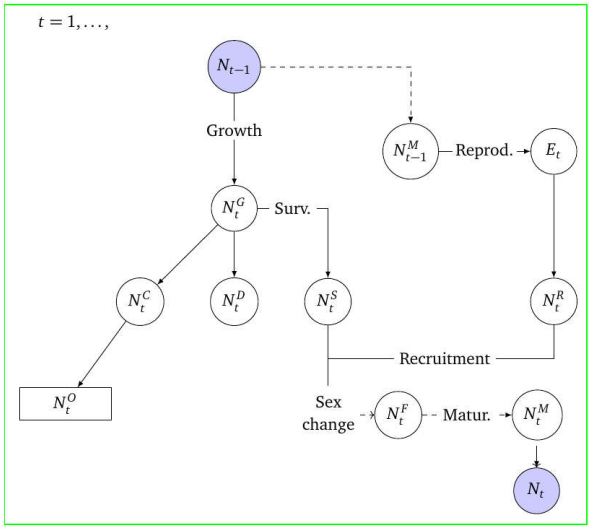


FIGURE 1. Representation of the biological subprocesses sequence that cause the evolution of the population from year $t-1$ to year t.

be written as

$$\pi(N_{0:T}, \theta \mid N^O_{1:T}) \propto \pi(\theta) g_0(N_0 \mid \eta) \times \prod_{t=1}^{T} g_t(N_t \mid N_{t-1}, \eta) f_t(N^O_t \mid N_t, \psi).$$

The evolution from N_{t-1} to N_t will be the result of a sequence of stochastic and deterministic subprocesses which we assume they are acting on the population sequentially but instantaneous.

2.1 Formulation

The yearly state process consists of six main subprocesses: growth, survival, mortality, reproduction and recruitment adn also sex change.
Figure 1 depicts an overview of the population evolution over period t, where N^G_t denotes the state of the population after growth; N^S_t, N^C_t and N^D_t denote the respective subpopulations relating to survival, fishing mortality (caught) and natural mortality. E_t denotes the number of eggs produced at the beginning of period t by the mature females in the end of year before, N^M_{t-1}, and finally, N^R_t denotes the population of recruits that enter the population at the end of period t. The state of the population N_t at the end of the year t is then obtained as $N_t = N^S_t + N^R_t$.
We assume that the numbers of catched fish per length-class, $N^O_{t,i}$, are centered around the value of $N^C_{t,i}$, i.e. the true value $N^C_{t,i}$ is observed with

errors. So we cast this assumption as the following Log-Normal model:

$$N_{t,i}^{O} \sim \text{Log-Normal}\left(\log\left(N_{t,i}^{C}+1\right)-\frac{\sigma_O^2}{2},\sigma_O^2\right).$$

3 Preliminary results

This is a first attempt to model the population dynamics of the Blackspot Seabream. New developments can still be pursued. Several assumptions were made because the model was fitted with very limited observed data. Adding the knowledge of the biologists, in the form of highly informed prior distributions, improved the results. The model was implemented in the R package `nimble` (de Valpine *et al.* (2017)).
The state process defined above is highly parametrized, reflecting the complexity of the system and the questions of interest. Thus, the influence of prior distributions for some parameters must be assessed.
A major goal of the modelling was to estimate the stock size using catches counts alone. Even with a short time-series and a relatively complex model, we have gained useful knowledge on the biology of the fish.

Acknowledgments: This work has been funded by Fundação para a Ciência e a Tecnologia (FCT) (UID/00006/2020). We thank Instituto Português do Mar e da Atmosfera (IPMA) for providing the observational data.

References

Mäntyniemi, S.P., Whitlock, R.E., Perälä, T.A., Blomstedt, P., et al. (2015) General state-space population dynamics model for Bayesian stock assessment. *ICES Journal of Marine Science*, **72**(8), 2209–2222.

Farias, I. and Figueiredo, I. (2019) Pagellus bogaraveo in Portuguese continental waters (ICES Division 27.9.a) *Working Document 14 to the 2019 ICES Working Group on the Biology and Assessment of Deep-Sea Fisheries Resources (WGDEEP)*

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C. et al. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, **6**(2), 403–413.

Binomial Confidence Intervals for Rare Events: Importance of Defining Margin of Error Relative to Magnitude of Proportion

Owen McGrath[1], Kevin Burke[1]

[1] University of Limerick, Ireland

E-mail for correspondence: Owen.McGrath@ul.ie

Abstract: Confidence intervals are usually assessed in terms of coverage probability and interval width (or margin of error). Using these criteria, we examine the performance of binomial proportion estimators when the success probability, p, is small. We discuss the importance of defining the margin of error relative to the proportion to avoid intervals that are impractically wide, or unnecessarily narrow (requiring very large sample sizes). To obtain valid interval estimates in a small-p regime, we propose a relative margin of error scheme that ensures that the margin of error is compatible with the order of magnitude of p.

Keywords: Confidence interval; Proportion; Rare event; Margin of error.

1 Introduction

The problem of estimating a small binomial proportion, p, is frequently encountered in applied statistics. For example, in manufacturing, the occurrence of a defect in a manufactured component is often considered as a rare event. In this work we consider $p \leq 10^{-1}$ as a rare-event probability, and assess the performance of four common proportion estimators: Wald, Agresti-Coull, Clopper-Pearson and Wilson (Agresti and Coull (1998); Clopper and Pearson (1934); Wilson (1927)), when the success probability is small.
Typically, the sample size, n, required for estimation is chosen based on setting the confidence interval margin of error to a fixed value, ϵ, and then solving for n. An inherent challenge with this technique is that ϵ must be defined in advance. This is not such an issue if p is moderately large, but when p is small, the definition of ϵ is crucial to the validity of the resulting interval. For example, $\epsilon = 0.05$ might be considered as a reasonable margin of error for $p = 10^{-1}$, but is too large for a proportion of the order $p = 10^{-3}$.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

In order to avoid such mismatches between ϵ and p, we propose a margin of error scheme that is considered relative to the order of magnitude of p. When interval performance is assessed in terms of both coverage probability and the proposed relative margin of error, we have found that, for a 95% confidence level, the four estimators perform similarly in many cases.

2 Sample Size and Performance Criteria

The most common sample size calculation for estimating a proportion with specified margin of error is based on the Wald confidence interval, and is given by

$$n = \left\lceil \frac{z^2_{\alpha/2} p^*(1-p^*)}{\epsilon^2} \right\rceil,$$

where $\lceil \cdot \rceil$ denotes the ceiling function, $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution, and p^* denotes an initial estimate of p. Such an initial estimate is standard practice in sample size planning (unless the 'default' $p^* = 0.5$ is used). For a given sample size n, and success probability p, the expected coverage probability is

$$CPr(n,p) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} 1(L_x \leq p \leq U_x),$$

where L_x and U_x are the lower and upper interval bounds calculated using x successes, and $1(\cdot)$ is an indicator function taking the value 1 when its argument is true, and 0 otherwise. The expected width, EW, is given by

$$EW(n,p) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} (U_x - L_x).$$

The expected margin of error, $EMoE$, is given as $EMoE = EW(n,p)/2$.

3 Relative Margin of Error

The importance of setting a margin of error that is consistent with the magnitude of p is illustrated in Table 1. Shown are the calculated Wald sample sizes (rounded to 2 significant figures) and coverage probabilities for $p^* = p = 10^{-3}$, corresponding to five margin of error schemes. When ϵ is large relative to p (schemes 1, 2 and 3), the coverage is unsatisfactory, whereas, when ϵ is small relative to p (scheme 5), the desired coverage is achieved, but a large sample size is required. Setting an ϵ that is consistent with the order of magnitude of p (scheme 4) produces reasonable coverage, and in comparison to scheme 5, requires a smaller sample size.
To maintain compatibility between ϵ and p^*, we consider a relative margin of error, $\tilde{\epsilon}_R = \epsilon/p^*$. We impose $\tilde{\epsilon}_R \leq 1$, however, $\tilde{\epsilon}_R$ values close to the

TABLE 1. Schemes with ϵ set independently of p^*

	Margin of Error Scheme				
	1	2	3	4	5
	$\epsilon = 4 \cdot 10^{-1}$	$\epsilon = 4 \cdot 10^{-2}$	$\epsilon = 4 \cdot 10^{-3}$	$\epsilon = 4 \cdot 10^{-4}$	$\epsilon = 4 \cdot 10^{-5}$
n	$1.0 \cdot 10^0$	$3.0 \cdot 10^0$	$2.4 \cdot 10^2$	$2.4 \cdot 10^4$	$2.4 \cdot 10^6$
CPr	0%	0.3%	21.3%	93.1%	95.0%

bound of 1 result in very wide intervals, and values close to the bound of 0 result in very narrow intervals (requiring excessively large sample sizes). We recommend $\tilde{\epsilon}_R \in [0.1, 0.5]$ and Table 2 shows that good coverage is achieved in this scheme.

TABLE 2. Schemes with ϵ set relative to p^*

	$\tilde{\epsilon}_R$						
	0.05	0.1	0.2	0.3	0.4	0.5	0.75
n	$1.5 \cdot 10^6$	$3.8 \cdot 10^5$	$9.6 \cdot 10^4$	$4.3 \cdot 10^4$	$2.4 \cdot 10^4$	$1.5 \cdot 10^4$	$6.8 \cdot 10^3$
CPr	95%	95%	95%	94.7%	93.1%	93.1%	90.4%

Figure 1 provides a confidence interval performance comparison for $p^* = p = 10^{-3}$ and $n \in [10,000, 80,000]$ in terms of CPr, and the expected relative margin of error, $\epsilon_R = EMoE/p^*$. Acceptable CPr and ϵ_R values are considered as $95 \pm 1\%$ and $\epsilon_R \leq 0.5$ respectively.

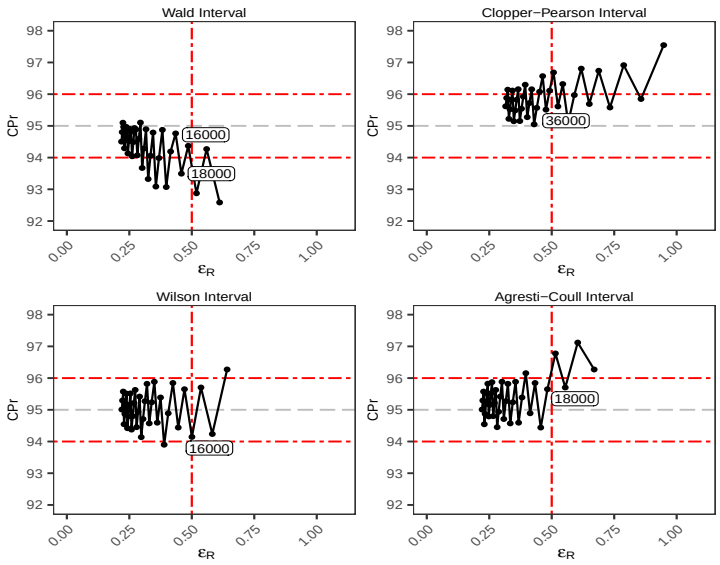


FIGURE 1. CPr versus ϵ_R for $p^* = p = 10^{-3}$. Dashed (grey) line represents the nominal CPr value. Dot-Dashed (red) lines represent the CPr and ϵ_R tolerances. Sample sizes decrease from left to right, in steps of 2000.

From Figure 1, we see that the coverage of all four estimators fluctuates as the sample size varies. For example, the coverage of the Wald interval is good at $n = 16,000$, but drops below 94% at $n = 18,000$ – this phenomenon of coverage oscillation has been previously discussed in the literature, e.g., Agresti and Coull (1998). In this particular scenario, the Wilson and Agresti-Coull intervals perform best, with the Wilson interval producing better coverage for $n < 16,000$. In this n range, none of the intervals satisfy the $\epsilon_R \leq 0.5$ requirement. However, when $\epsilon_R \leq 0.5$, the Wilson and Agresti-Coull intervals have similar performance.

4 Discussion

When constructing confidence intervals for a proportion, p, it is important that the margin or error, ϵ, be considered relative to the magnitude of p. When p is small, failure to consider ϵ relative to p can result in poor coverage, and/or intervals that are unnecessarily narrow or excessively wide. To ensure consistency between ϵ and p, we propose the use of a relative margin of error, $\epsilon_R = \epsilon/p$. We suggest restricting the range of values to $\epsilon_R \in [0.1, 0.5]$, as values outside this range can lead to imprecision and poor interval coverage, or impractically large sample sizes.
We have evaluated the proposed ϵ_R range in a variety of scenarios (values of p, n, and nominal coverage), and found that the four interval estimators perform similarly in many cases – albeit, here, we have only shown the case of a 95% confidence interval for $p = 10^{-3}$ due to space constraints. Similarly, we have also found the ϵ_R criterion useful for evaluating the adequacy of existing studies (also omitted here for brevity). However, ultimately, ideally the approach should feed into study planning in advance of data collection.

Acknowledgments: The authors would like to thank the Technological University of the Shannon for supporting this work, and the Confirm Smart Manufacturing Centre (https://confirm.ie/) funded by Science Foundation Ireland (grant number: 16/RC/3918).

References

Agresti, A., and Coull, B.A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, **52**(2), 119 – 126.

Clopper, C.J., and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404 – 413.

Wilson, E. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209 – 212.

Penalized Power-Generalized Weibull Distributional Regression

Laura McQuaid[1], Shirin Moghaddam[1] and Kevin Burke[1]

[1] University of Limerick, Ireland

E-mail for correspondence: laura.mcquaid@ul.ie

Abstract: Distributional regression for survival data refers to the approach whereby covariates enter the hazard function via multiple distributional parameters (e.g., scale and shape) simultaneously; this is also known as multi-parameter regression (MPR). We develop the MPR Power Generalized Weibull (PGW) model, which, with three parameters (one scale, two shapes), encompasses various common survival models and hazard shapes. Variable selection is challenging in this setting (and distributional regression more generally) since covariates can enter the model in various ways. Thus, we propose the use of a computationally feasible adaptive lasso penalized estimation procedure for variable selection and explore its performance using numerical studies.

Keywords: distributional regression; multi-parameter regression; parametric modelling; penalty; survival analysis

1 Introduction

The three parameter Power Generalized Weibull (PGW) model encompasses key shapes of hazard function (constant, increasing, decreasing, up-then-down, down-then-up) and a variety of survival distributions (Weibull, log-logistic, Gompertz) which makes it a highly flexible model, particularly in our proposed penalized MPR approach.

The Power Generalized Weibull (PGW) hazard is given by

$$\lambda(t) = \tau\gamma t^{\gamma-1}\left(1+\frac{t^{\gamma}}{\kappa+1}\right)^{\kappa-1} \tag{1}$$

where $\tau > 0$ and $\gamma > 0$ are scale and shape parameters, and $\kappa > -1$ is an additional shape parameter controlling the baseline distribution: $\kappa = 0 \Rightarrow$

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

log-logistic; $\kappa = 1 \Rightarrow$ Weibull; $\kappa \to \infty \Rightarrow$ Gompertz (Burke et al., 2020).

Taking a distributional regression approach, we then have that

$$\log(\tau) = x^T\beta, \qquad \log(\gamma) = x^T\alpha, \qquad \log(\kappa + 1) = x^T\omega$$

where $x = (1, x_1, \ldots, x_p)^T$ is a vector of covariates with regression coefficients, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)^T$, and $\omega = (\omega_0, \omega_1, \ldots, \omega_p)^T$, respectively; the log-link functions for τ and γ ensure their positivity, while the slightly different link function for κ ensures that $\kappa > -1$. Burke et al., (2020) found that the full distributional regression generality of allowing κ to depend on covariates is typically not required, and, therefore, we follow their suggestion that $x^T\omega = \omega_0$; however, the methods of this paper extend straightforwardly to the case where κ depends on covariates.

2 Penalized PGW Model

Variable selection is naturally more involved in this setting since there are various possibilities for a given covariate: it may appear in the *scale* (τ) but not the shape (γ), the *shape* but not the scale, *both* parameters, or *neither* parameter. Indeed, more classical variable selection procedures are hampered here by the fact that there are 2^{2p} sub-models; this, of course, is a general issue in distributional regression where there are $2^{d\times p}$ sub-models when there are d regression components. Moreover, the use of p-value-based selection is also more subtle since the net effect of a covariate may require that both scale and shape effects should be included even if one is non-significant when examined on its own (Burke and MacKenzie, 2017). Additionally, even in classical *single* parameter regression (i.e., $d = 1$), stepwise procedures are known to be unstable due to their inherent discreteness (i.e., covariates are either "in" or "out").

For all of the above reasons, we consider a penalized regression approach to carry out simultaneous parameter estimation and variable selection (via parameter shrinkage). More specifically, we make use of the adaptive lasso, which has been found to perform well in the Weibull model (Jaouimaa et al., 2019) that we extend via the PGW. Thus, the penalized likelihood for the vector of parameters $\theta = (\beta^T, \alpha^T, \omega^T)^T$ is

$$\ell_\lambda(\theta) = \sum_{i=1}^{n}\{\delta_i \log h(t_i) - H(t_i)\} - n\lambda\sum_{j=0}^{p}(w_{\beta_j}|\beta_j| + w_{\alpha_j}|\alpha_j|), \qquad (2)$$

where n is the sample size, $\delta_i \in \{0, 1\}$ is the censoring indicator, $h(t)$ is the hazard given in (1), $H(t) = (1 + 1/\kappa)[\{1 + t^\gamma/(\kappa + 1)\}^\kappa - 1]$ is the cumulative hazard, the dependence on covariates is implicitly assumed, λ

is the penalty tuning parameter, and $w_{\beta_j} = 1/\hat{\beta}_{0j}$ and $w_{\alpha_j} = 1/\hat{\alpha}_{0j}$ are the adaptive weights where $\hat{\beta}_{0j}$ and $\hat{\alpha}_{0j}$ maximize the unpenalized likelihood $\ell_0(\theta)$, i.e., $\ell_\lambda(\theta)$ with $\lambda = 0$. The presence of the absolute value function in (2) prevents gradient-based optimization. Thus, for practical purposes, we replace $|z|$ with the differentiable approximation $a_\epsilon(z) = \sqrt{z^2 + \epsilon^2} - \epsilon^2$, e.g., as in Lloyd-Jones et al., (2018), which is close to $|z|$ when ϵ is small.

The tuning parameter λ controls the level of parameter shrinkage such that larger values lead to greater shrinkage, and, hence, sparser models. In practice, one aims to choose the "optimal" λ value, which is akin to model selection. We use the BIC criterion

$$\text{BIC}(\lambda) = -2\ell_0(\hat{\theta}_\lambda) + e_\lambda \log n, \tag{3}$$

where we note that $\ell_0(\hat{\theta}_\lambda)$ is the unpenalized likelihood function evaluated at the penalized estimates $\hat{\theta}_\lambda$ (i.e., those which maximize (2)), and $e_\lambda = \text{tr}[\{I_\lambda(\hat{\theta}_\lambda)\}^{-1} I_0(\hat{\theta}_\lambda)]$ is the effective degrees of freedom where $I_\lambda(\theta)$ and $I_0(\theta)$ are the negative hessian matrices for $\ell_\lambda(\theta)$ and $\ell_0(\theta)$, respectively. We define λ^* to be the minimizer of (3), which we obtain using a simple grid search due to the one-dimensional nature of the problem, and, in turn, the estimated parameter vector $\hat{\theta}_{\lambda^*}$. This defines the selected model since the zero coefficients in $\hat{\theta}_{\lambda^*}$ correspond to the model components (scale and shape) from which particular covariates have been dropped. In fact, since we approximate $|z|$, no coefficient is set exactly to zero, but it can be made arbitrarily close to zero by decreasing ϵ.

3 Simulation Study

We explore the performance of this procedure on simulated PGW data with

$$\log(\tau_i) = x_i^T(-1.5,\ -1.0,\ 0.5,\ 0.0,\ 0.0,\ 0.0,\ 1.0,\ 0.0)^T,$$
$$\log(\gamma_i) = x_i^T(0.5,\ 0.75,\ 0.0,\ 0.0,\ 0.25,\ 0.0,\ -0.6,\ 0.0)^T,$$

where $x_i = (1, x_{i1}, \ldots, x_{i7})^T$ is a vector of covariates for the ith individual. We have considered a range of κ values, sample sizes, and censoring proportions, but here, we report only the results for the Weibull distribution ($\kappa = 1$) with 25% censoring.

In Table 1, we display the performance of our proposed procedure using a variety of metrics: C, the average number of true zero coefficients correctly set to zero; IC, the average number of true non-zero coefficients incorrectly set to zero; and PT, the probability of choosing the true model.

TABLE 1. Simulation results

n	θ	C	IC	PT
oracle values:		4	0	1
500	β	3.91	0.00	0.93
	α	3.90	0.00	0.91
1000	β	3.94	0.00	0.96
	α	3.94	0.00	0.95
2000	β	3.98	0.00	0.99
	α	3.97	0.00	0.98

4 Discussion

We can see from Table 1 that the proposed approach performs well with C tending towards the oracle value of 4 as the sample size increases (and IC is always equal to zero for the scenario considered). Similarly, the probability of selecting the true set of covariates tends towards one. We have also found that this favourable performance is maintained across a range of κ values and censoring proportions. Moreover, although not shown here, we have found that the inferential properties are also favourable, i.e., the bias is low and reduces with sample size, the standard errors reduce with sample size and are accurately estimated. This is particularly noteworthy given the largely automated nature of the estimation/selection procedure along with the overall flexibility of the PGW distributional regression model, which covers various common survival distributions and (covariate-dependent) hazard shapes. Thus, we anticipate that our proposal will be useful in practice, and, indeed, have found this to be the case in our own real data analysis (omitted here for brevity).

Acknowledgments: The first author would like to thank the Irish Research Council (www.research.ie) for supporting this work (GOIPG/2020/1307).

References

Burke, K and Jones, MC and Noufaily, A (2020).*A flexible parametric modelling framework for survival analysis*. Journal of the Royal Statistical Society: Series C (Applied Statistics).

Burke, K and MacKenzie, G (2017). *Multi-parameter regression survival modeling: An alternative to proportional hazards*. Biometrics.

Jaouimaa, F-Z and Ha, ID and Burke, K (2019). *Penalized Variable Selection in Multi-Parameter Regression Survival Modelling.* arXiv preprint arXiv:1907.01511.

Lloyd-Jones LR, Nguyen HD, McLachlan GJ. (2018). *A globally convergent algorithm for lasso-penalized mixture of linear regression models.*. Computational Statistics & Data Analysis.

On new classes of tests for the Pareto distribution based on the empirical characteristic functions

L. Ndwandwe[1], J.S. Allison[1], M. Smuts[1], I.J.H. Visagie[1]

[1] School of Mathematical and Statistical Sciences, North-West University, South Africa

E-mail for correspondence: lethani.ndwandwe@nwu.ac.za

Abstract: We propose new classes of tests for the Pareto type I distribution. These tests utilise the empirical characteristic function and is based on a lesser-known characterisation of the Pareto distribution. The finite sample performances of the newly proposed tests are evaluated and compared to some of the existing tests, where it is found that the new tests are competitive in terms of empirical powers.

Keywords: Characteristic function; Goodness-of-fit; Pareto type I distribution.

1 Introduction and motivation

The Pareto type I (henceforth referred to as the Pareto) distribution is a popular model in economics, finance and actuarial science, especially where phenomena characterised by heavy tails are studied, see, e.g. Ismaïl (2004). In finance, an application is to model stock price returns while in actuarial science it is frequently used to model excess of losses in insurance, see Rytgaard (1990). Due to its heavy tail, this distribution also plays a pivotal role in extreme value theory see, Beirlant et al. (2004). A number of characterisations for the Pareto distribution have been proposed in the literature, see, e.g. Gupta (1973). However, only a small number of goodness-of-fit tests have been developed in order to test the hypothesis that an observed dataset is compatible with the assumption of being realised from this distribution. Due to the increasing popularity of the Pareto distribution we propose new goodness-of-fit tests based on a characterisation and utilising the empirical characteristic function (ecf).

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Test statistic

Before proceeding, we introduce some notation. Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) copies from a continuous random variable, X, with unknown distribution function F. X is said to follow the Pareto distribution with parameter β, denoted by $X \sim P(\beta)$, if it has distribution function $F(x) = 1 - x^{-\beta}, \quad x \geq 1, \quad \beta > 0$. Throughout, the value of β is estimated by its method of moments estimator $\widehat{\beta_n} = \overline{X}_n/(\overline{X}_n - 1)$, where $\overline{X}_n$ is the sample mean. Let $X_{1:n} < X_{2:n} < \cdots < X_{n:n}$ denote the order statistics of $X_1, X_2, \ldots, X_n$. The composite null hypothesis to be tested is

$$H_0 : X \sim P(\beta), \tag{1}$$

for some $\beta > 0$, against general alternatives. We propose new classes of goodness-of-fit tests for testing (1) based on the following characterization given in Allison et al. (2021),

Theorem 1: , *Let $X, X_1, \ldots, X_n$ be i.i.d. random variables from a continuous distribution with distribution function F. Let m be an integer such that $2 \leq m \leq n$. $X^{1/m}$ and $\min(X_1, \ldots, X_m)$ have the same distribution if, and only if, $F(x) = 1 - x^{-\beta}, \quad x \geq 1, \quad \beta > 0$.*

Define the empirical characteristic function of $X^{1/m}$ by

$$\phi_n(t) = \frac{1}{n} \sum_{j=1}^{n} \exp(itX_{j:n}^{1/m})$$

and the ecf of the $\binom{n}{m}$ random variables $\min(X_{k_1}, \ldots, X_{k_m})$, $1 \leq k_1 < k_2 < \cdots < k_m \leq n$ as

$$\psi_n(t) = \binom{n}{m}^{-1} \sum_{1 \leq k_1 < k_2 < \cdots < k_m \leq n} \exp(it\min(X_{k_1}, X_{k_2}, \ldots, X_{k_m}).$$

After some algebra and combinatorics it follows that $\psi_n(t)$ can be expressed as a single summation

$$\psi_n(t) = \binom{n}{m}^{-1} \sum_{j=1}^{n-m+1} \binom{n-j}{m-1} \exp(itX_{j:n}).$$

From the above characterisation it follows that if $X_1, X_2, \ldots, X_n$ is a random sample from the Pareto distribution, then the difference between $\phi_n(t)$ and $\psi_n(t)$ should be close to zero. We thus suggest the following test statistic:

$$T_{n,m,a} = n \int_{-\infty}^{\infty} |\phi_n(t) - \psi_n(t)|^2 w_a(t) \mathrm{d}t,$$

where $w_a(t)$ is an appropriate weight function which depends on a user defined parameter a. After some algebra, we obtain easily calculable test

TABLE 1. Various choices of the alternative distributions.

Alternative	$f(x)$	Notation
Gamma	$\frac{1}{\Gamma(\theta)}(x-1)^{\theta-1}e^{-(x-1)}$	$\Gamma(\theta)$
Weibull	$\theta(x-1)^{\theta-1}\exp\left\{-(x-1)^{\theta}\right\}$	$\mathrm{W}(\theta)$
Lognormal	$\exp\left(-\frac{1}{2}\left(\log(x-1)/\theta\right)^2\right)\Big/\left\{\theta(x-1)\sqrt{2\pi}\right\}$	$\mathrm{LN}(\theta)$
Linear failure rate	$(1+\theta(x-1))\exp(-(x-1)-\theta(x-1)^2/2)$	$\mathrm{LF}(\theta)$

statistic based on the choices $w_a(t) = e^{-a|t|}$ and $\widetilde{w}_a(t) = e^{-at^2}$ denoted by $T^{(1)}_{n,m,a}$ and $T^{(2)}_{n,m,a}$, respectively. The calculable form of the test statistics can be expressed in terms of functionals of $X_1, X_2, \ldots, X_n$. The null hypothesis is rejected for large values of $T^{(1)}_{n,m,a}$ and $T^{(2)}_{n,m,a}$. The proposed tests are consistent against fixed alternatives. However, we omit the proof due to page limitations.

3 Simulation study

We compare the finite sample performance of the newly proposed tests to the traditional Kolmogorov-Smirnov (KS_n), Cramér-von Mises (CM_n) and Anderson-Darling (AD_n) tests as well as to a test based on the likelihood ratio proposed by Zhang (2002) denoted by ZA and an integral-type test based on a characterisation of the Pareto distribution proposed by Obradović et al. (2015), denoted by OJ.

3.1 Simulation setting and results

Power (and size) estimates are calculated at a significance level of 5% for sample size $n = 20$ using 50 000 independent Monte Carlo replications. Since the null distribution of all the test statistics depends on an unknown parameter, the parametric bootstrap will be used to calculate critical values for the different tests (the number of bootstrap replications used is $B = 1000$). We obtain power estimates for the various alternative distributions given in Table 1. These estimates are displayed in Table 2.
From Table 2 it is clear that each of the tests maintain the specified significance level of 5% closely. Considering the powers obtained against the various alternatives we see that the tests KS and OJ are less powerful than the other tests considered. It is also evident that CV is quite powerful for the gamma and linear failure rate alternatives, whilst the new tests obtained the highest powers against gamma, log-normal and weibull alternatives.

TABLE 2. Empirical powers for $n = 20$.

Dist	KS	CV	AD	ZA	OJ	$T_{3,1}^{(1)}$	$T_{3,2}^{(1)}$	$T_{4,1}^{(2)}$	$T_{4,2}^{(2)}$
Par(2)	5	5	5	5	5	5	5	5	5
Par(5)	5	5	5	5	5	4	4	4	4
$\Gamma(0.8)$	15	**16**	**16**	11	11	**17**	**16**	14	14
$\Gamma(1)$	39	**44**	42	34	34	40	**43**	41	40
W(0.8)	9	9	**12**	9	6	**14**	11	8	7
LN(1)	71	81	82	**90**	72	76	79	**87**	**87**
LN(2.5)	11	9	26	22	26	31	43	**49**	**50**
LN(3)	18	15	51	56	50	48	64	**73**	**75**
LFR(0.2)	46	**53**	50	41	45	49	**51**	48	48
LFR(0.5)	52	**60**	57	47	53	57	**58**	54	54
LFR(1)	56	**65**	61	53	61	62	**63**	60	60
LFR(1.5)	60	**68**	66	58	66	**67**	**67**	63	64

References

Alliso, J., Milošević, B., Obradović, M., and Smuts, M. (2021). Distribution-free goodness-of-fit tests for the pareto distribution based on a characterization. *Computational Statistics,* (pp. 1–16).

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons.

Giacomini, R., Politics, D. N., and White, H. (2013). A warp-speed method for conducting monte carlo experiments involving bootstrap estimators. *Econometric theory*, 567– 589

Gupta, R. C. (1973). A characteristic property of the exponential distribution. *Sankhyā: The Indian Journal of Statistics, Series B* , 365– 366.

Ismail, S. (2004). A simple estimator for the shape parameter of the Pareto distribution with economics and medical applications. *Journal of Applied Statistics,* **31** (1), 3– 13.

Obradović, M., Jovanović, M., and Milošević, B. (2015). Goodness-of-fit tests for pareto distribution based on a characterization and their asymptotics. *Statistics,* **49** (5), 1026– 1041.

Rytgaard, M. (1990). stimation in the pareto distribution. *ASTIN Bulletin: The Journal of the IAA,* **20** (2), 201–216.

Zhang, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **64** (2), 281–294.

Covariate-adjusted Association of Sensor Outputs

Lizzie Neumann[1], Jan Gertheiss[1]

[1] Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany

E-mail for correspondence: neumannl@hsu-hh.de

Abstract: Sensor data obtained from structural health monitoring of highway bridges is dependent on environmental influences such as temperature. In this paper, an approach for adjusting the corresponding covariances of sensor outputs by use of penalized regression splines is presented.

Keywords: Sensor Data; Partial Covariance; Penalized Regression Splines.

1 Introduction

Structural health monitoring uses sensor data from structure buildings such as bridges to monitor them. As these are not measured under laboratory conditions, these data are dependent on environmental influences such as temperature. Therefore, a model to adjust for these covariates is required before considering the association between the sensor outputs.

FIGURE 1. Valley Bridge Sachsengraben (Bundesanstalt für Straßenwesen, 2016)

The OSIMAB *(Online Safety Management System for Bridges)* (OSIMAB mCLOUD, 2020) data set consists of sensor measurements of the valley

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

bridge Sachsengraben on the motorway A45 in Germany. Among other things, strain was measured with six strain gauges in 50 hertz and temperature with six structure temperature sensors and one outer temperature sensor in 1 hertz on September 1st 2020. The strain data was downsampled to 1 hertz. In Figure 1 a picture of the bridge is shown. The strain senors are evenly distributed on the right and left sides of the road, with the odd named ones on the north (left) side. Temperature sensor 1 is in the middle of the street between strain sensor 1 and 2, and temperature sensor 10 is next to strain sensor 1.

2 Conditional and Partial Covariance

Let u and v be two random variables describing two different sensor outputs and let z denote a potentially confounding covariate, such as temperature. First, let us assume that u, v and z are jointly normal, i.e.

$$\begin{pmatrix} u \\ v \\ z \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_u \\ \mu_v \\ \mu_z \end{pmatrix}, \begin{pmatrix} \sigma_{uu} & \sigma_{uv} & \sigma_{uz} \\ \sigma_{vu} & \sigma_{vv} & \sigma_{vz} \\ \sigma_{zu} & \sigma_{zv} & \sigma_{zz} \end{pmatrix} \right). \tag{1}$$

Further let

$$\mu_{uv} = \begin{pmatrix} \mu_u \\ \mu_v \end{pmatrix}, \quad \Sigma_{uv} = \begin{pmatrix} \sigma_{uu} & \sigma_{uv} \\ \sigma_{vu} & \sigma_{vv} \end{pmatrix}, \quad \Psi = \begin{pmatrix} \sigma_{uz} \\ \sigma_{vz} \end{pmatrix}.$$

Then for the conditional distribution of (u, v) given z we have

$$\begin{pmatrix} u \\ v \end{pmatrix} | z \sim N\left(\mu_{uv} + \frac{1}{\sigma_{zz}} \Psi (z - \mu_z), \Sigma_{uv} - \frac{1}{\sigma_{zz}} \Psi \Psi^\top \right).$$

For estimating the *conditional* covariance of u and v given z

$$\sigma_{uv|z} = \sigma_{uv} - \frac{\sigma_{uz}\sigma_{vz}}{\sigma_{zz}},$$

we can use the empirical versions of σ_{uv}, σ_{uz}, σ_{vz} and σ_{zz}.

Alternatively, u and v can be regressed on z, and then we can calculate the covariance of the residuals. This approach is known under the name *partial* covariance. The covariance of u and v regressed on z is consistent with the conditional covariance if we assume the joint normal distribution assumption of (u, v, z):

$$\begin{aligned} & E\left(\left(u - \left(\mu_u + \frac{\sigma_{uz}}{\sigma_{zz}}(z - \mu_z) \right) \right) \left(v - \left(\mu_v + \frac{\sigma_{vz}}{\sigma_{zz}}(z - \mu_z) \right) \right) \right) \\ = \quad & \sigma_{uv} - \frac{\sigma_{uz}\sigma_{vz}}{\sigma_{zz}}. \end{aligned}$$

However, assuming joint normal distribution (1) for the derivation of the conditional covariance and the use of linear regression for partial covariance might be too restrictive. Therefore, as a generalization of the partial covariance approach, we allow nonlinear regression functions $f_u(z)$ and $f_v(z)$ when modeling the association between the sensor outputs u and v and the covariate z, respectively. Then we have for the sensor outputs u and v

$$\begin{aligned} u &= f_u(z) + \epsilon, \\ v &= f_v(z) + \xi, \end{aligned} \tag{2}$$

where the regression functions $f_u(z)$ and $f_v(z)$ are fitted by penalized splines as implemented in `mgcv` R-package (Wood, 2022). The *partial* covariance is then estimated as the covariance of the residuals ϵ and ξ.

3 OSIMAB

The data of strain sensor 1 and temperature sensors 1 and 10 are shown in Figure 2. Looking at it, the strain could be dependent of the temperature.

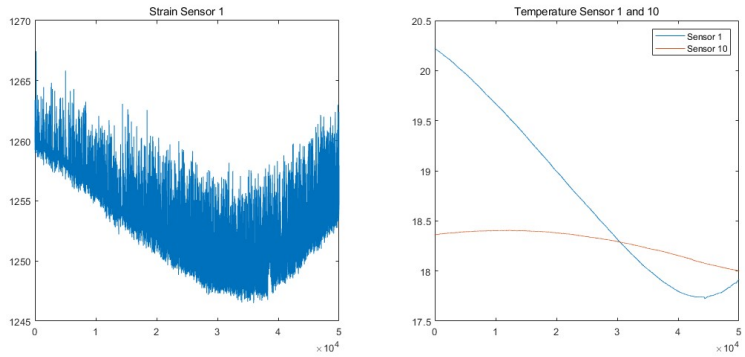


FIGURE 2. Strain (Sensor 1) and Temperature (Sensor 1 and 10) Data

Therefore, for estimating the partial covariance, the data is modeled as in Equation (2) where the response variables u and v are the outputs of the respective strain sensors, the covariate z is the output of the temperature sensors 1 or 10 and f_u and f_v are penalized cubic regression splines; k-folds cross validation is used to determine the tuning parameter and the `gam` function of the `mgcv` R-package is used to fit the model.
Then the empirical covariance of the residuals are calculated. In Figure 3 the covariance of the strain data, i.e. the marginal covariance, and the partial covariance, i.e. the covariance of the residuals is shown in dependency of the 1st and 10 temperature sensor, respectively. As can be seen in Figure 3, the covariance of the strain data is much higher than the covariance

of the residuals, with the covariance of the more fluctuating temperature (sensor 1) being lower than that of the more constant temperature (sensor 10). Furthermore, there seem to be some rather structural differences between the marginal covariance (top) and the partial version if regressed on sensor 1 (bottom left).

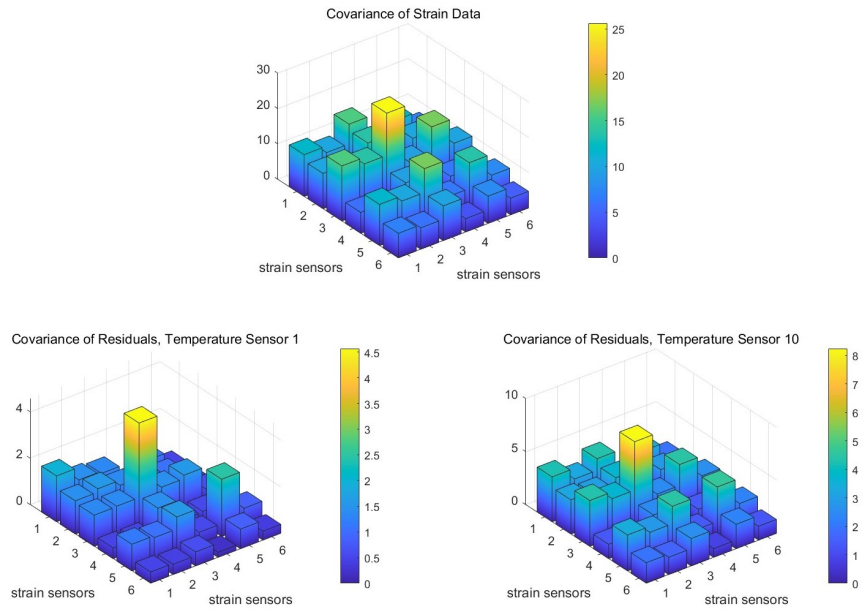


FIGURE 3. Covariance of the Strain Data and Residuals (Temperature sensor 1 and 10)

References

Bundesanstalt für Straßenwesen (2016) Intelligente Brücken. url = bast: https://www.bast.de/Forschungsplanung/DE/Verkehrsinfrastruktur/Beiträge/I-Bruecken.html?nn=2965460

OSIMAB mCLOUD (2020) text of license: http://www.govdata.de/dl-de/by-2-0, url = https://www.mcloud.de/downloads/mcloud/0489759B-1168-4 EF5-A26D-488DE44816FE/osimab_sample.zip

Wood, S. (2022) here: mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R package version 1.8-40. url = https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

Investigating different parameter estimation techniques for the Lomax distribution

Thobeka Nombebe[1], Leonard Santana[1], James S. Allison[1], Jaco Visagie[1]

[1] School of Mathematical and Statistical Sciences, North-West University, South Africa

E-mail for correspondence: Thobeka.Nombebe@nwu.ac.za

Abstract: We investigate the performance of a variety of estimation techniques for the scale and shape parameter for the Lomax distribution. These methods include the L-moment estimator, the probability weighted moments estimator, the maximum likelihood estimator, maximum likelihood estimator adjusted for bias, method of moments estimator and three different minimum distance estimators. The comparisons will be done by considering the variance and the bias of these estimators. Based on an extensive Monte Carlo study we found that the so-called minimum distance estimators are the best performers for small sample sizes, however for large sample sizes the maximum likelihood estimators outperform these minimum distance estimators.

Keywords: Lomax distribution; L-moments estimator; Maximum likelihood estimator; Method of moments estimator; Minimum distance estimators; Probability weighted moments estimator.

1 Introduction

The Pareto distribution is a heavy-tailed distribution that was originally developed in the 19th century to model the distribution of income among individuals Pareto (1897). However, in the years prior to its introduction, it has been extensively modified and changed to produce several variants, referred to as the Type I, II, III, and IV Pareto distributions. The focus of this paper is on the Pareto Type II distribution where the location parameter is zero, also known as the *Lomax* distribution.
We say a random variable X follows a Lomax distribution with scale parameter $\sigma > 0$ and shape parameter $\beta > 0$, if its cumulative distribution

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

function (CDF) is given by $F(x;\sigma,\beta) = 1 - \left[1 + \left(\frac{x}{\sigma}\right)\right]^{-\beta}$, $x > 0$.
This paper aims to discuss and study a number of different methods for obtaining the estimators of the scale (σ) and shape (β) parameters of the Lomax distribution. We start by considering the myriad different traditional methods of estimation proposed for related distributions, including the L-moment estimator, the probability weighted moments estimators (PWM), the maximum likelihood estimators (MLE), and method of moments estimators (MME). We then go on to propose the use of so-called 'minimum distance estimators' (MDEs) in the hope that these will be competitive alternatives to the more traditional options. The estimators are compared to one another via a comprehensive numerical study involving Monte Carlo simulations where the finite sample variance, bias, and mean squared error (MSE) of each estimator is approximated. In addition, an omnibus measure allowing one to gauge the MSE of the estimation of both σ and β simultaneously is also employed in the simulation study.

2 Estimation of parameters

For the remainder of the section, we assume that we have data $X_1, X_2, \ldots, X_n$ which is i.i.d. from the Lomax distribution with parameters $\beta > 0$ and $\sigma > 0$. The order statistics based on this sample are denoted using $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$. The estimators used are:

- **The traditional MME and MLE**
- **L–moment estimators (LME):**

$$\widehat{\beta}_{LM} = \frac{l_2}{2l_2 - l_1} \quad \text{and} \quad \widehat{\sigma}_{LM} = \frac{l_1^2 - l_1 l_2}{2l_2 - l_1}.$$

- **Probability Weighted Moment Estimators (PWM):** Proposed by Greenwood et al. (1990), the PWMs are given by

$$\widehat{\beta}_{PW} = \frac{2\widehat{M}_{1,0,1} - \widehat{M}_{1,0,0}}{4\widehat{M}_{1,0,1} - \widehat{M}_{1,0,0}} \text{ and } \widehat{\sigma}_{PW} = \frac{2\widehat{M}_{1,0,0}\widehat{M}_{1,0,1}}{\widehat{M}_{1,0,0} - 4\widehat{M}_{1,0,1}}.$$

  with $\widehat{M}_{1,0,0} = \frac{1}{n}\sum_{j=1}^{n} X_{j:n}$, and
  $\widehat{M}_{1,0,v} = \frac{1}{n}\sum_{j=1}^{n} \frac{(n-j)(n-j-1)\cdots(n-j-v+1)}{(n-1)(n-2)\cdots(n-v)} X_{j:n}$ $v = 1, 2, \ldots$.

- **MLE of the Lomax distribution adjusted for bias:** we consider a bias-adjusting approach for the MLEs used to reduce the bias of the MLE's to order $O(n^{-1})$ proposed by Giles et al., 2011. Denote these estimators by σ_{MLB} and β_{MLB}.

- **Cramér-von Mises (CVM) minimum distance estimators:** This distance measure permits a simple calculable form given by

$$D_n^{CM}(\sigma,\beta) = \frac{1}{12n} + \sum_{j=1}^{n}\left(1 - \left(\frac{\sigma}{X_{(j)}+\sigma}\right)^{\beta} - \frac{2j-1}{2n}\right)^2.$$

  The resulting estimators for σ and β are denoted by $(\widehat{\sigma}_{CM}, \widehat{\beta}_{CM}) = \arg\min_{(\sigma,\beta)} D_n^{CM}(\sigma,\beta)$.

- **'Squared difference' (SD) minimum distance estimators:** An alternative to the Cramér-von Mises distance measure with the following tractable calculation form:

$$D_n^{SD}(\sigma,\beta) = \sum_{j=1}^{n}\left(1 - \left(\frac{\sigma}{X_{(j)}+\sigma}\right)^{\beta} - \frac{j}{n+1}\right)^2.$$

  $(\widehat{\sigma}_{SD}, \widehat{\beta}_{SD}) = \arg\min_{(\sigma,\beta)} D_n^{SD}(\sigma,\beta)$.

- **The ϕ--divergence minimum distance estimators:** represent a broad class of distance measures that describe the distance between two densities f and g, and is defined as:

$$\delta(f,g) = \mathrm{E}\left[\phi\left(\frac{f(X)}{g(X)}\right)\frac{g(X)}{f(X)}\right], \tag{1}$$

  where X is a random variable from a distribution function with density $f(x)$ and $\phi(\cdot)$ is a convex function such that $\phi(1) = 0$ and $\frac{\partial^2\phi(x)}{\partial x^2}\Big|_{x=1} = \phi''(1) > 0$. In this setting, we will estimate the true density using the kernel density estimator, $\widehat{f}_h(x)$ with choices $\phi(t) = t\log(t)$ for Kullback-Liebler (Phi.kl) and we denote the estimator by $(\widehat{\sigma}_{KL}, \widehat{\beta}_{KL})$, $\phi(t) = (t-1)^2$ for the chi-square (Phi.X2) and we express the resulting estimator as $(\widehat{\sigma}_{CS}, \widehat{\beta}_{CS})$ and $\phi(t) = |t-1|$ for total variation (Phi.tv) ϕ-divergence distance measure, we designate the resulting estimator by $(\widehat{\sigma}_{TV}, \widehat{\beta}_{TV})$.

3 Results

3.1 Monte Carlo Simulation Settings

The Monte Carlo was conducted by simulating $MC = 10\,000$ samples of size $n = 50$ from the Lomax distribution using a variety of parameter settings, $\sigma = 2$ and $\beta = 1.1, 1.5$. All calculations were conducted using R Core Team (2021). In addition to calculating the MSE for each parameter separately, a combined MSE yielding a single value was also calculated as follows:

$$\text{Accuracy} = \left(\widehat{\beta}_{mc} - \beta\right)^2 + \left(\widehat{\sigma}_{mc} - \sigma\right)^2,$$

TABLE 1. Comparison of different estimation methods for different values of β, $\sigma = 2$ and $n = 50$.

	METHODS	MLE	MLE.b	LME	MDE.CvM	MDE.LS	MDE.Phi.X2	MDE.Phi.tv	MDE.Phi.kl	MME
					$\beta = 1.1, \sigma = 2\ n = 50$					
Mean	$\widehat{\beta}$	1.3013	1.083	1.591	1.2593	1.1278	-2.2916	1.1914	1.3013	2.7323
	$\widehat{\sigma}$	2.3834	1.7327	3.1678	2.2708	1.9488	4.3988	5.9274	2.3834	10.9364
Var	$\widehat{\beta}$	0.2896	0.218	0.1823	0.2943	**0.1605**	402.8793	16.8154	0.2896	0.6134
	$\widehat{\sigma}$	2.3852	1.7039	1.6936	2.4577	**1.3252**	54.7068	390.7247	2.3852	42.2581
Bias	$\widehat{\beta}$	0.2013	**-0.017**	0.491	0.1593	0.0278	-3.3916	0.0914	0.2013	1.6323
	$\widehat{\sigma}$	0.3834	-0.2673	1.1678	0.2708	**-0.0512**	2.3988	3.9274	0.3834	8.9364
MSE	$\widehat{\beta}$	0.3301	0.2183	0.4234	0.3197	**0.1613**	414.3418	16.8221	0.3301	3.2779
	$\widehat{\sigma}$	2.532	1.7751	3.0573	2.5308	**1.3277**	60.4555	406.1103	2.532	122.1127
	Accuracy	2.8621	1.9934	3.4808	2.8505	**1.489**	474.7972	422.9323	2.8621	125.3906
					$\beta = 1.5, \sigma = 2\ n = 50$					
Mean	$\widehat{\beta}$	1.8075	1.4907	1.9476	1.5532	1.3706	1.2414	2.3827	1.7724	3.3631
	$\widehat{\sigma}$	2.5945	1.9332	2.8333	2.1127	1.794	4.8198	8.1012	2.5359	7.4092
Var	$\widehat{\beta}$	28.2018	28.2386	1.8872	2.046	**1.7671**	55.6171	88.5542	6.9246	57.97
	$\widehat{\sigma}$	83.7788	83.8265	6.9077	7.4746	**6.3601**	309.369	1114.9427	24.3918	170.2952
Bias	$\widehat{\beta}$	0.3075	**-0.0093**	0.4476	0.0532	-0.1294	-0.2586	0.8827	0.2724	1.8631
	$\widehat{\sigma}$	0.5945	**-0.0668**	0.8333	0.1127	-0.206	2.8198	6.1012	0.5359	5.4092
MSE	$\widehat{\beta}$	28.2935	28.2358	2.0873	2.0486	**1.7837**	55.6784	89.3244	6.9981	61.4352
	$\widehat{\sigma}$	84.1239	83.8226	7.6013	7.4865	**6.4019**	317.2894	1152.0563	24.6766	199.5376
	Accuracy	112.4173	112.0584	9.6886	9.5352	**8.1856**	372.9678	1241.3807	31.6748	260.9727

3.2 Discussion

When comparing the class of estimators based on minimum distance measures with all of the remaining methods, one can readily see that the MDE.LS and MDE.CvM estimators are the best performers in terms of MSE for the small sample size ; the LME is still an excellent competitor, but the MLE and MLE.b estimators perform relatively poorly in terms of MSE in these cases. However, for larger sample sizes,(tables not shown here) the MLE and MLE.b clearly outperform the entire MDE class of statistics. Interestingly, the LME remains competitive here too, but is almost never found to have the best MSE performance. The overall worse performing estimator is MME regardless of which value of β or σ is used.

References

Giles, D. E., Feng, H., & Godwin, R. T. (2011). On the bias of the maximum likelihood estimator for the two-parameter lomax distribution. *Econometrics Working Paper*, *4*, 3–5.

Pareto, V. (1897). The new theories of economics. *Journal of Political Economy.*, *4*, 485–502.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
`https://www.R-project.org/`

A method for partitioning trends in genetic mean and variance

Oliveira, T.P.[1], Obšteter, J.[2], Pocrnic, I.[1], Gorjanc, G.[1]

[1] The Roslin Institute and Royal (Dick) School of Veterinary Studies, UK; [2] Agricultural Institute of Slovenia, Slovenia

E-mail for correspondence: thiago.oliveira@ed.ac.uk

Abstract: Quantifying sources of genetic change is essential for identifying key breeding actions and optimising breeding programmes. However, the observed genetic change is a sum of contributions from different groups of individuals (often referred to as selection pathways), which are difficult to disentangle and quantify due to the complexity of breeding programmes. Here we extended a simple method to analyse the contributions of groups to the genetic variance. Our approach showed the importance of analysing the partition of the genetic mean and variance rather than just the genetic mean and demonstrated that the contributions are not necessarily independent.

Keywords: Partitioning method; Genetics; Mixed-Effects Models.

1 Background

We aim to genetically improve populations in animal breeding by selecting the best individuals as the next generation's parents. Ideally, we would select the parents based on their true genetic/breeding value, but we can never know that values. Alternatively, we can select parents based on i) phenotypic value, which is the expressed trait and has a medium/low accuracy; ii) estimated breeding value, which may have high accuracy since it considers the phenotypic values of the individuals and all its relatives. Thus, an important step is to understand where genetic progress comes from and which group of animals creates the most genetic gain.
Let $\boldsymbol{a}$ be a vector of breeding values sampled from a normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{A}\sigma_a^2$. We can write $\boldsymbol{a}$ as a linear combination of the individual's ancestors breeding values and individual's deviation from ancestors $\boldsymbol{a} = \boldsymbol{T}\boldsymbol{w}$. We can define $\boldsymbol{T}$ as a triangular matrix of expected

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

gene flow between ancestors and individuals, and $\boldsymbol{w} \sim N\left(\boldsymbol{0}, \boldsymbol{W}\sigma_a^2\right)$ as the Mendelian sampling terms representing deviations, with $\boldsymbol{W}$ being a diagonal matrix of variance coefficients and σ_a^2 the base population genetic (additive) variance. Assuming a factor with p groups and for any set $\sum_{j=1}^{p} \boldsymbol{P}_j = \boldsymbol{I}$, García-Cortés et al. (2008) partitioned the genetic mean into contributions of each level by defining $\boldsymbol{T}_j = \boldsymbol{T}\boldsymbol{P}_j$, and further partitioned the contribution of each group to breeding values *a priori* using the equality $\boldsymbol{a} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_p)\, \boldsymbol{w} = \boldsymbol{a}_1 + \boldsymbol{a}_2 + \ldots + \boldsymbol{a}_p$. García-Cortés et al. (2008) further showed that these partitions can be estimated from data collected in breeding programmes (*posteriori*) by first estimating the breeding values $\widehat{\boldsymbol{a}} = E\left(\boldsymbol{a}|\boldsymbol{y}\right)$ from phenotype data ($\boldsymbol{y}$). Since $\boldsymbol{a}$ is a function of variance components in the mixed-effects model, we can estimate $\boldsymbol{a}$ and $\boldsymbol{w}$ by replacing their REML estimates $\widehat{\boldsymbol{a}} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_p)\, \widehat{\boldsymbol{w}} = \widehat{\boldsymbol{a}}_1 + \widehat{\boldsymbol{a}}_2 + \ldots + \widehat{\boldsymbol{a}}_p$. By summarising these partitions, they quantified the contribution of each group (i.e. males vs. females, countries, AI centres) to the time-trend in genetic mean.

2 Methods

2.1 Partitioning of genetic trends

Here we extend the partitioning method to analyse the contribution of groups to genetic variance. Variance of breeding values is, *a priori*, $Var\left(\boldsymbol{T}\boldsymbol{w}\right) = \boldsymbol{T}\boldsymbol{W}\boldsymbol{T}^T\sigma_a^2$. Thus, we can partition genetic variance as

$$\begin{aligned} Var\left(\boldsymbol{a}\right) = Var\left[\left(\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_p\right)\boldsymbol{w}\right] &= \sum_{j=1}^{p} \boldsymbol{T}_j \boldsymbol{W} \boldsymbol{T}_j^T \sigma_a^2 + 2\sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} \boldsymbol{T}_j \boldsymbol{W} \boldsymbol{T}_{j'}^T \sigma_a^2 \\ &= \sum_{j=1}^{p} \sigma_{a_j}^2 + \sum_{j=1}^{p-1}\sum_{j'=j+1}^{p} \sigma_{a_j, a_{j'}} \end{aligned} \tag{1}$$

While this "theoretical" partitioning involves matrix products, we can also summarise partitions $\boldsymbol{a}_1 + \boldsymbol{a}_2 + \ldots + \boldsymbol{a}_p$ (calculated via $\boldsymbol{T}^{-1}$) by calculating variance of each group level contribution $f\left(\boldsymbol{a}\right) = Var\left(\boldsymbol{a}_j\right)$ and covariance of each pair of group level contributions $f\left(\boldsymbol{a}_j, \boldsymbol{a}_{j'}\right) = Cov\left(\boldsymbol{a}_j, \boldsymbol{a}_{j'}\right)$. The partitions can be summarised in many ways to quantify the contribution of different groups to change in genetic variance over time. The partitioning method can be then only applied for *a priori* or true breeding values. Although the methodology has been developed to *a priori* or true breeding values, we can use methods from Sorensen et al. (2001) to estimate partitions of genetic variance from data collected in a breeding programme (*posteriori*).

2.2 Statistical model and computational approaches

In the previous subsection, we assumed we knew the true breeding values. Consequently, the same assumption is applied to additive genetic mean and

variance contributions. However, in reality, we use phenotype, pedigree, and genomic information to predict the breeding values $\hat{\boldsymbol{a}}$ and make inferences. We fitted standard animal model:

$$\begin{gathered}\boldsymbol{y}=\boldsymbol{X}\boldsymbol{b}+\boldsymbol{Z}\boldsymbol{a}+\boldsymbol{e}, \\ \boldsymbol{a}\sim N\left(\mathbf{0},\sigma_a^2\boldsymbol{A}\right), \text{ and } \boldsymbol{e}\sim N\left(\mathbf{0},\sigma_e^2\boldsymbol{I}\right)\end{gathered} \tag{2}$$

where $\boldsymbol{y}$ is a vector of observed phenotypes, $\boldsymbol{b}$ is a vector of fixed effects with design matrix $\mathbf{X}$, $\mathbf{a}$ is a vector of random animal effects with design matrix $\mathbf{Z}$, and $\mathbf{e}$ is a vector of random residuals. It is assumed that $\mathbf{a}\sim N\left(\mathbf{0},\mathbf{A}\sigma_a^2\right)$ and $\mathbf{e}\sim N\left(\mathbf{0},\mathbf{I}\sigma_e^2\right)$, where $\mathbf{A}$ is the pedigree-based numerator relationship matrix, while σ_a^2 and σ_e^2 are respectively known additive and residual variances.

The directed acyclic graph (DAG) representation of the model (2) considering only intercept as the fixed effect is illustrated in Figure 1, where pedigree and phenotypic records are displayed in separate plates as a generalization of the case where animals might not have phenotypic records. In addition, the dotted lines indicate a possibly missing parent in the pedigree. In the pedigree plate we have K individuals represented by founders and non-founders, where founders is a `priori` sampled from $a_k|\sigma_a^2\sim\left(0,\sigma_a^2\right)$. Non-founders individuals given the information of their parents are then represented by $a_k=1/2\left(a_{f(k)}+a_{m(k)}\right)+w_k$, where $a_{f(k)}$ and $a_{m(k)}$ are parent's breeding value and w_k represents the Mendelian sampling term $(w_k|\boldsymbol{W}_{k,k}\sim N\left(\mathbf{0},\sigma_a^2\boldsymbol{W}_{k,k}\right))$.

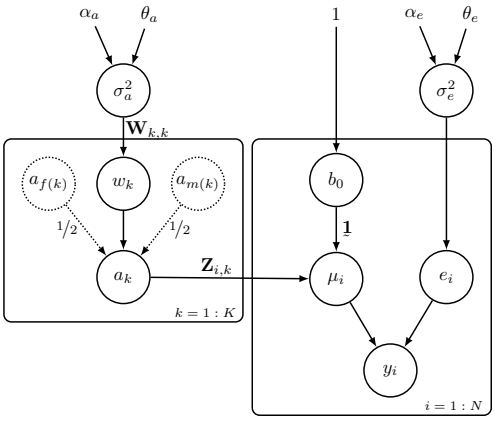


FIGURE 1. Directed acyclic graph of the animal model with nI individuals and nY phenotypic records (y_i) with explicit representation of Mendelian sampling terms (w_k) and error term (e_i), where σ_a^2 is the additive genetic variance, $a_{f(k)}$ and $a_{m(k)}$ are parent's breeding value, $\mathbf{1}$ represents a vector of ones, μ_i the linear predictor, and σ_e^2 the variance of the error term

In this sense, matrix $\boldsymbol{A}$ can be decomposed as $\boldsymbol{A}=\boldsymbol{T}\boldsymbol{W}\boldsymbol{T}^T$ using LDL decomposition Golub and Van Loan (1996), as described in section 2.1. The

diagonal elements of $\boldsymbol{W}$ can be computed according to specific scenarios described by Mrode (2005) as i) $\boldsymbol{W}_{k,k} = \frac{1}{2} - \frac{1}{4}\left(F_{f(k)} + F_{m(k)}\right)$ when both parents are known; ii) $\boldsymbol{W}_{k,k} = \frac{3}{4} - \frac{1}{4}F_{m(k)}$ or $\boldsymbol{W}_{k,k} = \frac{3}{4} - \frac{1}{4}F_{f(k)}$ when one parent are known; and iii) $\boldsymbol{W}_{k,k} = 1$ when both parents are unknown, where $F_{f(k)}$ and $F_{m(k)}$ are the coefficients of inbreeding related to the father and mother identification of the individual k, respectively Kennedy et al. (1988); Falconer and Mackay (1996); Mrode (2005).
Accounting for inbreeding when computing $\boldsymbol{A}^{-1}$ may impact the partitioning results for genetic variance according to the inbreeding level because, for any domain D, we have $Var\left(a_k|\boldsymbol{W}_{k,k}\right) = \int_D a_k^2 \Pr\left(a_k\right) da_k - \left[\int_D a_k \Pr\left(a_k\right) da_k\right]^2 = \left(1 + F_k\right)\sigma_a^2$, where $\Pr(.)$ represents a probability density function, and F_k is the inbreeding coefficient of the kth individual. Thus, we decided to include two more scenarios i) accounting and ii) not accounting for inbreeding when constructing $\boldsymbol{A}^{-1}$. In the case of ignoring inbreeding the $\boldsymbol{W}_{k,k}$ is equal to $\frac{1}{2}$, $\frac{3}{4}$ and 1, respectively Mrode (2005).
We used the full Bayesian approach by specifying prior distribution for all model parameters, as shown in Figure 1. Thus, $\boldsymbol{b}$, σ_a^2 and σ_e^2 are assumed to have a joint prior density of the form $p\left(\boldsymbol{b}, \sigma_a^2, \sigma_e^2\right) = p\left(\boldsymbol{b}\right) p\left(\sigma_a^2\right) p\left(\sigma_e^2\right)$, where $p\left(\tau_a = 1/\sigma_a^2|\alpha_a, \theta_a\right) \propto \tau_a^{\alpha_a - 1}\exp\left(-\theta_a\tau_a\right)$, $p\left(\tau_e = 1/\sigma_e^2|\alpha_e, \theta_e\right) \propto \tau_e^{\alpha_e - 1}\exp\left(-\theta_e\tau_e\right)$, and $p\left(\boldsymbol{b}\right) \propto 1$, with $\tau_a > 0$, $\alpha_a \geq 0$, $\theta_a \geq 0$, $\tau_e > 0$, $\alpha_e \geq 0$ and $\theta_e \geq 0$. In this case, we are assuming a flat prior distribution for β which is independent of σ_a^2 and σ_e^2. On the other hand, the inverse-gamma(α, θ) is a natural candidate for the prior distributions for variance components, and when α and θ are set to a value such as 0.1^3, it can be considered as vague prior within the conditionally conjugate family $\sigma_a^{-2}, \sigma_e^{-2} \sim$ Gamma $\left(0.1^3, 0.1^3\right)$. The posterior distribution can be obtained by applying the form of Bayes' theorem conditional on the data:

$$p\left(\boldsymbol{b}, \boldsymbol{a}, \sigma_a^2, \sigma_e^2|\boldsymbol{y}\right) \propto p\left(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{a}, \sigma_e^2\right) p\left(\boldsymbol{b}\right) p\left(\boldsymbol{a}|\boldsymbol{A}, \sigma_a^2\right) \times p\left(\sigma_a^2|\alpha_a, \theta_a\right) p\left(\sigma_e^2|\alpha_e, \theta_e\right).$$

We used Markov Chain Monte Carlo (MCMC) to generate samples from the posterior distribution using Gibbs sampler algorithm Sorensen et al. (2001). It was considered one chain with 80,000 samples, from which 20,000 iterations are burn-in while the remaining 60,000 were stored using a thinning of length 40. Consequently, 1,500 samples of EBV's are computed observing the posterior distribution $p\left(\boldsymbol{a}|\boldsymbol{A}, \sigma_a^2\right)$, which are passed as input for the `AlphaPart` package. We assessed MCMC convergence by looking at trace and autocorrelation function plots. Gibbs sampling was executed by GIBBS1F90 software Misztal et al. (2018).

3 Results and Discussion

The simulated cattle breeding programme illustrated the power of the partitioning method to summarise genetic trends in mean and variance, although some care is needed when using the proposed methodology. By partitioning the genetic mean and variance we showed that in a high accuracy scenario the covariance between females (F) and selected males (M) plays an important role in the contribution to the genetic variance and, consequently, in this case $Var\left(\boldsymbol{a}\right) < Var\left(\boldsymbol{a}|\mathrm{F}\right) + Var\left(\boldsymbol{a}|\mathrm{M}\right)$. In this sense, we demonstrated that the choice of groups is essential and that contributions are not necessarily independent; hence, they should not be analyzed in isolation from each other.

The advantage of combining the MCMC approach with the partition method presented here is related to drawing samples from the posterior distribution $p\left(\boldsymbol{a}|\boldsymbol{A}, \sigma_a^2\right)$ and using them to compute the point estimate partitions for genetic mean and variance and also access their uncertainty. Although the methodology presented here works fine for the extreme example proposed using medium accuracy, we again expect the reproducible inaccuracy showed in Figure 2 can be overcome with an extension of the partition method using genomic models.

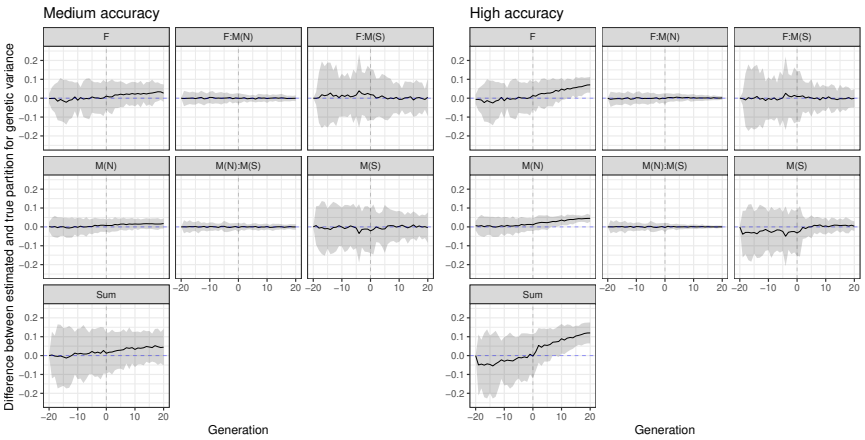


FIGURE 2. Distribution of the difference between true and estimated partitions for the total additive genetic variance (Sum) over generations by gender (male (M) and female (F)) and status (selected males (S) and non-selected males (N)) considering 30 simulations replicate

4 Conclusion

We developed a method for quantifying sources of genetic variance. This is a powerful and valuable method for understanding how different breeding

groups interact within a breeding programme, and hence for optimising breeding programmes. By partitioning the genetic variance in a simulated cattle breeding programme, we showed that the covariance between paths can make a substantial contributions to the genetic variance. Hence, to comprehend and manage the genetic variance in a breeding programme, we should not consider the contribution of different groups in isolation but should perform a holistic analysis and partition of the observed genetic variance instead.

Acknowledgments: The authors acknowledge support from the BBSRC, European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801215, Roslin Institute Strategic Programme (ISP) grants, and Core financing of the Slovenian Research Agency (grant P4-0133).

References

Falconer, D. S. and T. F. C. Mackay (1996). *Introduction to Quantitative Genetics* (4 ed.). Essex: Longman.

García-Cortés, L., J. Martínez-Ávila, and M. Toro (2008). Partition of the genetic trend to validate multiple selection decisions. *Animal 2*(6), 821–824.

Golub, G. H. and C. F. Van Loan (1996). *Matrix Computations*. Baltimore: John Hopkins University Press.

Kennedy, B. W., L. R. Schaeffer, and D. A. Sorensen (1988). Genetic Properties of Animal Models. *Journal of Dairy Science 71*, 17–26. Publisher: Elsevier.

Misztal, I., S. Tsuruta, D. A. L. Lourenco, Y. Masuda, I. Aguilar, A. Legarra, and Z. Vitezica (2018). Manual for BLUPF90 family programs.

Mrode, R. A. (2005). *Linear models for the prediction of animal breeding values* (2 ed.). Wallingford: CAB International.

Sorensen, D., R. Fernando, and D. Gianola (2001, February). Inferring the trajectory of genetic variance in the course of artificial selection. *Genetical Research 77*(1), 83–94.

A combination technique for unbalanced binary classification using artificial neural network

Hyebin Park[1], Juyong Hong[1]

[1] Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea

E-mail for correspondence: hyebinpark000@gmail.com

Abstract: The classification problems for the imbalance data occur frequently in our lives, and it is important to solve them well. Therefore, we propose a method combining the generalized extreme value (GEV) activation function and the cost-sensitive learning method and over-sampling in a simple neural network model. In order to check the performance of the proposed method, 65 data sets were employed and 5 evaluation metrics were considered. Five models including the proposed model were created and performance evaluation was performed. Then the number of the best result and second best result for each evaluation indicators in the entire datasets was counted and compared. As a result, generally excellent results were obtained in five evaluation indicators when the proposed method model was used.

Keywords: Activation function; Class imbalance; Over-sampling; Sigmoid function; KEEL imbalance dataset.

1 Introduction

Nowadays, classification problem is a very important problem that occurs very often. Traditional classification algorithms assume that the number of samples between classes is approximately equal. But in reality, that is rarely the case. Such a case in which a specific class appears more frequently than other classes is said to be a class imbalance problem, and it exists in real life such as medical diagnosis, fire detection and fraudulent transaction detection. Recently, there are some trials using the GEV activation function to solve class imbalance problem. Wang et al.(2010) used GEV as the link function of generalized linear model(GLM), and Munkhdalai et

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

al. (2020) improved classification performance by using a neural network model that has Gumbel distribution as an activation function. Recently, Bridge et al. (2020) used GEV activation function in a convolution neural network (CNN) model that diagnoses COVID-19.

2 Methods

When solving a classification problem using a multi-layer perceptron, we often use sigmoid as activation function. The sigmoid function is calculated as follows and has a symmetrical structure as shown in Figure 1 (right).

$$sigmoid(x) = 1/(1 + e^x) \tag{1}$$

Our proposed method is to use the cumulative distribution function (CDF) of the GEV distribution as the activation function instead of the sigmoid function, because it makes all real inputs to a value between 0 and 1. The GEV distribution has three parameters and in this study, each parameters were estimated using back propagation method with the weights of the neural network model. The GEV activation function is calculated as follows, and has an asymmetric structure as shown in Figure 1 (left).

$$G(x) := \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{2}$$

where $-\infty < (x-\mu)/\sigma < \infty, -\infty < \mu < \infty, \sigma > 0, -\infty < \xi < \infty$.

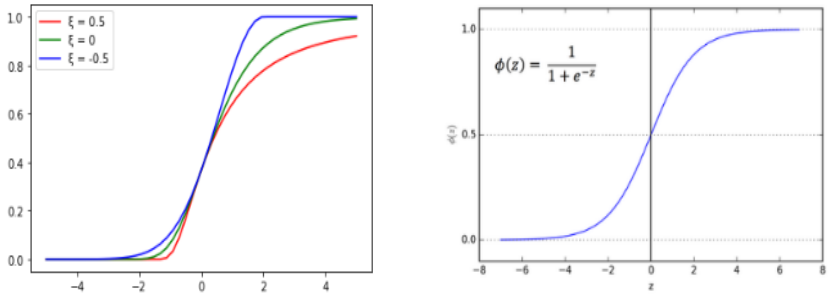


FIGURE 1. The CDF of generalized extreme value (GEV) activation function (left) and sigmoid activation function (right).

To compare the performance of the proposed method, we considered the following 5 multi-layer perceptron (MLP) models for 100 KEEL imbalanced data sets.

1. (MLP) sigmoid activation (baseline)
2. (MLP) GEV activation function

3. (MLP) GEV activation and Thresholding
4. (MLP) GEV activation, Thresholding and Focal Loss
5. (MLP) GEV activation, Thresholding, Focal Loss and Over-Sampling

The data used in this experiment are shown in Table 1. The asymmetry ratio was calculated by dividing the number of majority class samples by the number of minority class samples (N), larger this value means the more severe asymmetry.

TABLE 1. 5 out of 65 KEEL imbalance data sets (using data)

Data name	N	Input variable number	Imbalance ratio
abalone19	4,174	8	129.44
abalone20	1,916	8	72.69
kr vs k zero vs fifteen	2,193	19	80.22
pocker 8 vs 6	1,477	9	85.88
pocker 8 9 vs 5	2,075	9	82.00

For a more reliable result, the average of the results obtained by changing the seed (30 times) was compared, and for each data, 5 evaluation indicators (follows) suitable for unbalanced data were evaluated. The neural network model was used in the experiment, and the 5 evaluation indicators are shown in follows. All of indicators, the higher the value, the better. For the reliability of comparison, all hyper parameters such as batch size were made the same.

F1-score $= \frac{2\times(\text{Recall}\times\text{Precision})}{\text{Recall}+\text{Precision}}$
Geometric-Mean (GM) $= \sqrt{\text{TRP(Recall)} \times \text{TNR(Specificity)}}$
Balanced Accuracy (BA) $= \frac{1}{2} \times$ TPR+TNR
Area Under the ROC Curve (AUC)
Brier Inaccuaracy (BI) $= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{l}(\hat{p}(c=j,x^i) - p(c=j,x^i))^2$

3 Results

A summary of the experimental results is shown in Table 2. we counted the number of better results when we compared method 1 and 5. As a result, the results of the proposed method received higher scores in more evaluation indicators than other models. However it's never been nice if the GEV activation function was used alone. In combination models, the larger the model number, the better the results. So we can see that using combination models performs better on imbalanced data.

TABLE 2. Example of experiment result (data: abalone19)

Method	F1-score	GM	AUC	BC	2-BI
(1)	0.0	0.0	0.794	0.5	1.984
(2)	0.0	0.0	0.659	0.5	1.81
(3)	0.033	0.662	0.659	0.657	0.81
(4)	0.045	0.733	0.76	0.757	1.963
(5)	0.044	0.762	0.781	0.770	1.961

4 Summary and Discussion

We combined a GEV activation function with oversampling in a cost-sensitive learning method and a simple neural network model to better predict imbalanced data. Performance evaluation was considered through 5 evaluation indicators by creating 5 models including the proposed model. As a result, generally excellent results were obtained when the proposed method model was used. However the superiority of the model is shown differently depending on the evaluation index, a comparison method that considers the characteristics of the model and data is needed. Better results can be expected if the two hyperparameters for Focal Loss are adjusted.

Acknowledgments: This study was supported by the NRF grant funded by the Korea government (MSIT)(No.2020R1I1A3069260) and BK21 FOUR (NO.5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

References

Wang, X., and Dey, D.K. (2010). *Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption.*. Ann. Appl. Stat. **4(4)**, 2000–2023.

Munkhdalai, L., Munkhdalai, T., and Ryu, K.H. (2020). *GEV-NN: A deep neural network architecture for class imbalance problem in binary classification.*. Knowledge-Based Systems. 194, 1–14.

Bridge.J,Meng.Y, Zhao.Y, Du.Y, Zhao.M and Sun.R. (2020). In: *Introducing the GEV Activation Function for Highly Unbalanced Data to Develop COVID-19 Diagnostic Models.*, IEEE Journal of Biomedical and Health Informatics. **24(10)**, 2776–2786.

Johnson, J.M., and Khoshgoftaar, T.M. (2019). *Survey on deep learning with class imbalance.*, Journal of Big Data. **6(27)**, 1–54.

Fast computation for performance and independence weighting of climate multi-models

Jeong-Soo Park[1], Thanawan Prahadchai[1]

[1] Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea

E-mail for correspondence: tanawanp.st@gmail.com

Abstract: Scientists occasionally predict the future changes in climate using multi-model ensemble methods that combine predictions from individual simulation models. We employed a model weighting method that accounts for model performance and independence (PI-weighting). In calculating the PI-weights, two shape parameters should be determined, but usual perfect model test method requires a considerable computing time. To address this trouble, we suggest simple ways for selecting two shape parameters based on the chi-square statistic and the entropy, which reduce the computing time greatly. Our method is applied to 21 CMIP6 (the Coupled Model Inter-Comparison Project Phase 6) models for five climate variables over East Asia.

Keywords: Climate change; Dirichlet distribution; Generalized extreme value distribution; Leave-one-out cross validation; Return period.

1 Introduction

Studies on the projection of future climate change have used ensembles of multiple climate simulations. Model averaging or ensemble is a statistical method in which unequal or equal weights are assigned to those models. Despite some arguments, the equal weighting or "model democracy" has been criticized because it does not take into account the performance, uncertainty, and independency of each model in constructing an ensemble.

In addition to the performance, some researchers have considered other criteria such as model independency (Knutti et al., 2017; Lorenz et al., 2018). A weighting scheme that accounts for both the independence and performance simultaneously is called the PI-weighting. In this study, we

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

employ PI-weighting to robustly quantify uncertainty in MME. In applying the PI-weighting, we have to determine two shape parameters. One way to select the shape parameters is a leave-one-out perfect model test (Lorenz et al., 2018; Brunner et al., 2019), but it requires huge computing time. To overcome this trouble, we suggest simple ways to determine these parameters based on the entropy and p-values of the chi-square statistic.

2 Performance and independence weighting: Determination of σ_S and σ_D

Weights are calculated for each model based on a combination of the distance D_i (informing the performance) and the model similarity S_{ij} (informing the dependence) (Knutti et al., 2017):

$$w_i = \frac{exp(-\frac{D_i}{\sigma_D})}{1 + \sum_{j \neq i}^{M} exp(-\frac{S_{ij}}{\sigma_S})}, \tag{1}$$

with the total number of model runs M and the shape parameters σ_D and σ_S. The shape parameters are often determined through a perfect model test (or a model-as-truth experiment) using the continuous rank probability score (Lorenz et al., 2018; Brunner et al., 2019). This leave-one-out procedure requires huge computing time. To address this computational trouble, we consider relatively simple ways to determine the shape parameters.

To select an appropriate value of the shape parameter σ_S for the I-weights, we consider an entropy-based approach. Denote $I_i(\sigma_S)$ as a normalized I-weight for model i and for the given σ_S. The entropy of the I-weights as a measure of uncertainty (Ross, 2010) from these weights is defined by the following:

$$E(\sigma_S) \; = \; -\sum_{i=1}^{M} I_i(\sigma_S) \log I_i(\sigma_S) \tag{2}$$

as a function of σ_S. When all $I_i(\sigma_S)$s are almost equal, the entropy has a high value. We thus expect the entropy to increase because σ_S has a large value. Figure 1 presents the entropy function of σ_S computed from the data used for this study, which indicates that it is minimum at $\sigma_S = 0.4$. It is interesting to note that the entropy function increases as σ_S decreases from 0.4 to zero. Thus, s_i moves toward one, and I_i is close to $1/M$ for all i. Because we want to have a shape parameter σ_S that can differentiate the I-weights most distinctly with minimum uncertainty, the value $\sigma_S = 0.4$ minimizing the entropy is chosen in this study.

To determine σ_D, a technique based on the p-value of the chi-square statistic is considered in this study. Denote $P_i(\sigma_D)$ as a normalized P-weight for model i and for the given σ_D. For testing the hypothesis frame, the null hypothesis is that all weigths are equal, and the alternative hypothesis is that some weights are not equal. For the given P_i,

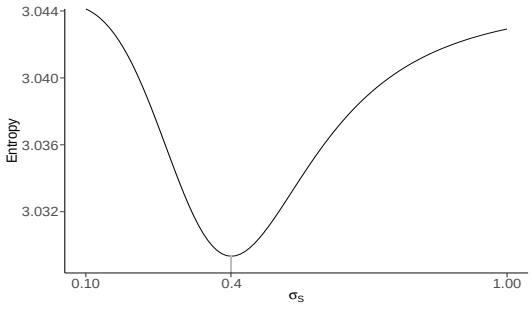


FIGURE 1. Plot of the entropy as the parameter σ_S changes from 0.1 to 1.0, and the selected $\sigma_S = 0.4$.

the chi-square statistic used to test the above hypothesis is as follows: $\chi_0^2(\sigma_D) = \sum_{i=1}^{M} (\frac{1}{M} - P_i(\sigma_D))^2 / \frac{1}{M}$.

Because we do not want to accept equal weights, σ_D should be selected to reject the null hypothesis. In addition, because we also do not want aggressive weights, a σ_D can be selected as the maximum value of σ_D in which we still reject H_0 with α level. That is, our selection is $\sigma_D^* = max\ \{\sigma_D : p - value\ (\sigma_D) < \alpha\}$, where $p - value\ (\sigma_D) = Pr[\chi^2 > \chi_0^2(\sigma_D) |\ H_0]$.

The p-values are computed by a Monte-Carlo simulation in which random numbers of weights are generated from the Dirichlet distribution. Figure 2 depicts the chi-square statistic values computed from AMP1 with some p-values as σ_D. We calculated the σ_D for each of the five climate variables, and then calculated the average from those five σ_Ds. When $\alpha = 0.05$ as is usually applied in statistics, the averaged σ_D^* from five different σ_D is 0.21.

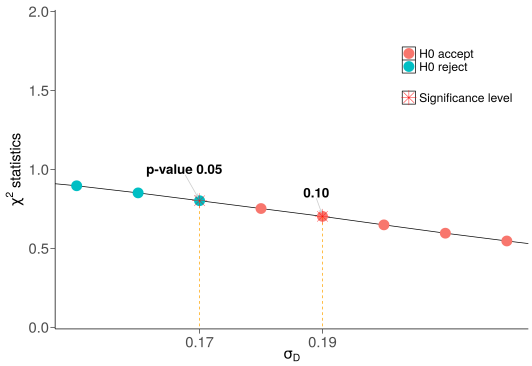


FIGURE 2. Plot of the chi-square statistic values as the parameter σ_D changes, for the annual maximum daily precipitation (AMP1). The selected σ_D is 0.17 (0.19) for p-value 0.05 (0.01).

3 Results: model weights

The normalized PI-weights are obtained using Eq.(1), with $\sigma_S = 0.4$ and $\sigma_D = 0.21$. Figure 3 demonstrates the distributions of the P-, I-, and PI-weights. The variability of the I-weights is smaller than that of the P-weights. The high P-weights of the CanESM5 and EC-Earth3-Veg models decrease in PI-weights owing to the low I-weights. The PI-weights of BCC-CSM2-MR, FGOALS-g3, and GFDL-ESM4 models increase owing to a relatively high independency. The performance is more influential to the PI-weights than the independency.

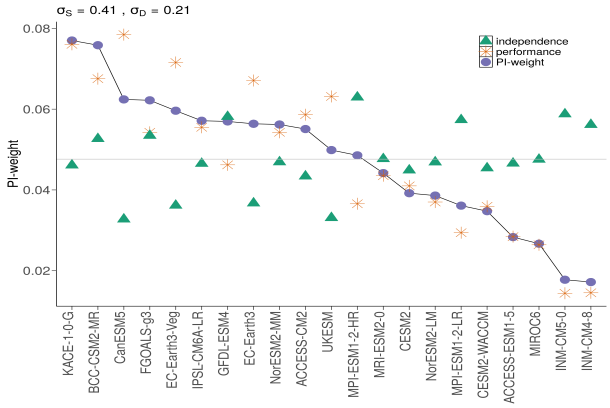


FIGURE 3. Spread of the weights for 21 CMIP6 (the Coupled Model Inter-Comparison Project Phase 6) models obtained based on the performance only, the independence only, and by both the performance and independence. The weights are obtained from five climate variables over East Asia.

Acknowledgments: This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research, NO.5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

References

Knutti, R., Sedlacek, J., Sanderson, B.M., Lorenz, R. et al. (2017). A climate model projection weighting scheme accounting for performance and independence. *Geophys. Res. Lett.*, **44**, 1909 – 1918.

Lorenz, R., Herger, N., Sedlacek, J., Eyring, V. et al. (2018). Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *J. Geophys. Res. Atmos.*, **123**, 4509 – 4526.

Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R. (2019). Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environ. Res. Lett.*, **14**, 124010.

Ross, S. (2010). *A First Course in Probability, 8th ed.; Pearson Prentice Hall: Upper Saddle River*. USA, NJ.

Performance of spatio-temporal hierarchical P-spline models using simulated data

Diana M. Pérez-Valencia[1], María Xosé Rodríguez-Álvarez[2], Martin P. Boer[3], Fred A. van Eeuwijk[3]

[1] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[2] Dept. of Statistics and Operations Research, Universidade de Vigo, Vigo, Spain
[3] Biometris, Wageningen University & Research, Wageningen, The Netherlands

E-mail for correspondence: dperez@bcamath.org

Abstract: When comparing one- and two-stage P-spline-based approaches for analysing spatio-temporal hierarchical data from high-throughput phenotyping experiments, a critical issue is to develop a good data simulation strategy. We present a strategy that is independent from the statistical methods used to fit the data. We find that for most simulated situations there was no clear difference between the two approaches compared.

Keywords: P-splines; Hierarchical models; Space-time models; Data simulation

1 Introduction

This work is motivated by the need for comparing two P-spline-based (Eilers and Marx, 1996) approaches when analyzing spatio-temporal data from high-throughput phenotyping (HTP) experiments, with a three-level nested hierarchical structure (plants nested in genotypes, and genotypes nested in populations). The focus is on analysing the evolution over time of the genetic signal on a given phenotype. To simulate this data, we decompose the spatio-temporal variation of the phenotype of interest in three components (for simplicity, we consider one population): within genotypes and plant variation, and spatio-temporal correlated noise. We then compare two modelling strategies: the two-stage approach (TSapp) proposed by Pérez-Valencia et al. (2022), in which they correct for experimental design factors and spatial variation in the first stage, while estimating the evolution over time of the genetic signal in the second stage; and the full and one-stage spatio-temporal approach (OSapp, Pérez et al., 2021) that generalizes the

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

previous two-stage approach. When applying these approaches to real data sets, two problems arise: (i) the phenotype of interest is not measured at the population and genotype levels but only at the plant level; it makes evaluating models' performance at these levels difficult, and (ii) the approaches seem to be sensitive to the dimension of the B-spline bases used at each level of the hierarchy. This work presents a data generating mechanism and a simulation experiment to study the above-mentioned problems.

2 Data simulation strategy

Let $y_i(t)$ denote the (simulated) phenotype of interest for the ith plant, at time $t \in \{t_1, \ldots, t_n\}$, at the spatial (row and column) position $s = (r(i), c(i))$. We simulate HTP data assuming the following three-level nested hierarchical structure

$$y_i(t) = f_{p(i)}(t) + f_{g(i)}(t) + f_i(t) + \varepsilon_i(s,t),\ p = 1,\ 1 \leq g \leq L,\ 1 \leq i \leq M,$$

with $f_{p(i)}(\cdot)$ the population trajectory, and genotype-specific deviations, plant-specific deviations and spatio-temporal correlated noise curves given by $f_{g(i)}(\cdot) \sim N(\mathbf{0}, \mathbf{\Sigma}_{geno})$, $f_i(\cdot) \sim N(\mathbf{0}, \mathbf{\Sigma}_{plant})$, and $\varepsilon_i(\cdot,\cdot) \sim N(\mathbf{0}, \Sigma_\varepsilon)$, respectively. Figure 1 depicts the kind of curves that are obtained at each step of the simulation. The data is generated from the population to the plant level, as follows:

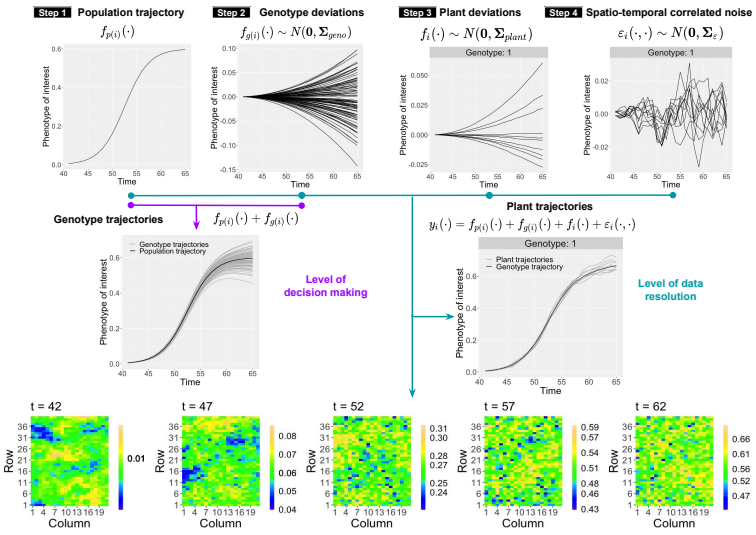


FIGURE 1. Data generating strategy.

Step 1. Generate one population trajectory from the growth logistic curve model, $f_{p(i)}(t) = \frac{a}{1+\exp(c(b-t))}$, where a is the asymptote, b is the inflection point, and c is the growth rate.

Step 2. Generate L genotypic deviations, $f_{g(i)}(\cdot)$. For the covariance matrix $\mathbf{\Sigma}_{geno}$, use an ARH(1) structure (Wolfinger, 1996) with $\mathbf{d}$, a euclidean distance matrix between time points ($d_{jk} = |t_k - t_j|$); σ^2_{geno}, the between genotypes variability; $S(\mathbf{t}) = \sigma^2_{geno} h(\mathbf{t})$, a heterogeneous variance function, with variance increasing as a quadratic function of time, $h(\mathbf{t})$; and autocorrelation parameter, ρ. Then, the ARH(1) covariance for two points j and k separated by distance d_{jk} (in time) and autocorrelation ρ is

$$\Sigma_{geno_{jk}} = \frac{1}{1-\rho^2} S_{jk} \rho^{d_{jk}}, \quad \mathbf{S} = (S(\mathbf{t}) S^{\mathrm{T}}(\mathbf{t}))$$

Step 3. Generate M plant deviations, $f_i(\cdot)$. For the covariance matrix $\mathbf{\Sigma}_{plant}$, follow the same ideas used in **Step 2**. Use σ^2_{plant} for the between plants variability, and $S(\mathbf{t}) = \sigma^2_{plant} h(\mathbf{t})$.
Step 4. Generate M spatio-temporal correlated noise curves, $\varepsilon_i(\cdot, \cdot)$. Use the space-time separable covariance model $\Sigma_\varepsilon = C(s_{rc})C(d_{jk})$, which is the product of a Matérn spatial covariance function (Guttorp and Gneiting, 2006) and a temporal ARH(1) covariance function, and s_{rc} is the distance between two plants locations, d_{jk} is the temporal lag between two timepoints, $S(\mathbf{t}) = \sigma^2 h(\mathbf{t})$, and

$$C(s_{rc}) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{s_{rc}}{\nu}\right)^{\kappa} \kappa_\kappa \left(\frac{s_{rc}}{\nu}\right), \quad C(d_{jk}) = \frac{1}{1-\rho^2} S_{jk} \rho^{d_{jk}},$$

where $\kappa_\kappa(\cdot)$, $\Gamma(\cdot)$, $\kappa > 0$, and $\nu > 0$ are parameters of the Matérn function. Genotypes are assigned to spatial positions following a completely randomised block design.

3 Performance assessment of spatio-temporal hierarchical P-spline models

In brief, data is simulated under eight different scenarios: four levels for the between genotype and plant variations (for two given variances σ^2_1 and σ^2_2, with $\sigma^2_1 < \sigma^2_2$, the four possible combinations of σ^2_{geno} and σ^2_{plant}) and two levels for the number of plants per genotype ($m_{pg} = 3, 10$). For each scenario, 100 datasets are generated. We then assessed the performance of the two modelling strategies (OSapp and TSapp) and five different configurations for the dimensions of the B-spline bases for the hierarchical components (b_p,b_g,b_i) while keeping fixed the B-spline bases for the spatio-temporal components of the approaches. We focus here in the results at the genotype level, which is the decision-making level in the agriculture context. We use the logarithm of the root mean square error (log(RMSE)) as performance measure to compare the simulated and the estimated genotype-specific deviations. Figure 2 shows that small differences appear between the two approaches, except when non-nested B-spline basis configuration (13,9,7) is used for scenarios with 10 replicates per genotype.

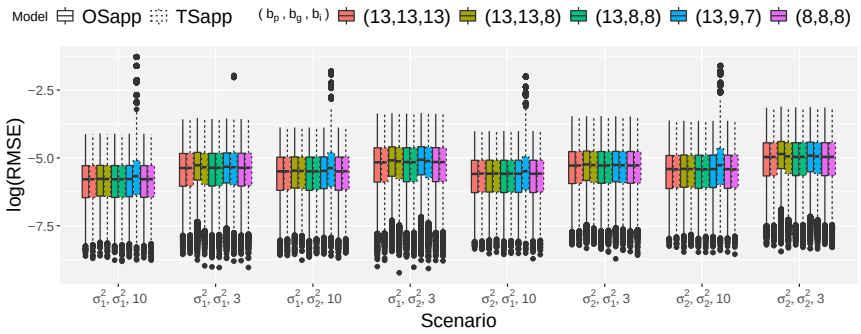


FIGURE 2. Comparison of the simulated and estimated genotype deviation curves for eight scenarios ($\sigma^2_{geno}, \sigma^2_{plant}, m_{pg}$) of data simulation, using the two modelling approaches (TSapp and OSapp), and five B-spline basis configurations (b_p,b_g,b_i) for functions at population, genotype and plant level, respectively.

Acknowledgments: This research was supported by the Basque Government through the BERC 2022-2025 program, by the Spanish Ministry of Science, Innovation and Universities through BCAM Severo Ochoa accreditation SEV-2017-0718.

References

Eilers, P. H., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89 – 102.

Guttorp, P. and Gneiting, T. (2006). Studies in the history of probability and statistics XLIX on the matérn correlation family. *Biometrika*, **93(4)**, 989 – 995.

Pérez, D. M., Rodríguez-Álvarez, M. X., Boer, M. P., Millet, E. J. and van Eeuwijk F. A. (2021). Spatio-temporal and hierarchical modelling of high-throughput phenotypic data In: *Proceedings of the 35th International Workshop on Statistical Modelling. Volume II*, Bilbao, Spain, 20 – 24.

Pérez-Valencia, D. M., Rodríguez-Álvarez, M. X., Boer, M. P., Kronenberg, L., Hund, A., Cabrera-Bosquet, L., Millet, E.J. and van Eeuwijk, F. A. (2022). A two-stage approach for the spatio-temporal analysis of high-throughput phenotyping data. *Scientific Reports*, **12(1)**, 1 – 16.

Wolfinger, R. D. (1996). Heterogeneous variance: covariance structures for repeated measures. *Journal of agricultural, biological, and environmental statistics*, 205 – 230.

'Predicted conversion rate' (PCR) of neoadjuvant treatment in relation to the (y)pN stage of HER2–positive breast cancer cases

Christian Pfeifer[1], Sabine Danzinger[2], Christian Singer[2]

[1] Department of Statistics and Economics, University of Innsbruck, Austria
[2] Department of Obstetrics and Gynecology, Medical University of Vienna, Austria

E-mail for correspondence: christian.pfeifer@uibk.ac.at

Abstract: In this approach we introduce a novel quantity which we call 'predicted conversion rate' (PCR) and give an application in medical statistics.

Keywords: HER2–positive breast cancer; Logistic regression.

1 Introduction

Fifteen to twenty percent of early breast cancer (BC) cases are HER2 (Human Epidermal Growth factor–Receptor 2) - positive which are associated with a more aggressive clinical course (including a higher risk of recurrence) (Wuerstlein and Harbeck (2017), Gianni et al. (2012)). There are several trials indicating that an adjuvant chemotherapy in combination with trastuzumab and pertuzumab before surgery significantly improves invasive disease-free survival in operable, HER2-positive BC with increased risk of recurrence (positive lymph node (LN) stage or negative hormone receptors; see von Minckwitz et al. (2012) or Boland et al. (2017)). It seems to be very likely that a neoadjuvant chemotherapy (NACT) results in 'downstaging' the axillary lymph node (LN) stage. Unfortunately, at present, the knowledge of LN stage prior to NACT is insufficient due to the not yet established bioptic-histological diagnostics prior to the therapy, for illustration see Fig. 1.
A question of fundamental medical interest, however, is: How many cases could be converted from pN-stage 'positive' to ypN-stage (y stage)

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

FIGURE 1. Therapeutic process in case of NACT and pN diagnosis

'negative' applying the NACT?

2 Data

The dataset used in our approach includes **153 early HER2+ BC cases from the Breast Health Center Vienna II (Vienna General Hospital) within 2012 – 2017** examined retrospectively. The data are stratified by NACT/without NACT. Inclusion criteria are:

- primary early BC, HER2+
- diagnosis between 2012 and 2017
- surgery carried out
- medical care by the Breast Health Center Vienna II
- female
- age equal to or larger than 18

Beside NACT/without NACT, age at diagnosis, tumor morphology, tumor grade, estrogen receptor, progesterone receptor, Ki67 of the pretherapeutic biopsy and (y)pN stage (pN0-pN3) were reported, where (y)pN-stage denotes: pN-stage in case of no NACT (no treatment in between, prior to the therapy) and ypN-stage in case of NACT (after treatment/therapy). As a result of the descriptive analysis we notice a substantially smaller (y)pN1-share in the NACT group.

3 Model

In a first step, we build a logistic regression model

$$(y)pN01 = \texttt{NACT} + \texttt{other meaningful predictors}$$

in order to estimate the number of pN1 cases without NACT, where (y)pN01 denotes the dichotomous variable {(y)pN equal to zero, (y)pN$\geq$ 1} and `NACT` the variable indicating NACT/without NACT. We, however, found only 1 additional predictor of significant relevance: The progesterone receptor status PR-negative(PR0) versus PR-positive(PR1) .
As mentioned above, we do not have any information about the pN stage before the treatment in the NACT group. To overcome this lack of information we are using the estimated probabilities in the no therapy group in order to predict the pN1 cases in the NACT group (=predicted number of cases in the NACT group if there would have been no therapy). As a result of this we are able to calculate the 'predicted number of conversions' as the difference of

- Predicted number of cases using the probabilities of the no therapy group (*est*pN1), and the
- Observed number of cases in the therapy group (ypN1).

Further on, we define the 'predicted conversion rate' (PCR), $PCR \leq 1$, according to:

$$\begin{aligned} PCR &= \frac{\text{Predicted number of conversions}}{\text{Predicted number of cases accord. to the no therapy group}} \\ \left(\right. &= 1 - \frac{\text{Observed number of cases in the therapy group}}{\text{Predicted number of cases accord. to the no th. group}} \left. \right) \end{aligned}$$

where

- $PCR > 0$ implies that there is a positive number of conversions,
- $PCR = 0$ implies that there are no conversions and
- ($PCR < 0$ implies an opposite effect)

3.1 Bootstrap and CI

In a further step, we are going to investigate the statistical confidence of the PCR-value. For this purpose we are employing a bootstrap approach resampling the cases of our data set, see Davison and Hinkley (1997) or Efron and Tibshirani (1993). Looking at the empirical distribution of the PCR-value we are calculating the α-percentiles. In order to get an unbiased estimate of the percentiles we are going to employ the BC_α-method as described in Efron and Tibshirani (1993).

TABLE 1. Probabilities and predicted number of pN1 cases and observed number of ypN1 cases

	prob.	cases	pred. # of pN1 cases	obs. # of ypN1 cases
PR0	0.351	52	18.25	8
PR1	0.537	37	19.86	7
total		89	38.11	15

4 Results

Using the data as described above we calculate the probabilities in case of PR0 and PR1 applying the logistic regression model which results in calculating the predicted number of pN1 cases in the 3rd column – see Table 1. Further on, we are able to calculate the predicted conversion rate as (see 'total' values of the last row):

$$PCR = 1 - \frac{15}{38.11} = 0.6064$$

Thus, the result for the predicted conversion rate (PCR) can be described in our case as follows: If we apply the NACT, about 60 percent of the pN1 cases are forecasted to be converted to pN0 cases! Applying bootstrap techniques we calculate a 95 % CI (corrected) according to: (0.317,0.779). Noticing that zero (which means 'no conversion') is outside of the CI implies that the PCR-result turns out to be significantly different to zero.

5 Discussion

As we look at the therapy of HER2-positive BC cases over time a preliminary chemotherapy seems to be meaningful. In this paper we were investigating the conversion of pN stage (pN1→pN0) which is assumed to be a result of a preliminary chemotherapy. Using a data set of 153 NACT/without NACT cases we are estimating a 'predicted conversion rate' (PCR) of about 60 percent.
Finally, we conclude that the PCR is an useful novel quantity for our special case with possible applications for other similar approaches.

References

Boland, M. et al. (2017). Impact of receptor phenotype on nodal burden in patients with breast cancer who have undergone neoadjuvant chemotherapy. *BJS Open*, **1**, 39–45.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Gianni, L. et al. (2012). Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial. *The Lancet Oncology*, **13**, 25–32.

von Minckwitz, G. et al. (2012). Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer. *The New England Journal of Medicine*, **377**, 122–31.

Wuerstlein, R. and Harbeck N. (2017). Neoadjuvant Therapy for HER2-positive Breast Cancer. *Reviews on Recent Clinical Trials*, **12**, 81–92.

Analysis of maximum precipitation in Thailand using ensemble of non-stationary extreme value models

Thanawan Prahadchai[1]

[1] Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea

E-mail for correspondence: tanawanp.st@gmail.com

Abstract: Non-stationarity in heavy rainfall time series is often apparent in the form of trends because of long-term climate changes. We have built non-stationary (NS) models for annual maximum daily (AMP) data observed by 79 stations over Thailand. Totally 16 time-dependent functions of the location and scale parameters of the generalized extreme value (GEV) model are considered. At each station, a model is selected by using Akaike information criteria (AIC) among these candidates. The return levels corresponding to some years are calculated and predicted for the future. The ensemble (ESB) model shows less variance in return level estimation than the best model.

Keywords: Generalized extreme value distribution; Akaike weight; Delta method; Heavy rainfall.

1 Introduction

At present, designing future rainfall forecasts is essential due to heavy rainfall causes economic, social, and environmental damage. Due to climate variations, rainfall patterns have changed, causing flooding problems in every region and becoming a natural disaster problem that has been occurring in Thailand. Since the changing times, the behavior of rainwater varies due to factors such as impact of climate change from global warming. Therefore, the analysis of NS rainfall patterns in time series has been studied in several pieces of research (Lee and Ouarda, 2010; Cannon, 2010). Ignoring the uncertainty of a single hydrological model affects the reliability of future forecast values. The ESB model provides a consistent mechanism for model uncertainty. Multi-model ESB method has been proven not only

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

to improve upon the bias and variance compared to the single model but also to solve the overfitting problem. Davison (2003) showed that the mean square error of the weighted combination model was less than that of the single model. In this study, we present the application of two statistical models: the NS and ESB models, with AMP data for 79 stations in Thailand to forecast the future rainfall.

2 Methodology

2.1 Time-dependent GEV Distribution

In this research, we use the GEV distribution (GEVD) for analyzing AMP data. The cumulative distribution function of the GEVD is Coles (2001):

$$\text{Model } GEV00: \ F(x) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu(t)}{\sigma(t)}\right)\right]^{-1/\xi}\right\}, \qquad (1)$$

where $1 + \xi(x - \mu(t)/\sigma(t)) > 0$ and $\mu(t)$, $\sigma(t) > 0$ and ξ are the location, scale and shape parameters, respectively.

The NS models are presented in Table 1. It consists of 16 models totally; 8 of GEVD and 8 of GD. We denote M_{ij} to represent the functional form of time. t_0 denotes the year that the data records start. The maximum likelihood estimation method is used to estimate parameters (Coles, 2001). We select the suitable model using smallest the AIC and AIC_C (for $n < 40$).

TABLE 1. Functional forms of parameters for time dependent NS GEV models.

Models	$\mu(t),\ \sigma(t)$
M00:	$\mu(t) = \mu_0$
M10:	$\mu(t) = \mu_0 + \mu_1 \times (t - t_0 + 1)$
M20:	$\mu(t) = \mu_0 + \mu_1 \times (t - t_0 + 1) + \mu_2 \times (t - t_0 + 1)^2$
M30:	$\mu(t) = \mu_0 + \mu_1 \times \exp(-\mu_2 \times (t - t_0 + 1))$
M01:	$\mu(t) = \mu_0$
	$\sigma(t) = \exp(\sigma_0 + \sigma_1 \times (t - t_0 + 1))$
M11:	$\mu(t) = \mu_0 + \mu_1 \times (t - t_0 + 1)$
	$\sigma(t) = \exp(\sigma_0 + \sigma_1 \times (t - t_0 + 1))$
M21:	$\mu(t) = \mu_0 + \mu_1 \times (t - t_0 + 1) + \mu_2 \times (t - t_0 + 1)^2$
	$\sigma(t) = \exp(\sigma_0 + \sigma_1 \times (t - t_0 + 1))$
M31:	$\mu(t) = \mu_0 + \mu_1 \times \exp(-\mu_2 \times (t - t_0 + 1))$
	$\sigma(t) = \exp(\sigma_0 + \sigma_1 \times (t - t_0 + 1))$

After the best models are determined, next step is to derive the return level (z_q) which is the level exceeded on average only once in every T years as follows (Coles, 2001):

$$z_q(t) = \mu(t) - (\sigma(t)/\xi)\Big\{1 - \Big[-\log\Big(1 - 1/T\Big)\Big]^{-\xi}\Big\}, \quad \text{for } \xi \neq 0; \qquad (2)$$

where $T = 1/p$. The variances of return levels ($\hat{z}_T$) is approximated from delta method (Obeysekera and Salas, 2014).

2.2 Ensemble model

The ESB here is a weighted average of all NS models ($m = 1, \cdots, 16$) from Table 1. The return level, $z_q^*(t)$, define as:

$$\hat{z}_q^*(t) = \sum_{m=1}^{M} w_m \hat{z}_{q,m}(t) \tag{3}$$

The variance of return level, $Var(z_q^*(t))$, define as:

$$Var(\hat{z}_q^*(t)) = \left\{ \sum_{m=1}^{M} w_m \sqrt{Var(\hat{z}_{q,m}(t)|b_m) + b_m^2} \right\}^2 \tag{4}$$

This variance may be estimated by substituting $\hat{b}_m = \hat{z}_{q,m}(t) - \hat{z}_{q,m}^*(t)$ and $\hat{Var}(\hat{z}_{q,m}(t)|b_m)$. The estimates $\hat{z}_q(t)$ and $\hat{Var}(\hat{z}_{q,m}(t)|b_m)$ are found by inference methods from section 2.1, assuming that model m is the true model, and $\hat{z}_q^*(t)$ is given by equation 3 (Buckland et al., 1997). The wieghts in this study are calculated based on Akaike weights (Buckland et al., 1997, Davison, 2003);

$$w_m = \frac{\exp\Big(-AIC_m/2\Big)}{\sum_{m=1}^{M} \exp\Big(-AIC_m/2\Big)} \tag{5}$$

where $\sum w_m = 1$. We use AIC_C instead of AIC values to calculate weights for small observations.

3 Results and Conclusions

The results of study were found as follows. In 79 stations, 27 NS GEV models were selected based on AIC value for AMP data. In the southeast region such as Nakon Si Tammarat and in the eastern region such as Trad are increasing with very heavy future rainfall (see Figure 1). We found that the ESB model shows lower variance in return level estimation than the best model. (see Figure 2).

Acknowledgments: This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research, NO.5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF). Special thanks to Jeong-Soo Park professor for his advice to improve this work.

References

Buckland, S.T., Burnham, K.P., and Augustin, N.H. (1997). Model selection: an integral part of inference. *Biometrics*, 603 – 618.

Cannon A.J. (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydro Process*, **24**, 673 – 685.

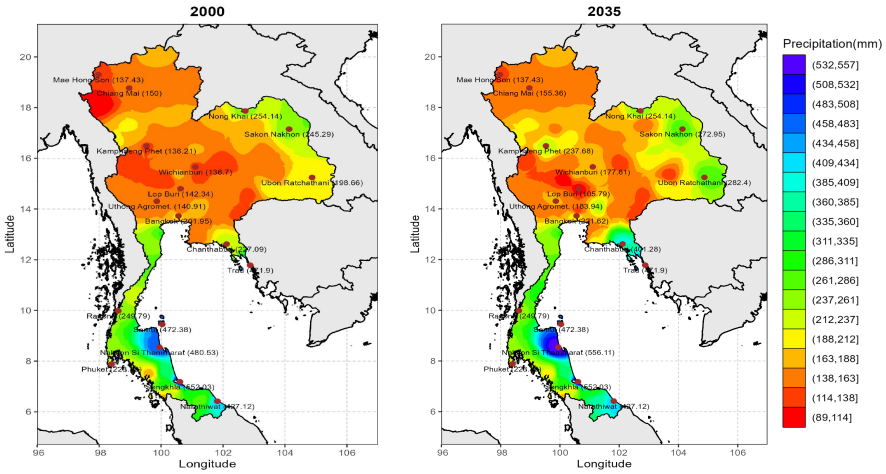


FIGURE 1. Maps of 50-year return levels of 2000 (past) and 2035 (future) (unit: mm) estimated for the AMP data in Thailand.

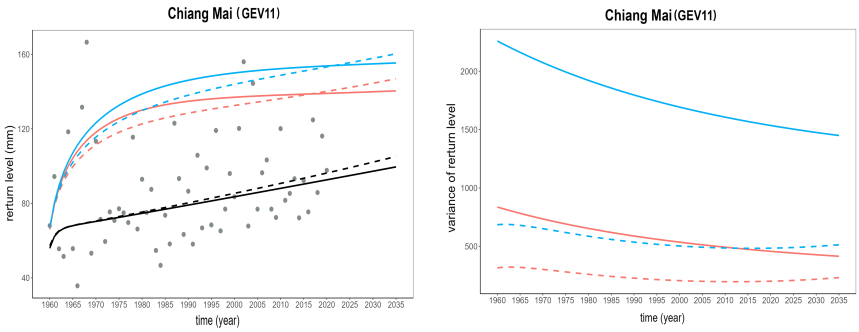


FIGURE 2. Comparison of 2-, 20-, and 50-year return levels (left) and variance of 20- and 50-year return level (right) between NS (dashed line) and ESB (solid line) models of Chiang Mai station. Black, red, and blue represent 2-, 20-, and 50-year, respectively.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. London: Springer.

Davison A.C. (2003). *Statistical models*. Cambridge: University Press.

Lee, T., and Ouarda, T.B.M.J. (2010). Long-term prediction of precipitation and hydrologic extremes with nonstationary oscillation processes. *Jour Geophys Res.*, **115**, D13107.

Obeysekera, J., and Salas, J.D. (2014). Quantifying the uncertainty of design floods under nonstationary conditions. *Journal of Hydrologic Engineering*, **19(7)**, 1438 – 1446.

Projecting Extreme Precipitation in the Philippines using the CMIP6 Multi-model Ensemble

Thanawan Prahadchai[1], Jeong-Soo Park[1]

[1] Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea

E-mail for correspondence: jspark@jnu.ac.kr

Abstract: To project the future changes of extreme precipitation in the Philippines, we investigated the observations based on 53 stations and 24 CMIP6 (Coupled Model Inter-Comparison Project Phase 6) models. We applied generalized extreme value (GEV) distribution and multivariate bias-correction to series of annual maximum daily precipitation (AMP1) data acquired from the observations and models under three shared socioeconomic pathway (SSP) scenarios (SSP2-4.5, SSP3-7.0, and SSP5-8.5). We employed an ensemble method that takes both independence and performance of model into account, which is named as the PI-weighting. From this study, we predict that the relative increases of 20-year return value of the AMP1 from the past to the year 2100 be about 8.5% in the SSP2-4.5, 11.6% in the SSP3-7.0, and 17% in the SSP5-8.5 scenarios, respectively, in the spatial median over the Philippines.

Keywords: Climate change; L-moments estimation; Relative change; Shape parameters.

1 Introduction

The Philippines is at high risk by the impacts of climate change, including increased frequency of extreme weather events, rising temperature, sea level rise, and extreme rainfall. This is because of its high exposure to natural risks (tropical cyclones, floods, landslides, droughts), reliance on climate-sensitive natural resources, and huge coastlines where most major cities and majority of the population resides. Heavy rainfall in the Philippines are usually due to monsoon surge (intensification of the monsoons) and slow-moving tropical cyclon in the area. The above typhoons brought extreme

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

rainfall and caused floods and landslides. Downpour events (with 350 mm or higher) have been more recurrent in the latter part of the 20th century (Villafuerte et al., 2015). In this study, we update the previous studies based on 24 CMIP6 models under the three shared socioeconomic pathway (SSP) scenarios: namely SSP2-4.5, SSP3-7.0, and SSP5-8.5. We predict the amount of changes in the largest precipitation

2 Data and method

We used the 24 CMIP6 climate models in this study. The considered scenarios are shared socioeconomic pathways SSP2-4.5, SSP3-7.0, and SSP5-8.5. Two periods are considered for the future, namely, period 1 (2021–2060) and period 2 (2061–2100). The observations for 40-year (1975–2014) reference period were obtained from the Philippines Atmospheric, Geophysical and Astronomical Services Administration (PAGASA) (PAGASA, 2011).

When our interest is in analyzing extreme events, the generalized extreme value (GEV) distribution is typically employed. The changes in extremes is usually described in terms of the changes in extreme quantile, which is called as the return level associated with the return period $1/p$ (Coles, 2001). The parameters in GEV distribution are estimated by the L-moment method (Hosking and Wallis, 1997).

In this study, we choose the multivariate bias correction (MBC) method by Cannon (Cannon, 2018) among some available BC techniques. The MBC is a multivariate extension of quantile delta mapping (QDM). QDM has an advantage of preserving the approximate trends of the model data.

A weighting method that accounts for both the independence and performance simultaneously is called the PI-weighting (Knutti et al., 2017). Weights are computed for each model based on a combination of the distance D_i (apprising the performance) and the model similarity S_{ij} (apprising the dependence):

$$w_i = exp(-\frac{D_i}{\sigma_D})/(1 + \sum_{j \neq i}^{M} exp(-\frac{S_{ij}}{\sigma_S})), \tag{1}$$

with the total number of models M and the shape parameters σ_D and σ_S. We follow a relatively simple method proposed by Shin et al. (2021) to determine the shape parameters σ_D and σ_S. They used Shannon's entropy and the Chi-square statistics.

3 Results

The normalized PI-weights are obtained using Equation (1) with $\sigma_S = 0.4$ and $\sigma_D = 0.86$. Figure 1 displays boxplots of the 20-year (50-year) return

values of the AMP1 in Philippines. The increasing trends from the past to the future are evident in every scenario.

Figure 2 displays boxplots for the 20-year and 50-year return periods, as compared to the reference years (1975–2014) for the two future periods under the three scenarios. We find out that a 1-in-20 year (1-in-50 year) AMP1 in Philippines will likely become 1-in-17 (1-in-37) year, 1-in-17 (1-in-33) year, and 1-in-14 (1-in-31) year events in the median by 2100 based on the SSP2-4.5, SSP3-7.0, and SSP5-8.5 scenarios, respectively.

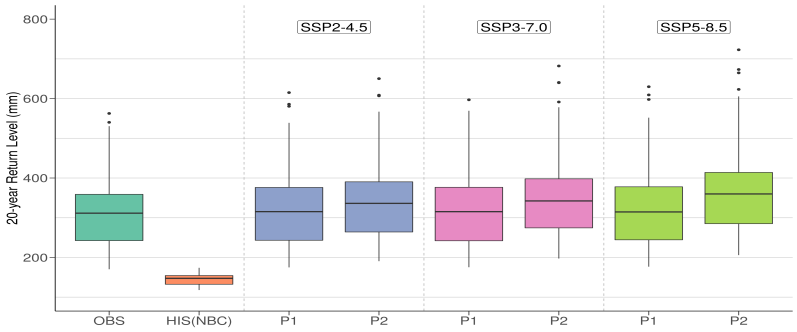


FIGURE 1. Boxplots of 20-year return values (in mm) of the AMP1 in Philippines for the future time: P1 (2021-2060), P2 (2061-2100), under three scenarios. HIST(NBC) and OBS indicate the historical data without a bias correction and the observations.

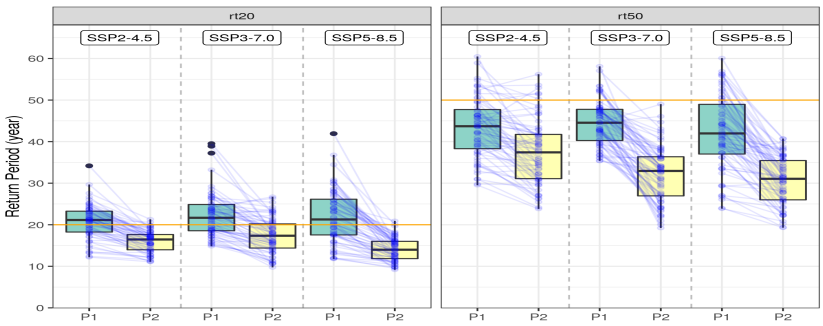


FIGURE 2. Parallel coordinated boxplots over the Philippines, similar to those in Figure 1, but for 20-year and 50-year return periods compared to the past from 1975 to 2014.

4 Summary

We estimated the future changes in precipitation extremes within Philippines using observations, 24 multiple CMIP6 models, generalized extreme

value distribution, the multivariate bias correction technique, and the model weighting method (PI-weighting), which account for both the performance and independence of the models.

From 20-year and 50-year return values of the annual maximum daily precipitation (AMP1) averaged over 64 grids in Philippines for two future time under three SSP scenarios, the increasing trends from P1 (2021-2060) to P2 (2061-2100) are evident in SSP3-7.0 and SSP5-8.5 scenarios. We foretell that the relative rises of 20-year return value of the AMP1 from the past to the year 2100 will be about 8.5% in the SSP2-4.5, 11.6% in the SSP3-7.0, and 17% in the SSP5-8.5 scenarios, respectively, in the spatial median.

We also found out that a 1-in-20 year (1-in-50 year) AMP1 within Philippines will likely become a 1-in-16 (1-in-37) year, a 1-in-17 (1-in-32) year, and a 1-in-14 (1-in-31) year event in terms of the median by the year 2100 under the SSP2-4.5, SSP3-7.0, and SSP5-8.5 scenarios, respectively, as compared to the observed data from 1975 through 2014.

Acknowledgments: This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research, NO.5120200913674) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

References

Villafuerte II, M. Q., Matsumoto, J., and Kubota, H. (2015). Changes in extreme rainfall in the Philippines (1911–2010) linked to global mean temperature and ENSO. *Intern J. Clim*, **35(8)**, 2033 – 2044.

PAGASA (2011). Climate Change in the Philippines. https://www.pagasa.dost.gov.ph/information/climate-change-in-the-philippines.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. London: Springer.

Hosking, J.R.M. and Wallis, J.R. (1997). *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge Univ Press: Cambridge, UK, 244.

Cannon, A.J. (2018). Multivariate quantile mapping bias correction: An N-dimensional probability density function transform for climate model simulations of multiple variables. *Clim. Dyn.*, **50**, 31 – 49.

Knutti, R., Sedlacek, J., Sanderson. B.M., Lorenz, R., et al. (2017). A climate model projection weighting scheme accounting for performance and independence. *eophys Res Lett*, **44**, 1909 – 1918.

Shin, Y., Shin, Y., Hong, J., et al. (2021). Future Projections and Uncertainty Assessment of Precipitation Extremes in the Korean Peninsula from the CMIP6 Ensemble with a Statistical Framework. *Atmosphere*, *12(1)*, 97.

Extending Generalized Additive Models for extreme value modeling: a software review

Ilaria Prosdocimi[1]

[1] Universitá Ca' Foscari Venezia, Italy

E-mail for correspondence: ilaria.prosdocimi@unive.it

Abstract: The terms distributional regression and Generalized Additive Model for Location, Scale and Shape both indicate a class of broad statistical models which allow for the modeling of a given response variable as a flexible function of some predictors. These models extend traditional regression models by allowing the response variable to follow any distribution indexed by one or more parameters. These parameters are then allowed to vary as unknown functions of the predictors: these functions are typically estimated using basis expansions. The analysis of extremes has proven to be an interesting area of application for distributional regression: there is for example an interest in assessing whether climate and other anthropogenic changes are having an impact on the measured records of some environmental extremes. These type of data is typically assumed to follow some highly skewed distribution which does not belong to the exponential family. Furthermore, there is little prior knowledge on the type of shape that the impacts of climate change could have on extremes: it is therefore preferable to allow for the relationship to be derived from observations. In this work, I provide a brief overview of distributional regression, extreme value statistic and of some off-the-shelf implementations available in the `R` statistical software for the estimation of distributional regression models for extremes.

Keywords: GAMLSS; Distributional Regression; Extremes; Statistical Software.

1 Generalized Additive Models and their extensions

Generalized Additive Models (GAMs, Wood, 2017) are a powerful statistical modeling tool which extends Generalized Linear Models by relaxing the assumption that the relationship between the (link-transformed) expected value of the response variable and the predictors can be expressed as a parametric function. Like in classical regression models, the focus of GAMs is to model the expected value of a variable of interest, typically assuming

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

that the response variable follows a certain distribution parametrized by a set a parameters. Rigby and Stasinopoulos (2005) proposed an extension to the traditional GAM framework, by allowing any of the parameters of the distribution (or their transformation by means of a link function) to vary as a flexible function of the predictors. These type of extensions can for example be used to model heteroschedasticity in linear regression or varying overdispersion for count data modeled using a Negative Binomial. This is achieved by allowing the parameters other than the location to vary as function of the predictors. In particular the parameters which are related to the scale and shape of the distributions might be modeled as functions of the covariates: this gives origin to the name Generalized Additive Models for Location, Scale and Shape (GAMLSS). The overall GAMLSS specification is therefore given as follow. Assume we have a sample of n independent observations from p explanatory variables $(X_1, \ldots, X_p)$ and one variable of interest Y: for the i^{th} element of the sample we have a vector of observations $(x_{1i}, \ldots, x_{pi}, y_i)$. Conditional on the covariates vector $\mathbf{x_i} = (x_{1i}, \ldots, x_{pi})$, the observations y_i are assumed to be independent realizations of a random variable Y_i which follows a distribution $\mathcal{D}$ indexed by K parameters: $(Y_i|\mathbf{X_i} = \mathbf{x_i}) \sim \mathcal{D}(\theta_{1i}, \ldots, \theta_{Ki})$. Each distribution parameter is assumed to be related to a linear predictor $\eta_i^{\theta_k} = g(\theta_{Ki})$ which is taken to be of the form:

$$\eta_i^{\theta_k} = g(\theta_{ki}) = \beta_0^{\theta_k} + f_1^{\theta_k}(x_{1i}) + \ldots + f_p^{\theta_k}(x_{pi}) \quad (1)$$

where the $f_j^{\theta_k}(x_{ji})$ components might represent parametric or unknown functions of the covariates. In the latter case, as in GAMs, a common approach to estimate such unknown functions is to approximate them by a basis function approximation: $f_J^{\theta_K}(x_{Ji}) = \sum_{b=1}^{B} \beta_{Jb}^{\theta_K} B_{Jb}^{\theta_K}(x_{Ji})$, where $B_{Jb}^{\theta_K}(x_{Ji})$ is a basis function, $\beta_{Jb}^{\theta_K}$ is the corresponding coefficient which needs to be estimated and B is the basis size. To avoid the need to choose the basis size, a common strategy is to use very rich bases and to use a penalized likelihood approach to avoid overfitting in the model estimation. The penalization can be derived as a constraints on the differences between consecutive $\beta_{Jb}^{\theta_K}$ or as a gaussian prior on the vector of the $(\beta_{J1}^{\theta_k}, \ldots, \beta_{JB}^{\theta_k})$ parameters. As a consequence, one can derive an estimate for the functions in equation (1) either by maximizing the penalized likelihood or by employing a Bayesian approach (see, among others, Umlauf et al. 2018). Although only univariate simple $f_j(\cdot)$ terms are included in equation (1) more complex predictor components can be included in the linear predictors, including multivariate smooth terms (eg $f_{ij}(x_i, x_j)$), (spatial) random effects or varying coefficient terms. Since it is sometimes the case that a distribution's parameter does not have a direct relationship with the location, scale or shape of the distribution GAMLSS-type models are also referred to as distributional regression models, to emphasize that what is modeled is the response variable's distribution by allowing its parameters to vary. A further extension of the GAMLSS framework is provided by

Vector Generalized Additive Models (VGAM, Yee 2015): which also allow for a multivariate response variable.

2 Statistical models for extremes

Extreme Value Statistic (Coles, 2001) aims at quantifying the atypical behavior of a process rather than the typical, average one which is the focus of traditional statistical inference. As such, extreme value inference is based on samples which can be deemed to be in some way representative of the atypical, extreme behavior of a process of interest. Two main types of sample are employed for inference on extreme values: block maxima and peaks over threshold. Block maxima are derived as the largest values in a fixed block of recording time, for example a year: yearly maxima of peak flow at a river gauge station are a classical example of such type of records. Block maxima are conceptually easy, but their use involve a great loss of information: an alternative definition of extremes is provided by peaks over threshold in which all values larger than a certain high threshold u are considered to be extreme. This implies, for example, that there could be several large events in the same year, while for some years no extreme events might be recorded, thus ensuring that all large events contribute to the the characterization of the extremes of a process. The challenge in the definition of the peaks over threshold records is to specify the threshold u above which records can be deemed to be extremes, although several methods have been proposed (Scarrot and MacDonald, 2012). One notable approach to select u is to specify its value as the value of a certain upper quantile of the entire sample: this can be generalized to a varying threshold approach by means of quantile regression, which can also be described as a distributional regression model in which it is assumed that the response variable follows an asymmetric Laplace distribution (ALD, see Youngman, 2020).
Using asymptotic arguments, different distributions are derived as the limiting distribution for either block maxima or peaks over threshold (Coles, 2001): these are respectively the Generalized Extreme Values (GEV) and the Generalized Pareto distribution (GP). The probability distribution functions (pdf) of the GEV is given as:

$$f(y;\mu,\sigma,\xi) = \frac{1}{\sigma}\left\{\left(1+\xi\frac{y-\mu}{\sigma}\right)^{\frac{\xi-1}{\xi}}\right\}\exp\left\{-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\}$$

for $y : 1+\xi(y-\mu)/\sigma > 0$. For the peaks over threshold, one typically models the value of the threshold exceedance conditional on the threshold being exceeded: $Z = (Y-u)|Y>u$, so that $Z>0$. Z is then assumed to follow a GP distribution with the following pdf:

$$f(z;\sigma,\xi) = \frac{1}{\sigma}\exp\left\{\left(1+\xi\frac{z}{\sigma}\right)^{-\frac{\xi+1}{\xi}}\right\}$$

and $z < -\sigma/\xi$, if $\xi < 0$. Notice that the domain over which the GEV and the GP distribution are defined depends on their parameters: in particular the sign of the shape parameter ξ determines whether the distribution has a lower or upper bound. There is an asymptotic correspondence between the two definitions of extremes: the Point Process (PP) representation of peaks over threshold provides a clear link between the GEV and the GP (see Coles 2001). Furthermore, when the shape parameter is null, both distributions reduce to simpler two-parameter distributions, respectively the Gumbel and Exponential distribution Although the GEV and GP distribution are derived as the limiting distribution of extremes, other distributions are employed for the analysis of extremes in several applications, either because they historically have been found to provide a good fit to the data or because they provide a simpler analytical framework. Popular alternative distributions are for example the Gamma, log-normal, Generalized Logistic, log-Pearson type III or Kappa distribution (Hosking and Wallis, 1997).

3 Flexible distributional models for extremes: statistical software implementation

It can be of interest to asses if some external variable influence the extremes of a process. A common approach to asses this type of question is to allow one or more of the parameters of the extreme value distribution to change as a function of some covariate, thus developing a distributional regression approach for extremes. In particular (see for example Villarini et al. 2010), it can be of interest to use flexible functions of the predictors like in equation (1). A working and reliable software which allows for the estimation of distributional regression is essential for the widespread usage of flexible modeling techniques in the applied sciences. Nevertheless, since extreme value distributions do not belong to the exponential family of distribution and their estimation can be challenging due to the fact that the distributions' domain depend on the distributions' parameters, many general purpose packages for GAMLSS-type estimation did not directly allow for the estimation of extreme value families. This has changed in the most recent years, with novel state-of-the-art routines for the estimation of distributional regression for extremes being made available as packages for the `R` Statistical computing language. Several of these packages also have ways for the user to add new distributions for which distributional regression models can be specified, thus increasing the possibility of applied scientists to use distributions which are relevant for their specific use case. A list of packages which have built-in functions for the estimation of distributional regression models for extremes is provided below, together with some indication of the estimation approaches, the distributions implemented in the package which are relevant for extreme value analysis and some further comments on whether functions specifically relevant for the analysis of extremes are

present. All packages are available on CRAN (https://cran.r-project.org/), the centralized repository for `R` software extensions.

- The `mgcv` (Wood, 2017) package implements (among others) the GEV, Gumbel and Gamma distributions. It is a CRAN recommended package. Adding novel distributions is possible but not straightforward. Various estimation approaches are implemented, although the default approach is mostly frequentist.

- The `gamlss` (Rigby and Stasinopoulos, 2005) package implements (among others) the Reverse GEV, Reverse Gumbel, Gamma and log-normal distributions: notice that the parametrization of extreme value distributions is somewhat different from the one in Coles (2001) typically used in the extreme value literature. Adding novel distributions is possible and not too complicated. The estimation approach is mostly frequentist.

- The `VGAM` (Yee, 2015) package implements (among others) the GEV, Frechet, GP, Gumbel and Gamma distribution. Adding novel distributions is possible but not straightforward. The estimation approach is mostly frequentist. Some extreme value dedicated functions.

- The `evgam` (Youngman, 2020) package implements (among others) the ALD, GEV, GP and a PP representation of peaks over threshold. Adding novel distributions is possible but not straightforward. The estimation approach is mostly frequentist. Several extreme value dedicated functions.

- The `bamlss` (Umlauf et al., 2018) package implements all families used in the `gamlss` package and, among others, the ALD, GEV, GP and Gumbel distribution. The main estimation approach is Bayesian, since the package aims to implement Bayesian approaches for the estimation of GAMLSS-type models. Adding novel distributions is possible and not too complicated.

- The `brms` package implements (among others) the ALD, GEV, Frechet, Gamma and log-normal distribution. `brms` is a general purpose package for the Bayesian estimation of regression models, which also allows for the estimation of flexible regression models. Adding novel distributions is possible and not too complicated (with abundant documentation in the package's vignette).

References

Coles, S.G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

Hosking, J.R.M. and Wallis, J.R. (1997). *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press.

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

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical journal*, **10**, 33 – 60

Umlauf, N., Klein, N., Zeileis, A. (2018). BAMLSS: Bayesian Additive Models for Location, Scale and Shape (and Beyond). *Journal of Computational and Graphical Statistics*, **27**, 612 – 627

Villarini, G., Smith, J.A. and Napolitano, F. (2010). Nonstationary modeling of a long record of rainfall and temperature over Rome. *Advances in Water Resources*, **33**, 1256 – 1267

Wood, S.N. (2017). *Generalized additive models: an introduction with R (2nd ed.)*. CRC press.

Yee, T.W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer.

Youngman, B.D. (2020). *Evgam: An R package for generalized additive extreme value models*. arXiv preprint 2003.04067.

Review of covariance structures of multiresponse and multisubject models from the point of view of Optimal Design of Experiments

Juan M. Rodríguez-Díaz[1]

[1] Universidad de Salamanca, Spain

E-mail for correspondence: juanmrod@usal.es

Abstract: In many fields of knowledge experiments are performed in order to get an insight of the models describing the objects of study. The choice of the optimal experimental conditions where runs are to be taken is essential in order to get the maximum information about these models. The literature offers a great number of works based on independent observations, and thus very often the researcher agrees with this assumption in order to be able to use the huge toolbox of results, algorithms and procedures already known for independent observations. However, different situations can make the researchers drop this independence assumption. Some examples are: repeated observations over the same subject at different temporal points, observing different responses on the same subject (multiresponse models) and/or observing several subjects that could share or not similar covariance structures. Quite often, the actual situation is a combination of the previous ones. In all of these cases it would be interesting to know the design that produces the best estimates of the model parameters, or minimizes the variance of the predicted response, or optimizes another characteristic of the model (according with a specific optimality criteria), that is the optimal designs for each model. In this work, most of the situations described above will be addressed, trying to find analytically the optimal sample plans for them using optimal experimental design techniques, and applying the results to convenient examples.

Keywords: Covariance matrix; Multiresponse Models; Multisubject Models; Optimal Design of Experiments.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

1 Optimal Design of Experiments

The one-response linear model will be denoted as $y = \mathbf{f}(x)^T\boldsymbol{\beta} + u$, where $\boldsymbol{\beta}$ is the parameter vector of size m, u is the error term, and $\mathbf{f}(x) = (f_1(x), \ldots, f_m(x))^T$, with the $f_i(x)$ linearly independent in the experimental domain $\mathcal{X}$. An exact design ξ is a collection of points $\{x_1, \ldots, x_n\}$ of the independent variable, which represents the experimental conditions, with x_i in $\mathcal{X}$. In matrix notation it can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + u \quad ,$$

where $\mathbf{Y} = \{y_1, \ldots, y_n\}^T$ is the observations vector, $\mathbf{U} = \{u_1, \ldots, u_n\}^T$ the error terms, and $\mathbf{X} = (\mathbf{f}(x_1), \ldots, \mathbf{f}(x_n))^T$ the design matrix. For normally distributed random errors $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, with $\boldsymbol{\Sigma} = Var(\mathbf{Y})$ and $Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}$. The *information matrix* of ξ will be

$$\mathbf{M}(\xi) = \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} \quad .$$

The standard method to obtain optimal designs requires to know the analytical expression of the model and to compute the derivatives with respect to the parameters in order to work with the linearized model. When no analytical expression of the model can be obtained, some methods for computing these derivatives can be employed, as described in Rodríguez-Díaz and Sánchez-León (2014).

When dealing with correlated observations the size of the design, n, should be fixed in advance. In many experiments it has no sense to take more than one observation to the same experimental unit at the same design point x, especially when the design variable is time, thus in the following it will be assumed that $x_i \neq x_j$ for all i, j. Usually, the aim is to find the points $\{x_1, x_2, ...\}$ where to take observations in order to get the best estimates of the parameters of the model, that is, the estimation with minimum variance, providing an *optimal design* for the model. The inverse of the information matrix is proportional to the covariance matrix (the generalized variance) of the parameter estimators) of the model; therefore the aim is usually to minimize (a convex function of) $\mathbf{M}^{-1}(\xi)$. However, there is not an only way of minimizing a matrix, giving rise to different *criterion functions*. A particular criterion function should be chosen depending on the objectives of the practitioners, for instance getting the best estimators of the parameters (one, some of them or all of them), or minimizing the variance of the predicted response. The most used criterion is D-optimality, which focuses on the determinant of the information matrix. A design ξ is D-optimal if maximizes this determinant, what is equivalent to minimize that of the covariance matrix. A-optimality pays attention to the trace of the covariance matrix, thus an A-optimal design minimizes the average of the variances of the estimators of model parameters. When the information matrix depends on unknown parameters, nominal values are needed

for them and thus the obtained designs will be *locally optimal*, that is, they are good for (or close) those nominal values used in the computation. Fedorov and Hackel (1992), Pukelsheim (1986) or Atkinson et al. (2007) are classic references on optimal design of experiments.

2 Multiresponse and multisubject models

In many studies, different kind of responses (say k of them) are measured, getting into the field of *multiresponse models*. These models have been studied from the point of view of optimality from different perspectives, but to date most of the literature about multiresponse models consider correlation among different variables observed on the same 'subject', whereas in every of the studies the measures taken at different points, $y(x)$ and $y(x')$, are assumed independent. This assumption could be sensible when the variable refers to individuals in a study, but not so much in other cases. In particular, when the interest is to analyze the evolution of a set of characteristics (variables) observed in a specific experimental unit at a different time moments. It seems clear that, apart from a *static* (in the same temporal point) or *intra* covariance structure among different type of observations taken at the same time, a *longitudinal* (along a period of time) or *inter* correlation among the same type of measures obtained at different times should be taken into account (Rodríguez-Díaz and Sánchez-León, 2019a, 2019b).
Let $S(x)$ denote the variance of the sample $y(x)$, that is $S(x) = cov[y_1(x), \ldots, y_k(x)] = (\sigma_{ij})_{i,j=1,\ldots,k}$. It is usual to assume that the relation among the different variables is similar for every x, thus in the following a constant covariance S will be considered (intra correlation). For other hand, the covariance between the same type of observations taken at different points will be assumed to be dependent only on the distance between points, thus the longitudinal covariance will be the same for the k different responses, $R = cov[y_i(x_1), \ldots, y_i(x_n)]^T$ $\forall i$. The points for measurements are supposed to be different, that is, a minimum distance $\delta > 0$ between samples is assumed, avoiding singular covariance matrices (longitudinal covariance).
To date, the double covariance structure has been considered only for studies carried out over one experimental unit, for which several variables were measured at different times (Rodríguez-Díaz and Sánchez-León, 2019a, 2019b). Now N subjects are supposed to be observed at different temporal points $t_1, \ldots, t_n$, which will be the design ξ. The design points t_i can denote any convenient temporal unit. It will be assumed that for each t_i in ξ the values of several characteristics $Y_1, \ldots, Y_k$ will be obtained for all of the subjects, and the aim will be to choose the 'best' design, the one giving the greatest information about the models describing the evolution of the response variables.
Different situations can be considered: the first one when the subjects have similar characteristics (e.g. students within the same classroom) and thus

the same covariance structure can be assumed for every subject; and another one when the previous assumption is no longer valid (e.g. different type of bacteria) with different covariance structures and thus more complex to deal with. In addition, the response variables may have the same model for every subject (or may not), and various evolution model can be considered for the response variables. Furthermore, when there are different types of subjects multiple subjects of each type could be observed, in this case usually assuming independence between subjects. Combining all of these factors, several analytical results have been obtained, and various examples of application are shown.

References

Atkinson, A.C., Donev A.N. and Tobias R.D. (2007). *Optimum experimental designs, with SAS*. Oxford University Press.

Fedorov, V.V. and Hackl P. (1997). *Model-oriented design of experiments*. New York: Springer-Verlag.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. Philadelphia, PA: SIAM.

Rodríguez-Díaz, J.M, and Sánchez-León G. (2014). Design optimality for models defined by a system of ordinary differential equations. *Biometrical Journal* 56 (5): 886–900.

Rodríguez-Díaz, J.M. and Sánchez-León G. (2019a). Efficient parameter estimation in mulltiresponse models measuring radioactivity retention. *Radiation and Environmental Biophysics* 58 (2): 167–182.

Rodríguez-Díaz, J.M. and Sánchez-León G. (2019b). Optimal designs for multiresponse models with double covariance structure. *Chemo. Intel. Lab. Syst.* 189: 1–7.

Rodríguez-Díaz, J.M. et al. (2020). Optimal designs for a linear-model compositional response. *Stochastic Environmental Research and Risk Assessment* 34, 139-148.

Rodríguez-Díaz, J.M., Pruneda, R.E. and Rodríguez-Hernández, M. (2022). Design Plan for an Evolution Study of Related Characteristics of a Population. *Mathematics 2022*, 10, 792.

Robust zero-inflated interval regression for cyber security survey data

Cristian Roner[1], Claudia Di Caterina[1], Davide Ferrari[1]

[1] Faculty of Economics and Management, University of Bozen-Bolzano, Italy

E-mail for correspondence: claudia.dicaterina@univr.it

Abstract: Zero-inflated interval regression models handle the excess of zeros in an ordinal response by combining a probit model with an ordered probit model. In case of violation of the usual distributional assumptions, standard maximum likelihood estimation is biased and inefficient. We propose a robust inferential approach based on exponential tilting, which weighs each observation according to its compatibility with the assumed model. This methodology is motivated by the analysis of UK survey data on cyber attacks. Our robust results clearly outperform classical inference and reveal the importance of cyber defence investments in reducing the costs from cyber security breaches.

Keywords: Exponential tilting; Likelihood inference; Model misspecification; Ordinal response variable.

1 Introduction

Cyber attacks are malignant assaults launched against single computers or a computer network in order to access an asset, like data, patents, or product specifics. Due to their increasing occurrence, the study of the economic impact of cyber security breaches has gained relevance in recent years. Here we focus on data from two waves of the UK Cyber Security Breaches Survey (CSBS). The response variable of interest is the cost associated with recent cyber breaches, and takes interval values. The first interval class containing the zero is observed with especially high frequency. Such features suggest to analyse the data via the zero-inflated interval regression (ZIIR) model (Brown et al., 2015). As in this case some key distributional assumptions are violated, the maximum likelihood (ML) estimator is known to be biased, so we develop a robust approach within the exponential tilting framework (Choi et al., 2000). This strategy allows to get reliable insights on the impact of investments in cyber defence on costs from cyber attacks.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Zero-inflated interval regression

Let Y^* be the unobservable actual monetary loss sustained by a firm due to cyber attacks. For each ith firm, we observe only the K ordered categories

$$Y_i = l \in \{0, 1, \ldots, K\} \quad \text{if} \quad \gamma_l < Y_i^* \leq \gamma_{l+1} \qquad (i = 1, \ldots, n)\,,$$

where $\gamma = (-\infty, \gamma_1, \ldots, \gamma_K, +\infty)^{\mathrm{T}}$ is a $(K+2)$-vector of given thresholds. The ZIIR assumes for Y^* a two-stage selection model. In the first stage, the binary variable S indicates whether any loss occurs; for the ith firm, $Y_i^* = 0$ if $S_i = 0$, and $Y_i^* > 0$ if $S_i = 1$. We consider the probit model

$$P(S_i = 1; \beta^{(1)}) = \Phi(x_i^{\mathrm{T}} \beta^{(1)}) \qquad (i = 1, \ldots, n)\,, \tag{1}$$

where $\beta^{(1)}$ is a p-vector of parameters, x_i is a p-vector of covariates, and $\Phi(\cdot)$ denotes the cumulative standard normal distribution function. The second stage predicts instead the entity of the loss, once this loss has occurred. To this aim, we use the latent regression model

$$Y_i^* = z_i^{\mathrm{T}} \beta^{(2)} + U_i^{(2)} \qquad (i = 1, \ldots, n)\,, \tag{2}$$

where $\beta^{(2)}$ is a q-vector of parameters, z_i is a q-vector of covariates, and $U_i^{(2)}$ are independent normal errors with zero mean and variance σ^2. Let $\theta = (\beta^{(1)}, \beta^{(2)}, \sigma^2)$ denote the overall parameter vector. Based on (1) and (2), the zero-inflated distribution of the observed Y_i $(i = 1, \ldots, n)$ is

$$\begin{aligned} P(Y_i = 0; \theta) &= \Phi(-x_i^{\mathrm{T}} \beta^{(1)}) + \Phi(x_i^{\mathrm{T}} \beta^{(1)}) P(Y_i = 0 | S_i = 1; \beta^{(2)})\,, \\ P(Y_i = l; \theta) &= \Phi(x_i^{\mathrm{T}} \beta^{(1)}) P(Y_i = l | S_i = 1; \beta^{(2)}), \quad l \in \{1, \ldots, K\}\,, \end{aligned} \tag{3}$$

with $P(Y_i = l | S_i = 1; \beta^{(2)}) = \Phi(\gamma_{l+1} - z_i^{\mathrm{T}} \beta^{(2)}) - \Phi(\gamma_l - z_i^{\mathrm{T}} \beta^{(2)})$ $(l \in \{0, \ldots, K\})$.

3 Robust inference by exponential tilting

Crucial assumptions of the ZIIR model are normality of the actual monetary loss and independence among units. Yet here Y^* has a non-negative support by definition and two observations in CSBS data $y_1, \ldots, y_n$ can refer to the same company which participated, anonymously, in both survey waves. Such issues clearly invalidate ML results and call for a strategy that is robust against model misspecification. We suggest to control the individual contribution to the global likelihood according to the compatibility of the specific unit with the assumed model. Following Genton and Hall (2015), for a fixed level of exponential tilting $\alpha \geq 0$, the tilted estimator $\hat{\theta}^{(\alpha)}$ is found by maximizing numerically the tilted log-likelihood function

$$\ell_{\hat{\pi}^{(\alpha)}}(\theta) = \sum_{i=1}^{n} \hat{\pi}_i^{(\alpha)}(\theta) \log P(Y_i = y_i; \theta)\,,$$

where $P(Y_i = y_i; \theta)$ is given in (3) and the weights $\hat{\pi}_i^{(a)}(\theta)$ are obtained by tilting uniform prior weights $\pi_i^u = 1/n$, for each $i = 1, \ldots, n$, in the direction that gives most emphasis to data enjoying larger likelihood. Setting $\alpha = 0$ corresponds to ML estimation, as $\hat{\pi}_i^{(0)}(\theta) = \pi_i^u = 1/n$ for all i.
Under suitable regularity conditions, the estimator $\hat{\theta}^{(\alpha)}$ is asymptotically normal (Choi et al., 2000). Its covariance matrix can be estimated by

$$\hat{V}_{boot}^{(\alpha)} = \frac{n}{R-1} \sum_{r=1}^{R} \big(\hat{\theta}_r^{(\alpha)} - \bar{\theta}^{(\alpha)}\big)\big(\hat{\theta}_r^{(\alpha)} - \bar{\theta}^{(\alpha)}\big)^{\mathrm{T}}, \tag{4}$$

where $\hat{\theta}_1^{(\alpha)}, \ldots, \hat{\theta}_R^{(\alpha)}$ are estimates of θ computed via standard nonparametric bootstrap techniques for some α, and $\bar{\theta}^{(\alpha)} = \sum_{r=1}^{R} \hat{\theta}_r^{(\alpha)}/R$.
The most appropriate value of α is chosen to minimize the bootstrap mean squared error (MSE) of the tilted estimator, i.e.

$$\widehat{MSE}_{boot}(\alpha) = \Big\|\hat{\theta}^{(\alpha)} - \bar{\theta}^{(\alpha)}\Big\|_2^2 + \frac{1}{n}\mathrm{tr}\big(\hat{V}_{boot}^{(\alpha)}\big), \tag{5}$$

where $\mathrm{tr}\big(V_{boot}^{(\alpha)}\big)$ indicates the trace of the covariance matrix (4).

4 Application to cyber security survey data

The above methodology is adopted on the CSBS data in order to investigate the determinants underpinning costs of cyber attacks. Focusing on for-profit companies, the response of interest is the cost (in £1000) related to security breaches experienced by the firm in the last 12 months. This variable is recorded through intervals only, and in the pooled sample from 2018 and 2019 survey waves the first class $[0, 0.5)$ is observed 22.7% of the times. In the first stage (1), the vector x_i contains 11 dummy variables to identify the ith firm's UK industrial sector; in the second stage (2), we considered in z_i all the remaining firm-level predictors ($i = 1, \ldots, n$). The main one is another banded variable, i.e. the amount invested in cyber security (Invest, in £1000). Other predictors are the number of breaches (Nbreach), the presence of a cyber security incident management process (Incid), and the sales turnover (Sales, in £million) that controls for the company's size. The time predictor is coded as a numerical variable, equal to 1 or 2 according to the first or second wave. To assess the effect of investments over time, we also include the interaction term between time and amount invested in security (Time×Invest). For fitting purposes, the values of banded variables are taken as the logarithm of the observed intervals mid-points.
The estimate of $\beta^{(1)}$ is not shown here to conserve space. However the robust ZIIR approach points out, differently from biased ML inference, that the Administration or Real Estate sector is significantly the most likely to incur a loss from cyber attacks. Estimates of the second-stage parameters

TABLE 1. ZIIR model fitted via ML (left) and robust exponential tilting (right). Statistical significance of coefficients at the 5% level is marked by an asterisk. Bootstrap standard errors given in parentheses and bootstrap MSEs in the last line are based on 2000 replications. The CSBS dataset is available at www.gov.uk/government/collections/cyber-security-breaches-survey.

$\hat{\beta}^{(2)}$	ZIIR $\alpha = 0$	Robust ZIIR $\alpha = 0.74$
Incid	0.09 (0.22)	-0.27 (0.18)
Sales	0.19 (0.07)∗	0.16 (0.05)∗
Invest	0.50 (0.16)∗	0.38 (0.10)∗
Time	1.28 (0.96)	1.28 (0.71)∗
Time×Invest	-0.16 (0.10)	-0.13 (0.07)∗
Nbreach	0.07 (0.04)	0.05 (0.04)
$\widehat{MSE}_{boot}(\alpha)$	735.43	33.78

in the robust ZIIR model for the optimal $\alpha = 0.74$ are reported in Table 1, along with ML estimates. The estimated coefficient on Sales suggests that larger companies typically sustain more substantial losses, as found by Romanosky (2016). The significant positive coefficient of the investments variable is evidently due to reverse causality and confirms previous findings (e.g., Woods and Böhme, 2021). Instead, the significance in the robust ZIIR of the negative interaction coefficient with time supports the claim that investing in cyber defence is effective for reducing the cost of cyber attacks. Overall, the superiority of the robust approach with respect to ML is attested not only by the smaller size of the bootstrap standard errors derived from (4), but also by a dramatically lower bootstrap MSE in (5).

References

Brown, S., Duncan, A., Harris, M.N., Roberts, J., and Taylor, K. (2015). A zero-inflated regression model for grouped data. *Oxford Bulletin of Economics and Statistics*, **77**, 822 – 831.

Choi, E., Hall, P., and Presnell, B. (2000). Rendering parametric procedures more robust by empirically tilting the model. *Biometrika*, **87**, 453 – 465.

Genton, M.G. and Hall, P. (2015). A tilting approach to ranking influence. *Journal of the Royal Statistical Society, Series B*, **1**, 77 – 97.

Romanosky, S. (2016). Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, **2**, 121 – 135.

Woods, D.W. and Böhme, R. (2021). SoK: Quantifying cyber risk. In: *2021 IEEE Symposium on Security and Privacy (SP)*, IEEE, pp. 211 – 228.

A new refined non-homogeneous dynamic Bayesian network with globally coupled network interaction parameters

Abdul Salam[1,2], Marco Grzegorczyk[1]

[1] Bernoulli Institute, Groningen University, Groningen, Netherlands
[2] Department of Statistics, University of Malakand, Chakdara, Dir Lower, KP, Pakistan

E-mail for correspondence: m.a.grzegorczyk@rug.nl

Abstract: Non-homogeneous dynamic Bayesian networks (NH-DBNs) are a popular class of statistical models for learning structures of networks that have non-constant network interaction parameters. For example, in some applications the network interaction parameters may vary over time and/or the interaction parameters may vary with changing experimental conditions. In this paper, we propose a new refined NH-DBN with globally coupled interactions parameters. The new model improves upon an earlier proposed globally coupled model, as it introduces segment-specific coupling strength parameters. Our empirical results show that the new NH-DBN can yield a higher network reconstruction accuracy than competing NH-DBN models, including the original globally coupled NH-DBN.

Keywords: Network learning; Dynamic Bayesian networks; Bayesian regression

1 Introduction

An important statistical task in computational systems biology is to learn the structures of cellular networks from experimental wet-laboratory data. Examples of important cellular networks are gene regulatory networks, protein signalling pathways and metabolic pathways. The traditional class of (homogeneous) dynamic Bayesian network (DBN) models is inappropriate if network interaction parameters are not constant. For example, in some applications the interaction parameters might vary with changing experimental conditions under which data have been collected, and/or the interaction parameters might be time-varying when data have been collected over a long time interval. With regard to our application to a small

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

benchmark gene expression data set from yeast (*S. cerevisiae*), we here describe the model in terms of the second case.
Many non-homogeneous DBN (NH-DBN) models employ multiple change-point processes to divide a long time series into shorter time series segments, and they assume the inferred short segments to have different ('segment-specific') network interaction parameters. NH-DBNs then infer the network structure along with a segmentation of the time series as well as the segment-specific network parameters from the data. To avoid model over-flexibility, it was proposed to couple the network interaction parameters among segments; so as to encourage them to stay similar across segments. Conceptually, two different coupling schemes can be distinguished: The sequential coupling scheme makes use of the temporal order of the segments and encourages neighbouring segments to have similar parameters. That is, the network parameters of segment h are encouraged to stay similar to those of the preceding segment $h-1$ (Grzegorczyk and Husmeier, 2012). The global coupling scheme treats the segments $h = 1, \ldots, H$ as interchangeable units and encourages the segment-specific network parameters to stay similar to each other by imposing a restrictive hierarchical hyper-prior (Grzegorczyk and Husmeier, 2013). Although the results of comparative evaluation studies suggest that the global coupling scheme leads to better results than the sequential coupling scheme (see, e.g., the results reported in Grzegorczyk and Husmeier (2019)), in recent years only advanced sequential coupling schemes were proposed; see, e.g., Shafiee Kamalabad and Grzegorczyk (2018) for an example. Here we fill a gap and focus on improving the global coupling scheme, and we propose a new refined model variant of the NH-DBN with globally coupled interaction parameters. Instead of coupling all segments with the same coupling strength parameter λ^2, like in the original work by Grzegorczyk and Husmeier (2013), we here propose to introduce segment-specific coupling strength parameters, $\lambda_1^2, \ldots, \lambda_H^2$, so that each time series segment h can be coupled with its own individual coupling strength λ_h^2. We note that the modelling idea is borrowed from the work by Shafiee Kamalabad and Grzegorczyk (2018), in which a similar extension for the NH-DBN with sequentially coupled parameters was proposed.

2 Statistical modelling

Consider a piece-wise linear Bayesian regression model with a response Y and a covariate set $\boldsymbol{\pi} = \{X_1, \ldots, X_k\}$. We assume that there are temporal data points that have been divided into disjoint segments $h = 1, \ldots, H$, where each segment h has a segment-specific regression coefficient vector, $\boldsymbol{\beta}_h = (\beta_{h,0}, \beta_{h,1}, \ldots, \beta_{h,k})^{\mathsf{T}}$. Let $\mathbf{y}_h$ be the response vector and $\mathbf{X}_h$ be the design matrix for segment h, where each $\mathbf{X}_h$ has a first column of 1's for

the intercept. We assume for the likelihood:

$$\mathbf{y}_h|(\boldsymbol{\beta}_h, \sigma^2) \sim \mathcal{N}(\mathbf{X}_h\boldsymbol{\beta}_h,\ \sigma^2\mathbf{I}) \qquad (h = 1, \ldots, H) \tag{1}$$

and for the segment-specific regression coefficient vectors we use the priors:

$$\boldsymbol{\beta}_h|\left(\boldsymbol{\mu}, \sigma^2, \lambda_h^2\right) \sim \mathcal{N}(\boldsymbol{\mu},\ \sigma^2\lambda_h^2\mathbf{I}) \qquad (h = 1, \ldots, H) \tag{2}$$

where $\sigma^{-2} \sim \text{GAM}(0.005, 0.005)$, $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\lambda_h^{-1} \sim \text{GAM}(2, 0.2)$.

For the density of the posterior distribution, we then have:

$$\begin{aligned} &p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H, \lambda_1^2, \ldots, \lambda_H^2, \sigma^2, \boldsymbol{\mu}|\mathbf{y}_1, \ldots \mathbf{y}_H) \\ \propto\ &p(\boldsymbol{\mu}) \cdot p(\sigma^2) \cdot \prod_{h=1}^{H} p(\mathbf{y}_h|\boldsymbol{\beta}_h, \sigma^2) \cdot p(\boldsymbol{\beta}_h|\boldsymbol{\mu}, \sigma^2, \lambda_h^2) \cdot p(\lambda_h^2) \end{aligned}$$

With regard to the model inference, we note that the marginal likelihood, $p(\mathbf{y}_1, \ldots, \mathbf{y}_H|\lambda_1^2, \ldots, \lambda_H^2, \boldsymbol{\mu})$, with σ^2 and the regression vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_H$ integrated out, can be computed analytically.

Reversible Jump Markov Chain Monte Carlo (RJMCMC)
In many applications the covariates $\boldsymbol{\pi}$ and the time series segmentation are unknown and have to be inferred from the data.
First, we assume all covariate sets with up to 3 covariates to be equally likely, $p(\boldsymbol{\pi}) = c$ if $|\boldsymbol{\pi}| \leq 3$, while $p(\boldsymbol{\pi}) = 0$ if $|\boldsymbol{\pi}| > 3$. Second, we assume the distance between changepoints to be geometrically distributed with hyperparameter $p \in (0, 1)$, and we identify H segments with a changepoint set, $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{H-1}\}$; data point t belongs to segment h if $\tau_{h-1} < t \leq \tau_h$.

RJMCMC simulations can then be used to generate posterior samples $\{\boldsymbol{\pi}^{(s)}, \boldsymbol{\tau}^{(s)}, \{\lambda_h^{2,(s)}\}_h, \boldsymbol{\mu}^{(s)}\}_{s=1,\ldots,S}$. Our sampling scheme uses the marginal likelihood, and for sampling the covariates $\boldsymbol{\pi}$ and the changepoints $\boldsymbol{\tau}$ we implement Metropolis-Hastings moves: Changepoint birth, death and re-allocation moves on the changepoint set $\boldsymbol{\tau}$, and covariate addition, deletion and exchange moves on the covariare set $\boldsymbol{\pi}$. As the latter moves are trans-dimensional, there is need for RJMCMC sampling techniques. We design the RJMCMC sampling moves along the lines of Shafiee Kamalabad and Grzegorczyk (2018). For lack of space, we cannot provide the mathematical details here.

Network learning and interaction scores
To learn a network among variables $Z_1, \ldots, Z_n$ we apply the regression model to each response $Y := Z_i$ separately. The potential covariate sets $\boldsymbol{\pi}_i$ for response $Y = Z_i$ are all subsets of the remaining variables, symbolically $\boldsymbol{\pi}_i \subset \{Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n\}$.

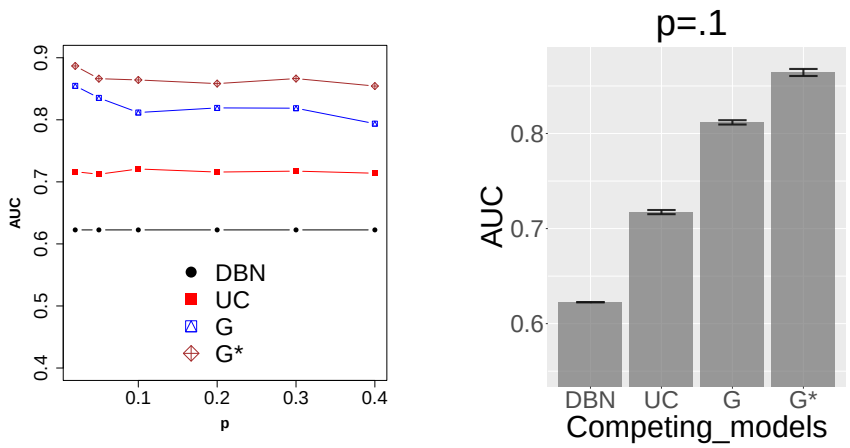


FIGURE 1. **Comparison of the network learning performances of four NH-DBN models.** *Left*: Model-specific average AUCs (vertical axis) for 6 hyperparameters p (horzontal axis). *Right:* Example bar plot of mean AUCs with 95% confidence interval for the hyperparameter $p = 0.1$. See main text for further details on AUC scores and the four NH-DBN models: DBN, UC, G and G*.

For each interaction, $Z_j \to Z_i$, we estimate an interaction score

$$\hat{e}_{j,i} = \frac{1}{S} \sum_{s=1}^{S} I_{j \to i}(\boldsymbol{\pi}_i^{(s)})$$

where $\boldsymbol{\pi}_i^{(s)}$ is the s-th sampled covariate set for Z_i, $I_{j \to i}(\boldsymbol{\pi}_i^{(s)}) = 1$ if $Z_j \in \boldsymbol{\pi}_i^{(s)}$, and $I_{j \to i}(\boldsymbol{\pi}_i^{(s)}) = 0$ else. When the true network structure (= the set of all network interactions) is known, the network reconstruction accuracy can be quantified in terms of areas under the precision-recall curve, $AUC \in [0, 1]$. The higher the AUC, the higher the network reconstruction accuracy.

3 Empirical results

We apply the newly proposed NH-DBN model to a small benchmark yeast gene expression data set. In this application, the gene regulatory processes are time-varying, because data were collected during an experimentally imposed transition from galactose to glucose metabolism. Moreover the network was designed by means of synthetic biology, so that the true network structure is known; cf. left panel of Figure 2. We refer to the work by Cantone et al. (2009) for further details on the yeast data.

We use the benchmark yeast data set to cross-compare the learning performances of four NH-DBN models. For each model we use six different

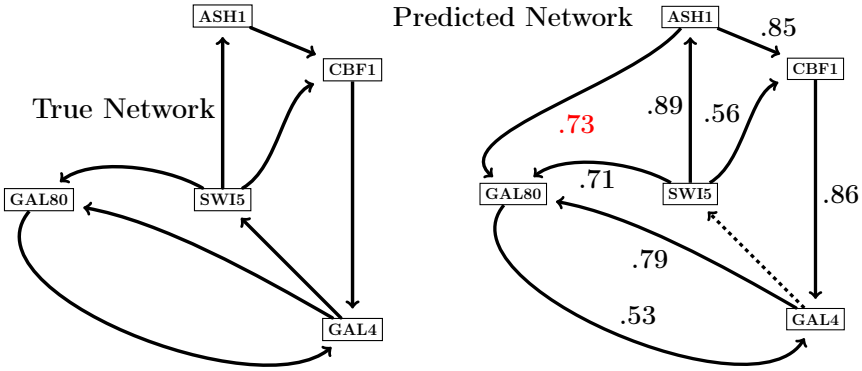


FIGURE 2. **True (left) and predicted (right) yeast network.** The true yeast network consists of 8 gene interactions. When imposing the threshold $\psi = 0.5$ on the interaction scores, $\hat{e}_{j,i}$, the newly proposed NH-DBN model ($\mathbf{G}^\star$) learns 8 interactions. The predicted network structure features seven true positive interactions, one false positive interaction (ASH1→GAL80) and one false negative interaction (GAL4→SWI5). Precision and recall are both equal to 0.875.

hyperparameters $p \in (0, 1)$ for the geometric distribution on the distances between changepoints. We vary the hyperparameter p, as it can be expected that the number of changepoints (and so the number of data segments) increases in the hyperparameter p.

The average network reconstruction scores (AUCs) are shown in the left panel of Figure 1. The new refined globally coupled NH-DBN ($\mathbf{G}^\star$) performs consistently better than the three competing models, namely:

- a standard homogeneous dynamic Bayesian network (**DBN**)
- an uncoupled NH-DBN with independent regression coefficient per data segment (**UG**)
- the less flexible originally proposed globally coupled model (**G**).

Moreover it can be seen that the AUC results are rather robust with respect to the hyperparameter p. The right panel of Figure 1 shows an exemplary error bar plot of the average AUCs with 95% confidence intervals for the hyperparameter $p = 0.1$. The tiny AUC variations refer to independent RJMCMC simulations that started from different random initialisations. From the error bar plot it can be seen that the improvements that are achieved with the new model are significant.

The two panels of Figure 2 show the true yeast network from Cantone et al. (left panel) and the network structure that was inferred with the newly

proposed NH-DBN model ($\mathbf{G}^{\star}$) using the hyperparameter $p = 0.1$ (right panel). It can be seen that the predicted network structure is rather close to the true network structure. Among the 8 gene interactions with the highest scores ($\hat{e}_{j,i} > 0.5$) there are 7 true positive gene interactions. Precision and recall are both equal to 0.875 (7/8). In terms of the area under the precision recall curve, the inferred network yields $AUC \approx 0.88$.

4 Conclusions

In this paper we have proposed a new refined variant of the globally coupled non-homogeneous dynamic Bayesian network (NH-DBN) model from Grzegorczyk and Husmeier (2013). Our empirical results on a small benchmark yeast gene expression data set from synthetic biology have shown that the proposed model refinement can lead to a higher network reconstruction accuracy. The yeast network structure that was inferred with the new NH-DBN model, shown in the right panel of Figure 2, is rather close to the true yeast network structure.

References

Cantone et al. (2009) A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell*, **137**, 172-181.

Grzegorczyk, M. and Husmeier, D. (2012). A non-homogeneous dynamic Bayesian network model with sequentially coupled interaction parameters for applications in systems and synthetic biology. *Statistical Applications in Genetics and Molecular Biology (SAGMB)*, **11**(4), Article 7.

Grzegorczyk, M. and Husmeier, D. (2013). Regularization of non-homogeneous dynamic Bayesian networks with global information-coupling based on hierarchical Bayesian models. *Machine Learning*, **91**(1), 105-154.

Shafiee Kamalabad, M. and Grzegorczyk, M. (2018). Improving nonhomogeneous dynamic Bayesian networks with sequentially coupled parameters. *Statistica Neerlandica*, **72**(3), 281-305.

Grzegorczyk, M. and Husmeier, D. (2019). Chapter 32: Modelling Non-homogeneous Dynamic Bayesian Networks with Piecewise Linear Regression Models. In: Balding, D.J., Moltke, I., Marioni J. (editors), *Handbook of Statistical Genetics, 4th edition, vol. 2*, John Wiley & Sons. 899-931.

Statistical Downscaling of Air Temperature using Variational Autoencoded Regression

Yire Shin[1], Jeong-Soo Park[1]

[1] Department of Mathematics and Statistics, Chonnam National University, Gwangju 61186, Korea

E-mail for correspondence: shinyire87@gmail.com

Abstract: This study was conducted to evaluate the effectiveness of a new downscaling technique based on variational autoencoder (VAE). This paper proposes a new urban-scale downscaling framework method by merging a VAE and VAE regression. The proposed method introduces a strategy to regressor the latent space of a complex urban-scale temperature image which is pre-trained by a VAE model. The regressed response in the latent space is embedded into a generative model so that local-scale temperature is able to estimate the urban-scale. This study demonstrate that the proposed technique is applicable to urban-scale meteorology research and potentially applicable other areas.

Keywords: Downscaling; Variational Autoencoder; Temperature

1 Introduction

Nowadays, most weather disaster occur at local or micro-scale level. Downscaling is necessary and increasingly being used as a method to handle detailed spatiotemporal meteorological information. The downscaling techniques with deep learning methods are cost-effective and easier to implement than nested mesoscale models and have proven to be as accurate as those with dynamical downscaling methods.

Accarino et al. (2021) presented a multi-scale gradients generative adversarial network for statistical downscaling of 2-meter temperature over the European domain, the results show an accurate and cost-effective solution of downscaling to the 2m temperature climate maps. Kingma and Welling (2014) introduced a powerful algorithm for efficient inference and learning Auto-Encoding variational Bayesian (AEVB) learned and an approximate inference model using the stochastic gradient variational Bayes (SGVB) estimator. Laubscher and Rousseau (2020) applied generative deep learning

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

to predict temperature and other data types. Zhao et al. (2019) introduced a generalized regression model based on the variational autoencoder (VAE) framework and applied it to the problem of predicting age from structural MR images, where this method provides more accurate brain predictions than regular feed-forward regressor network.
We studied the downscaling of air temperature with VAE regression model in deep learning.

2 Method

2.1 Overal Scheme

The study region, Seoul (605 km^2) is the capital city of and largest metropolis in the Republic of Korea, in Northeast Asia. The response y is defined as urban-scale temperature field(50×90 gird by 500 m) from observed data. The corresponding input x is defined local-scale temperature prediction field(29×20 grid by 1.5 km) from LDAPS. The data collection period was 2018-2020.

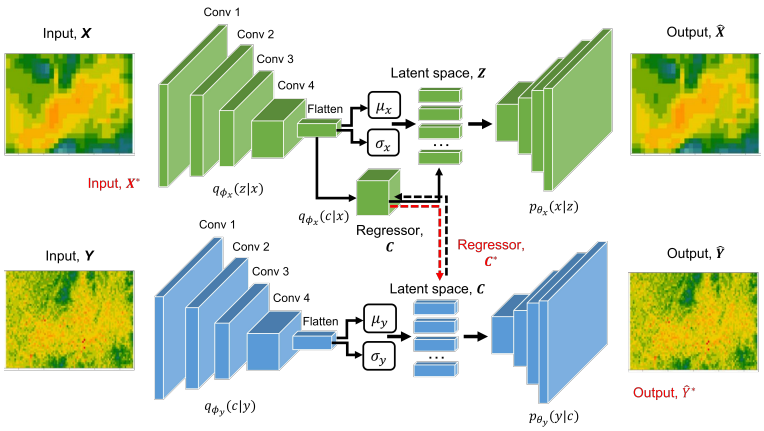


FIGURE 1. Overall scheme of the proposed method.

We propose a new downscaling framework method by merging VAE and VAE regression. As shown in Figure 1, for the observed data $y_i, i = 1, \cdots, N$ the encoder/decoder produces which is the reconstruction of an observed data y_i. For the observed data, the encoding network $q_{\phi_y}(z|y)$ produces mean (μ_y) and variance (σ_y^2) for a part of the latent vector c_i. Using sampled from $N(\mu_y, \sigma_y)$, the decoding network $p_{\theta_y}(y|c)$ reconstructs the output response $\hat{y}_i$. For the input data x_i, it is difficult to produce to y_i directly because y_i have a high dimension space than x_i. To produce

y_i, we estimate a low-dimensional latent vector c_i from the encoding network $q_{\phi_y}(c|y)$ using a regressor $q_{\phi_x}(c|x)$. After $\hat{c}$ is estimated, the output response $\hat{y}_i$ is reconstructed from $\hat{c}$ using the decoding network $p_{\theta_y}(y|\hat{c})$.

3 Experiment

We applied the proposed method to air temperature forecasts with 1.5-km resolution (LDAPS) for Seoul metropolitan area and downscaled it to 500-m resolution. We performed an accuracy test of the proposed downscaling method with regard to the hour temperature during the three month(July, August and September 2020, 09:00-23:00). We predicted temperatures were compared to the AWS observation. The respective mean value of RMSE and R^2 were calcuated (Table 1): 1.28 and 0.70, respectively, for $\hat{y}$, 0.03 and 0.67, respectively, for $\hat{c}$, 1.36 and 0.62, respectively, for $\hat{x}$. The estimation of the latent vector from the encoding network $q_{\phi_y}(c|y)$ was found to be well estimated using the regression term $q_{\phi_x}(c|x)$.

TABLE 1. The respective mean value of $RMSE$ and R^2 for the test periods

	$\hat{y}$	$\hat{r}$	$\hat{x}$
$RMSE$	1.28	0.03	1.35
R^2	0.70	0.67	0.62

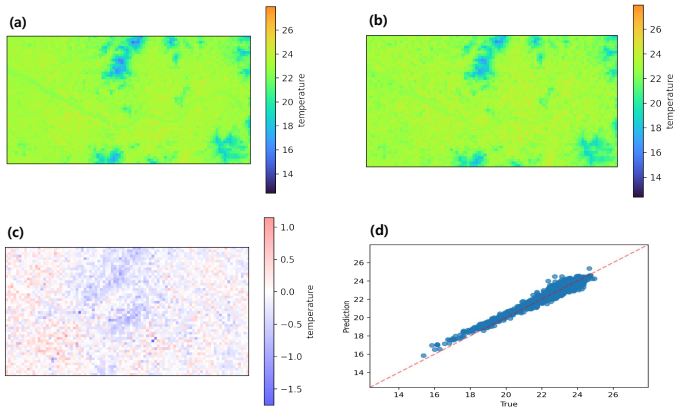


FIGURE 2. Spatial accuracy evaluation for test data (predict day : September 29, 2020, 14:00 KST). (a) True spatial distribution(y), (b) Predicted spatial distribution($\hat{y}$). (c) The difference between y and $\hat{y}$ (RMSE 0.294), (d) Scatterplot of y and $\hat{y}$ (R^2 0.928)

4 Conclusion

We proposed a new urban-scale downscaling framework method by merging a variational auto encoder (VAE) and VAE regression. We applied the proposed method to air temperature forecasts with 1.5-km resolution (LDAPS) for Seoul metropolitan area and downscaled it to 500-m resolution. The result of this experiment, the estimated $\hat{y}$ showed spatially high accuracy and similarity to y. In some results, however, there was a significant difference between y and $\hat{y}$. Further studies should be performed considering a time series such as LSTM, which can lead to robust results and a more precise quantification. This study suggest that the proposed a new downscaling approach can be applied to solving problems related to urban-scale meteorology, as well as its potentially applicable to other metropolitan areas.

Acknowledgments: This research was supported by Basic Science Research Program through the NRF funded by the Ministry of Education (2021R1A6A3A13044162), by the BK21 FOUR (NO.5120200913674) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea (NRF)

References

Accarino, G., Chiarelli, M., Immorlano, F., Aloisi, V., Gatto, A., Aloisio, G. (2021). MSG-GAN-SD: A Multi-Scale Gradients GAN for Statistical Downscaling of 2-Meter Temperature over the EURO-CORDEX Domain. *AI.* **2(4)**. 600 – 620.

Kingma, D. P., Welling, M. (2014). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*..

Laubscher, R., Rousseau, P. (2020). Application of generative deep learning to predict temperature, flow and species distributions using simulation data of a methane combustor. *International Journal of Heat and Mass Transfer*, **163**. 120417.

Zhao, Q., Adeli, E., Honnorat, N., Leng, T., Pohl, K. M. (2019). Variational autoencoder for regression: Application to brain aging analysis. *International Conference on Medical Image Computing and Computer-Assisted Intervention.*, Springer. **11765**, 823 – 831.

Option pricing using Hawkes Process

Shubhangi Sikaria[1], Rituparna Sen [2]

[1] Indian Institute of Technology, Madras, India
[2] Indian Statistical Institute, Bangalore, India

E-mail for correspondence: shubhangisikariya@gmail.com

Abstract: We propose a methodology for European options pricing in which the Hawkes process drives variations in asset prices. This construction preserves the Brownian diffusion behavior when it comes to the microstructure level. We associate two point processes corresponding to the sum of the asset's positive and negative jumps, respectively. The point processes have a self and mutually exciting stochastic intensities with an exponential kernel. We employ a mean signature plot to estimate the parameters. We examine the model's implementation in real data applications and compare it with the Black-Scholes formula.

Keywords: Microstructure noise; Point process; Signature plot.

1 Introduction

The recent availability of high-frequency financial data prevails the issue of controlling the market microstructure noise. To circumvent the effect of microstructure noise, Engle and Russell (1998) introduced the first point process with dependent arrival rates, namely Autoregressive Conditional Duration (ACD) model for irregular intervals. Later, Bowsher (2007) modeled multivariate market events such as the timing of trades and mid-quote changes using the Hawkes process with vector conditional intensity. Hawkes process, introduced by A.G. Hawkes (1971), is a class of multivariate point processes with self and mutually exciting stochastic intensities.
One can find wide applications of the Hawkes process in finance. Barcy (2015) provides an excellent survey of Hawkes process in finance, focusing on high frequency and market microstructure noise. Barcy (2013) introduced a tick-by-tick model using the multivariate Hawkes process, which deals with the effect of microstructure noise and also preserves the Brownian diffusion behavior on large scales. In this paper, we work on similar lines as Barcy (2013) to price the options. The model is associated with

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

two point process corresponding to positive and negative jumps of the asset prices with appropriate stochastic intensities.
The rest of the paper is organized as follows. Section 2 illustrates the Hawkes process in detail, explaining the bivariate counting process and intensity functions. Section 3 provides the methodology followed in this paper, including a detailed explanation of parameter estimation using signature plot and the option pricing algorithm. Section 4 presents the empirical results, comparing the proposed method and the Black Scholes model. Section 5 delivers the concluding remarks.

2 Hawkes Process

We have considered two counting process $N_1(t)$ and $N_2(t)$ over time horizon $t \in [0, T]$ representing the sum of positive and negative jumps of asset price respectively. Let $X(t)$ be the price of the asset for $t \in [0, T]$ such that:

$$X(t) = N_1(t) - N_2(t). \tag{1}$$

As defined by Barcy (2013), the bivariate process $\{N_1(t), N_2(t)\}$ is said to be linear Hawkes process if $N_1(t)$ and $N_2(t)$ have no common jumps and if there exist non-negative functions $\{\phi_{ij}\}_{i,j=1,2}$ such that:

$$\lambda_i(t) = \mu_i + \int_{-\inf}^{t} \phi_{ij}(t-s)dN_i(s) + \int_{-\inf}^{t} \phi_{ij}(t-s)dN_j(s). \tag{2}$$

We have considered a simplified version of the intensity function with only mean reverting terms, given as:

$$\begin{aligned} \lambda_1(t) &= \mu + \int_{-\infty}^{t} \phi(t-s)dN_2(s) \\ \lambda_2(t) &= \mu + \int_{-\infty}^{t} \phi(t-s)dN_1(s) \end{aligned} \tag{3}$$

where μ is an exogenous intensity and $\phi(t)$ is a right-sided exponential kernel defined as:

$$\phi(t) = \omega \exp(-\delta t) 1_{\mathrm{R}^+}(t) \tag{4}$$

where ω, $\delta > 0$ and $\phi(t)$ satisfy the stability condition:

$$||\phi||_1 = \frac{\omega}{\delta} < 1. \tag{5}$$

3 Methodology

3.1 Parameter estimation

The estimation of the parameters can be done using the best fit of the realized signature plot to the mean signature plot. For exponential kernel

defined in (4) with the stability condition (5), we have the mean signature plot as:

$$C(\tau) = \frac{2\mu\delta}{(\delta-\omega)(\omega+\delta)^3\tau}[\delta^2 + \omega(\omega+2\delta)(1-\exp\{-(\omega+\delta)\tau\}]. \quad (6)$$

The realized signature plot is the realized volatility of $X(t)$ over $[0,T]$, defined as:

$$\hat{C}(\tau) = \frac{1}{T}\sum_{n=0}^{T/\tau} |X((n+1)\tau) - X(n\tau)|^2. \quad (7)$$

There are mainly 3 parameters which we have defined using the parameter space $\theta = (\mu, \omega, \delta)$ and let $F(\theta)$ is the regression estimator given as:

$$F(\theta) = |\hat{C}(\tau) - C(\tau)|^2. \quad (8)$$

We applied Newton Rapson iterations on the regression estimator to obtain the parameter space θ.

3.2 Option Pricing

In this section, we explain the procedure to price the options:

1. Choose M large enough such that:

$$\lambda_1(t) < M \quad \text{and} \quad \lambda_2(t) < M, \forall t \in [0,T].$$

2. Simulate on $[0,T]$ a standard Poisson process with an intensity M and apply thinning procedure to each jump of the obtained process as follows:
   (a) Reject the point with probability $(M - \lambda_1(t) - \lambda_2(t))/M$
   (b) Mark the point as jump of $N_1(t)$ with probability $\lambda_1(t)/M$
   (c) Mark the point as jump of $N_2(t)$ with probability $\lambda_2(t)/M$
3. Obtain the cumulative sum of jumps corresponding to $N_1(t)$ and $N_2(t)$, written as $\tilde{N}_1(t)$ and $\tilde{N}_2(t)$ respectively. The simulated path $\tilde{X}(t)$ is then given as:

$$\tilde{X}(t) = \tilde{N}_1(t) - \tilde{N}_2(t), \quad \forall t \in [0,T]. \quad (9)$$

4. Using the simulated path of $\tilde{X}(t)$, the simulated option price $\tilde{O}_T$ is:

$$\tilde{O}_T = (\tilde{X}(T) - K)^+, \quad \text{where } K \text{ is the strike price.} \quad (10)$$

5. Repeat step 2 to step 4 for a large number N simulations and take an average of the simulated option prices.

4 Empirical Results

We apply the modeling framework to the Indian stock market index Nifty50 corresponding to one-day and one-week option prices for an empirical illustration. For our study, we have downloaded the tick-by-tick asset price from Dukascopy Swiss Banking Group. The options data corresponding to the same asset is obtained from NSE India website. The time period of the data is 3.5 months, from Jun 2021 to Sep 2021. An initial period of 3 months of this data is a training sample, which we used for stable parameters estimation. The remaining dataset is used as a test sample.
We have shown Nifty option prices data on 1-Sep-2021 expiring on 2-Sep-2021 and 9-Sep-2021 corresponding to various strike prices. Model price is the price obtained from applying the methodology described in section 3. Black Scholes price is the price obtained from employing the famous Black Scholes model. Price of Nifty50 on 1-Sep-2021 is 17076.25. In Table 1 and 2, we also present the last traded prices to test our model. It is observed that our model gives better prediction compared Black Scholes model which is generally upward biased.

TABLE 1. Nifty option prices on 1-Sep-2021 expiring on 9-Sep-2021 compared to purposed model and Black Scholes prices.

Strike Price	Last Traded Price	Model Price	Black Scholes Price
15000	2132.8	2383.921	2502.19
15500	1572.60	1883.92	2153.53
16000	1080.8	1384.38	1836.41
16500	567.05	894.09	1551.68
17000	88	495.17	1299.31

TABLE 2. Nifty option prices on 1-Sep-2021 expiring on 9-Sep-2021 compared to purposed model and Black Scholes prices.

Strike Price	Last Traded Price	Model Price	Black Scholes Price
15500	1589.1	1559.58	1658.89
16000	1089.25	1059.58	1242.76
16500	595.45	559.58	883.42
17000	168.2	40.99	592.66
17500	14.05	3.63	373.83

5 Conclusion

We have priced the options using a bivariate tick-by-tick asset price model with stochastic intensities. The actual data study shows that pricing using

Hawkes process produces better results than the Black-Scholes model.

References

Bacry, E., Delattre, S., Hoffmann, M., & Muzy, J. F. (2013). Modelling microstructure noise with mutually exciting point processes. *Quantitative finance*, **13(1)**, 65 – 77.

Bacry, E., Mastromatteo, I., & Muzy, J. F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, **1(01)**, 1550005.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, **141(2)**, 876 – 912.

Engle, R. F., & Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica,* 1127 – 1162.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58(1)**, 83 – 90.

MD-dating of pinewood using FTIR-spectroscopy and statistical learning algorithms like random forests and CNNs

Bernhard Spangl[1], Matthias Medl[1], Johannes Tintner[2]

[1] Institute of Statistics, Department of Landscape, Spatial and Infrastructure Sciences, University of Natural Resources and Life Sciences, Vienna, Austria
[2] Institute of Physics and Materials Science, Department of Material Sciences and Process Engineering, University of Natural Resources and Life Sciences, Vienna, Austria

E-mail for correspondence: bernhard.spangl@boku.ac.at

Abstract: More than 65 years after the publication of C14 dating we present a new chronometric method for the dating of wood based on molecular decay (MD). Infrared spectroscopy is used to detect the chemical changes over time. The presented prediction models cover a maximum of 815 years. The models considered here are valid for Scots pine (*Pinus sylvestris*).

Keywords: Dendrochronology; FTIR-spectroscopy; Random forest model; convolutional neural network (CNN); feature importance.

1 Introduction

In this study we applied the random forest method and one-dimensional convolutional neural networks (CNNs) to model the age of pinewood. Previous studies (Tintner et al., 2020a, 2020b) have shown that different models are needed to predict the age of different tree species.
In the field of machine learning, tree-based methods for regression are well established (Hastie et al. 2017; Gareth et al. 2021). Although tree-based methods are simple and useful for interpretation they lack prediction accuracy, i.e., they produce good predictions on the training set, but are likely to overfit the data, leading to poor test set performance. To overcome these drawbacks the random forest method was introduced by Breiman (2001).
On the other hand, since their introduction by LeCun et al. (1989) in the early 1990's, CNNs have demonstrated excellent performance at classifica-

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

tion and prediction tasks. However there is still no clear understanding of why they perform so well.
The objective of this work is to establish a dating tool for wooden artefacts based on the relation between chemical characteristics of the decay and time. The chemical decay is revealed by means of infrared spectroscopy—a rapid and cheap analytical method. Dendrochronology serves as the reference method. Additionally, we focus on the explainability of the models.

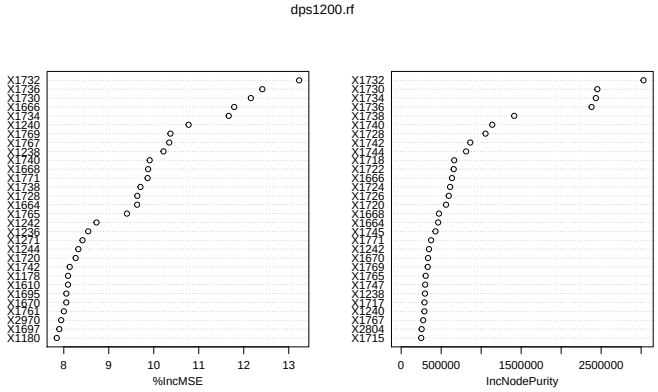


FIGURE 1. Variable Importance Plot of the random forest model; the 30 most important variables are plotted.

2 Materials and methods

2.1 Sample description

The pinewood samples (*Pinus sylvestris*) investigated in this study came from Europe. We restricted our analysis to pieces younger than AD 1,200. Approximately every tenth tree ring of the samples was recorded. The measured FTIR-spectra were smoothed, and the second derivative was applied. Both spectra, the smoothed as well as the second derivative one, were used for statistical analysis.

2.2 Statistical analysis

All statistical analysis was done using the statistical computer software language R. The R package randomForest (Liaw and Wiener, 2002) was used to fit random forest models to the data. CNNs were trained using the R package keras (Allaire and Chollet, 2022).
Hence, the data were analyzed as follows. First, either a random forest model or a CNN was fitted to the data. This step will be referred as basic model in the following. As the predicted values y_i of the basic model

often underestimate the true year x_i for all species, especially for very old probes, we additionally calibrate the predicted years. We call this combined approach 'MD-dating'.
Both approaches, basic modelling and MD-dating, were compared by 10-fold cross validation based on their averaged RMSEP.

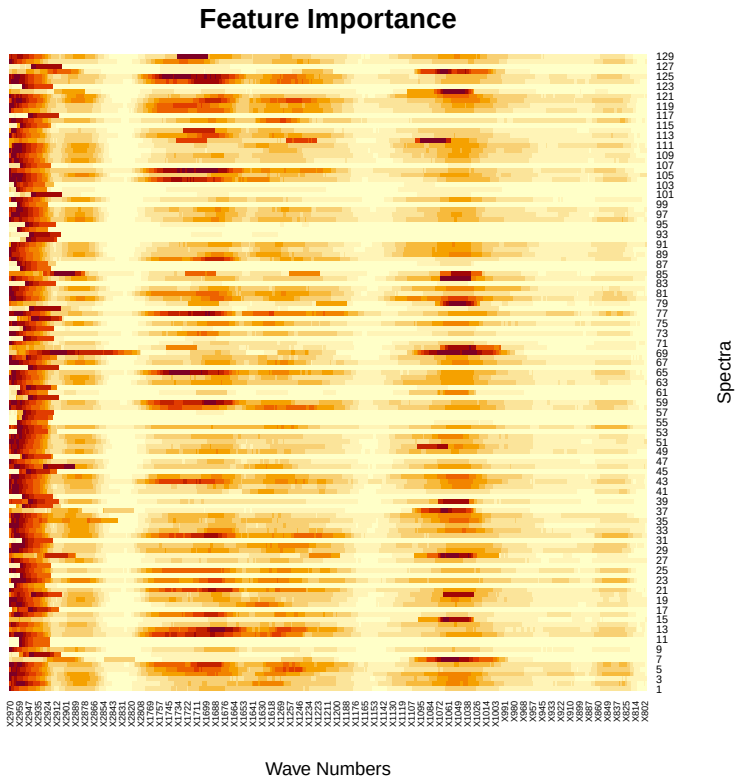


FIGURE 2. Measuring the feature importance in CNN models; dark colors indicate wave numbers that are important for predicting the age of pinewood.

To achieve explainability of the models, i.e. to decide which wave numbers are important for predicting the age of pinewood, usually the Variable Importance Plot is used for random forest models. In Figure 1 the 30 most important wave numbers are listed. For CNN models we adapted the occlusion sensitivity approach proposed by Zeiler and Fergus (2013) to measure the feature importance. The result is plotted in Figure 2. Dark colors indicate wave numbers that are important for predicting the age of pinewood. Comparing the results they agree on most wave numbers. However, the occlusion sensitivity approach found wave numbers between $1100\,\mathrm{cm}^{-1}$ and $1000\,\mathrm{cm}^{-1}$ to be additionally important.

3 Conclusions and outlook

A big strength of this method is the strictly monotonous molecular decay. This distinguishes it from other methods, especially from dendrochronology, where it might happen that different results far away from one other appear equally probable. But in C14 dating there are time spans at which calibration curves bend into a plateau as well. In comparison with C14, dating the comparably far lower costs here serve as another main advantage.
Future work will focus on the comparison with further statistical learning algorithms like generalized additive models (GAMs) and on refining the prediction accuracy taking the longitudinal structure of the measurements into account.
The combination of the spectral method and statistical learning algorithms are a promising approach to stimulate and support the work of building historians, archaeologists, and even environmental scientists.

References

Allaire, J.J. and Chollet, F. (2022). *keras: R Interface to 'Keras'*. R package version 2.8.0.

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

Hastie, T., Tibshirani, R., and Friedman, J.H. (2017). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Gareth, J., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An introduction to statistical learning, 2nd Edition*. New York: Springer.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., and Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, **1**, 541 – 551.

Liaw, A. and Wiener, M. (2002). Classification and regression by random forest. *R News*, **3**, 18 – 22.

Tintner, J., Spangl, B., Reiter, F., Smidt, E., and Grabner, M. (2020a). Infrared spectral characterization of the molecular wood decay in terms of age. *Wood Science and Technology*, **54**, 313 – 327.

Tintner, J., Spangl, B., Grabner, M., Helama S., Timonen M., Kirchhefer A.J., Reinig F., Nievergelt D., Krapiec M., and Smidt, E. (2020b). MD Dating—Molecular decay (MD) in pinewood as a dating method. *Scientific Reports*, **10**.

Zeiler, M.D. and Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *arXiv:1311:2901v3*.

Bayesian spatial modeling of extreme precipitation

Federica Stolf [1] and Antonio Canale[1]

[1] Department of Statistical Sciences, University of Padova, Padova, Italy

E-mail for correspondence: federica.stolf@phd.unipd.it

Abstract: Maps of return levels provide information about the spatial variations of the risk of extreme precipitation and are expected to be useful for infrastructure planning. In this paper we analyze a collection of spatially distributed time series of precipitation in Georgia (USA): exploiting a spatial hierarchical Bayesian model we can produce maps of precipitation return levels with uncertainty measures. Inference about the parameters and spatio-temporal predictions are obtained via Markov Chain Monte Carlo (MCMC) simulation.

Keywords: Extreme value, Bayesian Hierarchical model, Rainfall, Georgia.

1 Introduction

Extreme value theory finds wide application in environmental sciences. Extreme meteorological events such as high rainfall and windstorms arise due to physical processes and are spatial in extent. These events are usually characterized by limited predictability and can cause significant economical and social damages. We mention for example the catastrophic flood that impacted North Georgia, in particular the Atlanta metropolitan area, on September 2009 as a result of multiple days of prolonged rainfall. The flood is blamed for at least 10 deaths and $500 million in damage (National Weather Service). Although these extreme precipitation events are rare, understanding their frequency and intensity is important for public safety and long-term planning.
A rich statistical literature concerned with modeling extreme events is available and Coles (2001) provides a comprehensive introduction. Standard approaches utilize generalized extreme value (GEV) distributions (Fisher and Tippett, 1928) and Generalized Pareto Distribution (GPD) (Gnedenko,

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

1943). In Davison et al. (2012) are identified three main classes of statistical models for spatial extremes: Bayesian hierarchical models, copula based models, and max-stable process models. Latent variable models arise naturally in the Bayesian framework and have been widely used in the context of extremes (see e.g. Cooley et al., 2007). A particular issue when dealing with extremes is that although vast amounts of data may be available rare events are necessarily unusual and so the quantity of directly relevant data is limited. One of the advantages of the Bayesian approach is the possibility to incorporate reliable information supplementary to the data in the form of prior distributions. In this paper we focus on the Bayesian method recently proposed in Stolf and Canale (2022), a specification that does not make any asymptotic assumption and explicitly takes into account the spatial dependence of the data.

2 A hierarchical Bayesian Extreme Value Model

Let $x_{ij}(s)$ denote the magnitude of the i-th event within the j-th block for the site s, where $j = 1, \ldots, J$ with J the number of blocks in the observed sample, $i = 1, \ldots, n_j(s)$ with $n_j(s)$ the number of events observed within the j-th block for the site s and $s = 1, \ldots, S$ with S the total number of stations. Traditional approaches focus on the distribution of block maxima for each station $Y_j(s) = \max_i\{x_{ij}(s)\}$ only, discarding the 'ordinary values'. This approach has several limitations. First, the number of events per block may be often not large enough for the asymptotic argument to hold (Koutsoyiannis, 2004) and second the assumption of a constant parent distribution is unrealistic in many contexts. Based on these considerations a hierarchical Bayesian extreme value model that avoids the asymptotic argument and accounts for possible inter annual variability in the magnitude of the events was introduced in Zorzetto et al. (2020), building upon the results discussed in Marani and Ignaccolo (2015).
By considering some physical process such as rainfall, one can expect that nearby locations will exhibit similar behavior, and in the Bayesian framework the reduction in uncertainty gained from pooling over space is particularly useful. Thus, spatial modelling of extremes is expected to reduce the overall uncertainty in extreme values estimates, by borrowing strength across spatial locations. For these reasons starting from the approach proposed in Zorzetto et al. (2020), Stolf and Canale (2022) include the spatial dependence of the data in the model, incorporating in the layers of the hierarchical model geographical features, to make predictions at unobserved sites. We adopt the latter model specification, where the events within a block, $x_{ij}(s)$, conditionally on unobserved latent processes are assumed to be conditionally independent with common parametric cdf $F(\cdot; \theta_j(s))$, with $\theta_j(s) \in \Theta$ unknown parameter vector. For technical details see Stolf and Canale (2022).

3 Georgia rainfall data analysis

This study uses daily precipitation observations from the United States Historical Climatological Network (USHCN), a high-quality source of data sets freely available. We use all available stations in Georgia which contain more than 80 years of data from 1892 to 2021. The number of stations included is 20.

Data are analyzed applying the spatial hierarchical Bayesian model (sHMEV) of Stolf and Canale (2022). In particular, we consider as geographic covariates to be included in the model, in addition to latitude and longitude, also the altitude since there are some mountainous ranges in the northern part of the region. We fit the model only on the first 20 years of observations for each station and we use the remaining records to validate it.

Adopting a Bayesian methodology allows to make inference on any functional of the posterior distribution (like our target, the cumulative probability of block maxima) and uncertainty measures result naturally from the sampling procedure. Figure 1 shows maps of the predictive pointwise posterior mean for the 25 year return levels, with pointwise 90% credible intervals width. To create these maps the study region was divided into a grid of points, and considering the posterior draws and the values of the covariates for each point is straightforward to obtain draws for the posterior distribution at any grid point. We observe higher return levels for the north

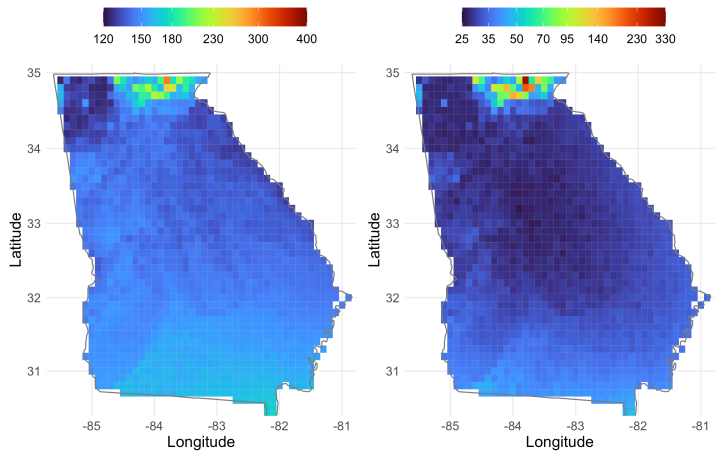


FIGURE 1. Maps of the predictive pointwise 25 year return level estimates for rainfall (mm). Predictive pointwise posterior mean (left) and the width of the 90% pointwise credible intervals (right).

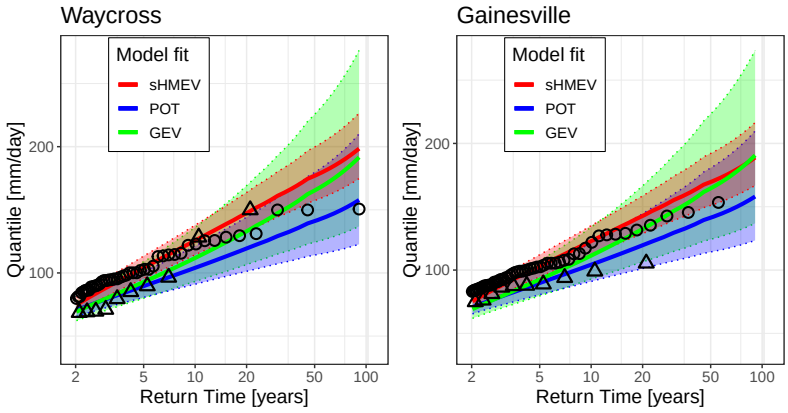


FIGURE 2. Quantiles predicted for the stations of Waycross and Gainesville by the GEV (green), POT (blue), and spatial hierarchical Bayesian (red) models. Solid lines show the expected value of the quantile for a given return time, while dashed lines represent the bounds of 90% credibility intervals. Triangles represent the maxima on the training set, while the circles represent the maxima on the test set.

mountain area and for the south area, in particular on the southeast of the region near the Atlantic Ocean. There are a few points in the mountainous area with rather high return values and great variability. This is due to the high values of altitude for those points, much greater than for the sites in the data, that lead to a more uncertain extrapolation.

In order to evaluate the performance of the model we compare it with Bayesian implementations of standard alternative methods: GEV and peak over threshold (POT). Since the focus is the right tail of the distribution of the block maxima, we evaluate the predictive accuracy in estimating the true distribution of block maxima on the test set. Figure 2 shows two representative examples. Specifically, the quantile versus return time plots obtained for the different methods for two sites, Waycross and Gainesville, are reported. The fist station is in south Georgia, while the second one is located north of Atlanta. For both stations the spatial hierarchical model presents an overall good agreement with the empirical frequencies associated to the annual maxima extracted from the entire record, and yields quantile estimates with narrower credibility intervals. POT and GEV appear to be more sensitive to the smallest observations in the training set and tend to underestimate the quantiles. This behavior is expected given the limited length of the records, as observed in Stolf and Canale (2022). In this case the spatial hierarchical model, exploiting also information from the other sites (*borrowing strength*), manages to obtain more accurate and less variable estimations than the competitors.

This analysis shows the potentiality of the spatial hierarchical models

framework for extreme precipitation, although a more detailed comparison via simulated and real data sets is needed. A major asset of latent variable models is flexibility: the approach discussed above can be generalized or extended to study different natural phenomena.

References

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer-Verlag.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, **102**, 824 – 840.

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Statistical Science*, **27 (2)**, 161 – 186.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In: *Mathematical Proceedings of the Cambridge Philosophical Society*, **24 (2)**, 180 – 190.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, **44 (3)**, 423 – 453.

Koutsoyiannis, D. (2004). Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation. *Hydrological Sciences Journal*, **49(4)**, – 590.

Marani, M. and Ignaccolo, M. (2015). A metastatistical approach to rainfall extremes. *Advances in Water Resources* **79**, 121 – 126.

Stolf, F. and Canale, A. (2022). A hierarchical Bayesian non-asymptotic extreme value model for spatial data. *arXiv:2205.01499.*

Zorzetto, E., Canale, A., and Marani, M. (2020). Bayesian non-asymptotic extreme value models for environmental data. *arXiv:2005.12101.*

A Model for Alcohol Consumption Trajectories

J. ten Dam[1], A.J. Rodenburg[1], K. Katona[1], A. van Giessen[1]

[1] National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands.

E-mail for correspondence: jasper.ten.dam@rivm.nl

Abstract: In the decision making process for policies in public health, models are becoming increasingly important to estimate the impact of policies on prevalences of future risk factors and related diseases. This study presents a modelling approach for both people's drinking status (never drinker, drinker or ex-drinker) as well as the amount of alcohol consumed by drinkers. Cross-sectional data was used in order to match the population distribution of the Netherlands. In addition, longitudinal data was used to capture individual trajectories, i.e., the evolution of any individual's drinking status over time.

Keywords: Alcohol; Health Modelling; Statistical Model

1 Introduction

The use of alcohol is a major risk factor for disease burden and causes substantial health loss (Griswold. Max G et al. 2016). To prevent harmful effects of alcohol use on health, countries are enrolling control policies, such as minimum unit pricing (Beeston 2020). In the decision making process for such prevention policies, models are an important support tool to evaluate the expected impact of an intervention or policy measure. A model that predicts trajectories of alcohol consumption is therefore useful to (1) predict the future prevalence of alcohol use without (extra) interventions and (2) to estimate the impact of interventions.
This study presents a model of alcohol use based on historical data. As such, it predicts future alcohol consumption under 'constant policy'. The aim is to later extend the model to include the effect of interventions.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Model Description

The model consists of two parts. Firstly, it models the *drinking status* of any person in the Dutch population: the person either has never drunk alcohol, drinks alcohol, or is an ex-drinker. Secondly, for any individual who drinks alcohol it models *how many* glasses of alcohol (s)he drinks on average per week.
The model parameters are fitted on a combination of a large cross-sectional dataset and a smaller longitudinal dataset. This ensures that the model captures both the population distribution in the Netherlands and individual trajectories of alcohol use. An individual trajectory here refers to the evolution of an individual's drinking status and alcohol consumption. We describe both fitting procedures one by one.

3 Population Level Distribution

The Public Health Monitor dataset is a large cross-sectional dataset that is representative of the Netherlands. It is based on a health-related questionnaire with more than 400,000 respondents (Community Health Services, Statistics Netherlands & RIVM, (2016)). We use the Public Health Monitor to determine the model parameters that describe the population distribution of drinking status and of the amount of alcohol consumption. In this dataset, being a drinker is defined as having drunk alcohol in the past 12 months. Figure 1 shows how the average number of glasses consumed per week by drinkers is distributed in the Public Health monitor dataset. Here, abstainers are removed from the data.
To determine the population distribution of drinking status, we fitted a multinomial model. The determinants age (continuous), sex (male, female) and level of education (low, middle, high) were used as predictors, as is common in public health models. The relationship between age and drink status was modelled through a cubic spline with 5 knots, and interactions between sex and education, age and sex, and age and education were included. Figure 2 shows the modelled proportion of drinkers together with the data, over age and stratified by sex and education level.
Next, we modelled the number of glasses that *alcohol drinking persons* drink on average per week based on the cross-sectional sample. Subsetting the data by omitting abstainers and ex-drinkers (blue in Figure 1) corresponds to a hurdle approach for dealing with zero-inflation. A negative binomial distribution was fitted to the remaining data (red in Figure 1) for drinkers, with age, sex and level of education as predictors. The relationship between age and the number of glasses was modelled through a cubic spline with 5 knots, and interactions between sex and education, age and sex, and age and education were again included. Figure 3 shows the modelled number of glasses for drinkers together with the data.

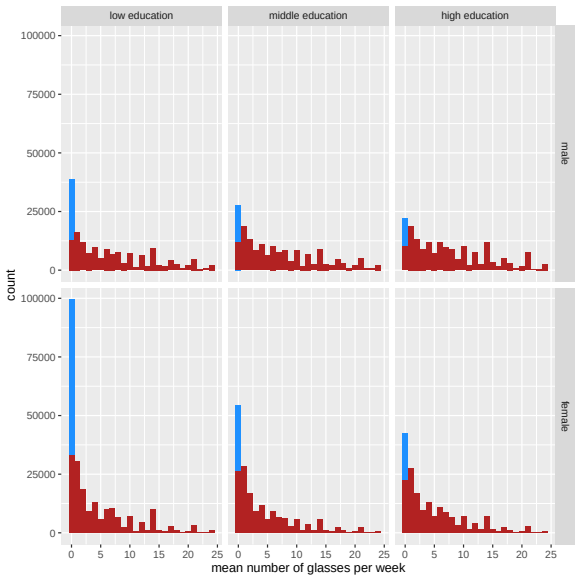


FIGURE 1. Histogram of the average consumed number of glasses of alcohol per week in the Public Health Monitor dataset, by sex and level of education (truncated to <25 glasses). Counts of zero's for abstainers and ex-drinkers are indicated in blue, counts for drinkers are indicated in red.

4 Transition Probabilities

In order to describe the evolution of an indivual's drinking status, we determine transition probabilities - or rather, transition *rates* - between the three drinking states. Describing the evolution of an individual's alchol *consumption* is also part of the study, but this is still work in progress. We use the Doetinchem Cohort Study to illustrate the transition probability model. The Doetinchem Cohort Study is a longitudinal study that followed a sample of individuals from the municipality of Doetinchem in the Netherlands for the past 30 years, in time intervals of approximately 5 years (Verschuren et al. 2008). A drawback of the Doetinchem cohort study in this approach is that the closed cohort contains no young individuals for recent years, and in general little young persons. The answers to the question "Do you drink alcohol?" in the questionnaire for participants were used to determine the drinking status of any participant during any measurement. Transitions from drinker to never-drinker were present in data and here, the end-point never-drinker was replaced by ex-drinker. We illustrate how we determine the rate to stop drinking - the other transition rates follow in a similar fashion. First, we list all *transitions* in the dataset. Any two successive drinking states of the same person form a

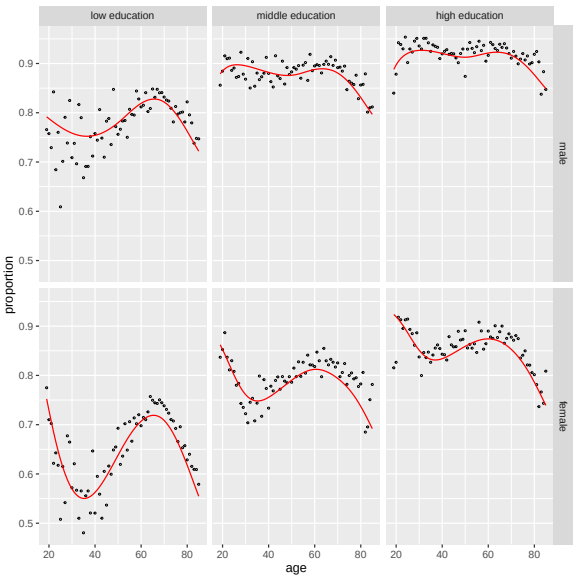


FIGURE 2. Proportion of drinkers over age, according to the model (red line) and according to the data (black points), by sex and level of education.

transition. Next, we select all transitions that have as initial drinking status 'drinker'. For every such transition i we define the variable $stops_i$ to indicate whether the person stopped drinking during the transition ($stops_i = 1$) or not ($stops_i = 0$). This variable is binomially distributed, $stops_i \sim B(n = 1, p_i)$, where p_i is the probability to stop drinking during the transition. We assume that - just like in survival analysis - p_i depends on the 'stopping rate' r_i and on the length of the time interval $(\Delta t)_i$ as

$$p_i = 1 - \exp(-r_i(\Delta t)_i). \tag{1}$$

Subsequently, we assume a linear dependence of the log-rate on age age_i, age^2 age_i^2, sex sex_i, education level edu_i and calender time t_i, i.e.,

$$\log(r_i) = b_1 \cdot age_i + b_2 \cdot age_i^2 + \sum_k \delta_{sex_i,k} b_{3k} + \sum_l \delta_{edu_i,l} b_{4l} + b_5 \cdot t_i, \tag{2}$$

where the first summation is over male and female, the second summation is over all education levels, and $\delta_{k,l}$ denotes the Kronecker delta. The coefficients b_1, b_{2k}, b_{3l} and b_4 follow by fitting a binomial regression model with the complementary log-log link function (and with $\log(\Delta t)_i$ as offset). Figure 4 shows the proportion of drinkers that stop drinking within a time interval of 5 years, both according to the model and according to the data.

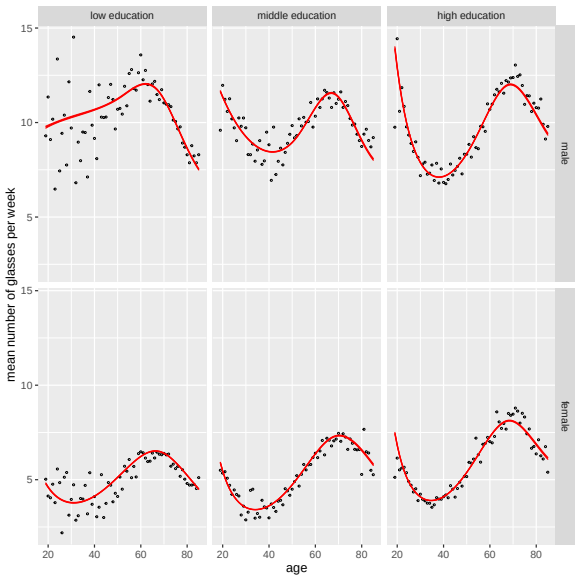


FIGURE 3. Mean number of glasses of alcohol per week of drinkers, according to the model (red line) and according to the data (black points), by sex and level of education.

5 Discussion and Conclusion

We designed and fitted complementary models to describe alcohol consumption in The Netherlands using age, sex and level of education as predictor variables. A cross-sectional dataset was used to describe alcohol consumption on a population level and transition probabilities were estimated based on a longitudinal dataset to describe the evolution of an individual's drinking status. The transition model and model for the number of glasses together form a hurdle modelling approach that is suitable for the zero-inflated count data. Describing the evolution of the number of glasses within a person who drinks is still work in progress. Other future research includes calibrating the transition probabilities and the time trend from the longitudinal data using cross-sectional data. Next, we used the Doetinchem dataset to illustrate our modelling approach while using a longitudinal dataset which includes young persons is recommended to better estimate transition probabilities for this group. Concluding, this approach is able to predict alcohol trajectories of a population, whilst providing the possibility to encoperate interventions. The approach can be used to model other risk factors as well, such as smoking.

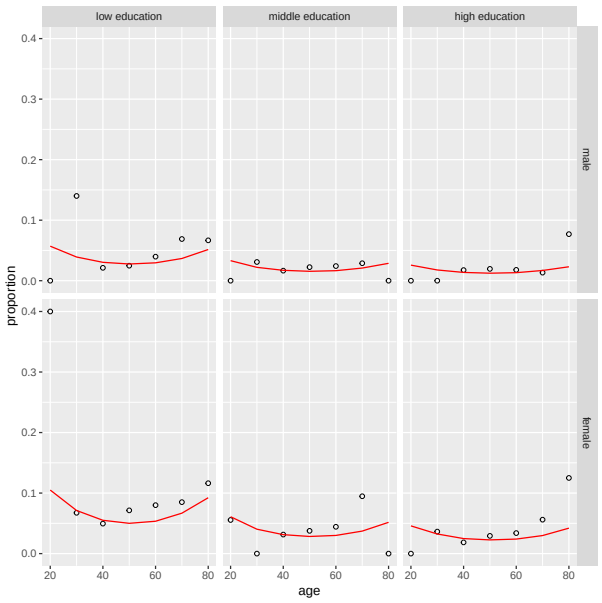


FIGURE 4. Share of people who stop drinking during a time interval of five years, by sex and level of education.

Acknowledgments: We would like to thank Susan Picavet for her help in supplying and interpretating the Doetinchem dataset and Noah Cnossen for her help on exploratory research.

References

Beeston, C., Robinson, M., Giles, L., Dickie, E., Ford, . . . Craig, N. (2020). *Evaluation of Minimum Unit Pricing of Alcohol: A Mixed Method Natural Experiment in Scotland.* International journal of environmental research and public health, Volume 17, Issue 10, 3394

Community Health Services, Statistics Netherlands, & RIVM *Public Health Monitor 2016..*

Griswold, Max G et al. (2016). *Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016.* The Lancet, Volume 392, Issue 10152, 1015 - 1035

Verschuren, W., Blokstra, A., Picavet, H., & Smit, H. (2008). *Cohort Profile: The Doetinchem Cohort Study.* International Journal of Epidemiology, Volume 37, Issue 6, 1236-1241

Model based clustering of households from the EU-SILC database

Jan Vávra[1], Arnošt Komárek[1]

[1] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

E-mail for correspondence: komarek@karlin.mff.cuni.cz

Abstract: In another IWSM contribution (Vávra and Komárek, 2022), we present the model-based clustering (MBC) approach towards segmentation of units based on multivariate mixed type longitudinal data. The method is based on a Bayesian approach and a multivariate variant of Generalized Linear Mixed Model (GLMM). In this paper, we use it to analyze the Czech subset of the European Union Statistics on Income and Living Conditions database (EU-SILC) to identify poverty and social exclusion temporal patterns of Czech households.

Keywords: EU-SILC; GLMM; Model-based clustering; Multivariate longitudinal data.

1 EU-SILC database and research problem

In 2003, core member states of European Union agreed to launch an instrument aiming at collecting timely and comparable cross-sectional and longitudinal multidimensional micro-data on income, poverty, social exclusion and living conditions leading to the EU-SILC project. In our analysis, we focus on the Czech subset of the longitudinal part of data gathered annually with the aim to identify the incidence and dynamic processes of the persistence of poverty and social exclusion among subgroups in population. The changes are followed up only for a limited duration – a period of four years. This is induced by a rotational panel, where each year a quarter of households is replaced by a set of newly observed households of comparable size. Since 2005, data from $n = 23\,360$ Czech households observed exactly for $n_i = 4$ consecutive years have been gathered until 2018.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

2 Model structure

To identify poverty and social exclusion temporal patterns in households economics, we exploit methodology proposed by Vávra and Komárek (2022), further referred to as [VK]. Therein a mixture of generalized linear mixed-effects models (GLMM) is assumed over a chosen set of outcomes of diverse nature. The goal of our analyzis of the EU-SILC database is to identify different groups of households with homogeneous evolution patterns. One of the reasons why we look for different groups within the data may be the economical crisis that has struck in 2008 and could have had significant impact on the prosperity of households. Increasing, stagnating or even decreasing trends in time (and their different combinations with respect to different outcomes) are expected to be discovered.
We will start first by introducing the chosen outcomes and the covariates that may help in explaining the outcomes behavior. Only household-level variables are used for this analysis to avoid nested effects; complex variables (income, education, ...) are originally measured at the personal level and then aggregated to construct household-level variables.

2.1 Outcomes of interest

First, we list outcomes of interest being primarily used for clustering the households, while distinguishing their several types.

- Numeric outcomes (modeled on logarithmic scale)
  - `HX090` – *Equivalised total disposable income* [EUR/year]
    The sum of gross personal income components of all household members divided by the *Equivalised household size*
  - `HS130` – *Lowest income to make ends meet* [EUR/month]
    The very lowest net monthly household income required to pay for the usual necessary expenses
- Binary outcomes (Yes / No)
  - `HS040` – *Affordability of a one week holiday*
    Capacity to afford paying for a one week annual holiday away from home
  - `HS060` – *Afford to pay unexpected expenses*
    Capacity to face unexpected financial expenses
- Ordinal outcomes (self-evaluation by the respondent)
  - `HS120` – *Ability to make ends meet*
    with great difficulty (1) $< \cdots <$ very easily (6)
  - `HS140` – *Financial burden of the total housing cost*
    a heavy burden $<$ a slight burden $<$ not a burden at all
- Categorical outcomes (Yes / No – cannot afford / No – other reason)
  - `HS090` – *Do you have a computer?*
  - `HS110` – *Do you have a car?*

2.2 Other covariates

The dataset also offers plenty of potential regressors to be used in the GLMM's underlying the households segmentation:

- *Time* will be considered as the most important and will be used for distinguishing the groups. We define the time covariate as the number of years past the beginning of 2005, which limits the time into the interval $[0, 14)$. Note that the interviews in the Czech republic were held in either Q1 or Q2.
- *Equivalised household size* expresses how large the household is while taking the age of its members into consideration. The head of the household (respondent) has a unit weight, while other members have either 0.5 (older than 14) or 0.3 (younger than 14).
- *Level of urbanisation* was divided by the population density and minimum population into the following categories:
  1. thinly-populated area (non-urban),
  2. intermediate area (at least 300 inhabitants per km^2, minimal population of 5 000),
  3. densely populated area (at least 1500 inhabitants per km^2, minimal population of 50 000),
  4. Prague (the highly-populated capital city).
- *The highest ISCED (education) level achieved* within the whole household rarely attains the lowest possible option of primary education. Hence, we merge it with lower-secondary education. Then, follows the most common upper-secondary education. Finally, the third category contains both post-secondary and the tertiary education level (a university degree).
- *Presence of student* or *baby* indicate whether some household member currently attends any educational institution or is younger than 3 years, respectively.

2.3 The model settings

In total, we deal with eight outcome variables for which a mixture of multivariate GLMM's is specified as outlined in Vávra and Komárek (2022). The evolution in time is captured by a quadratic spline parametrization and this is also a group-specific part of the model to capture possibly different patterns in data. Remaining covariates are included additively in the model formula without any interaction. Their effects are estimated to be common to all households regardless of the cluster allocation. The random effects part is formed solely by the random intercept term; not only do the random effects capture the specific level of a household, but they channel

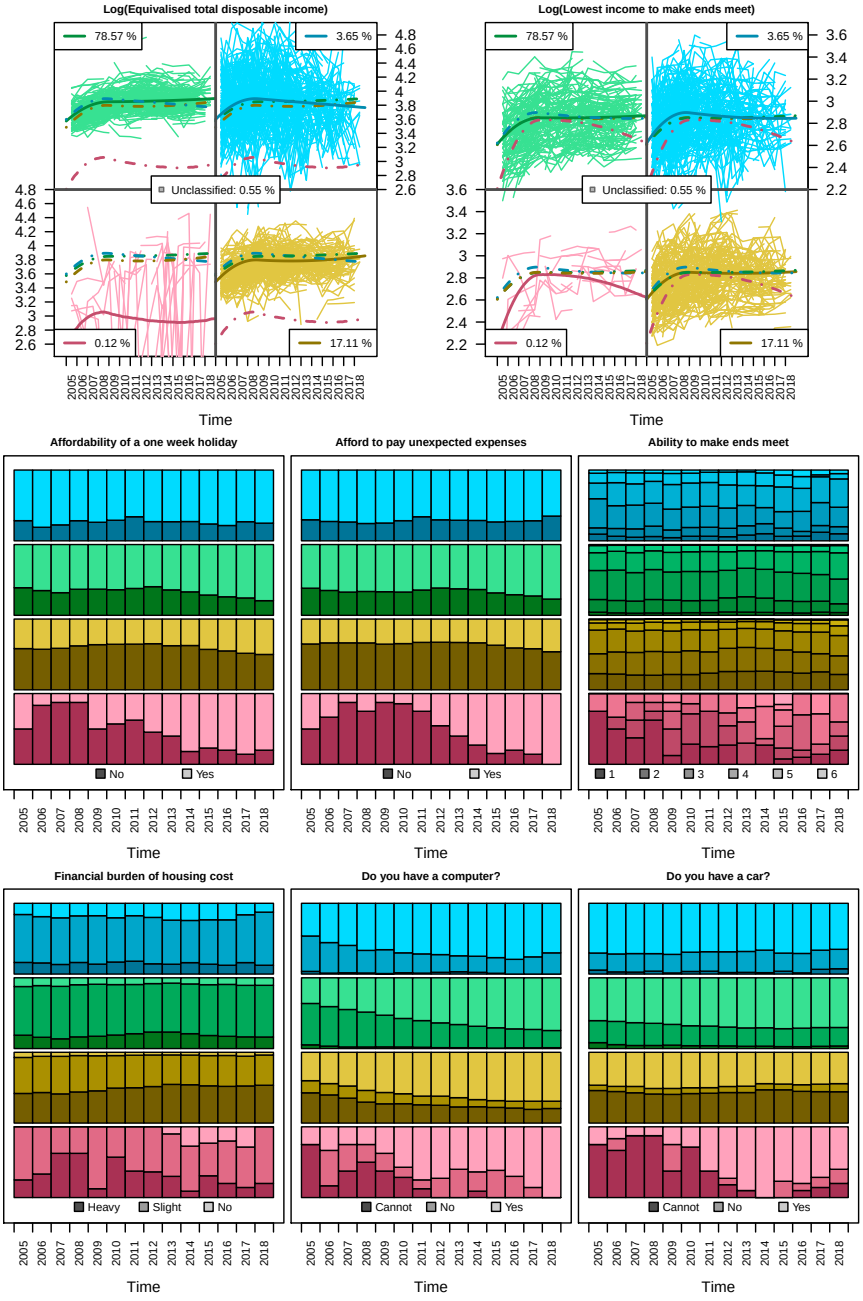


FIGURE 1. Evolution in time of all eight outcomes in discovered clusters.

the marginal associations among outcomes through a common covariance matrix ($\boldsymbol{\Sigma}$ in [VK]).
These correlations among outcomes are assumed to be the same across all clusters, hence $\boldsymbol{\Sigma}$ is common to all clusters. The precision parameters for the numeric outcomes ($\boldsymbol{\tau}$ in [VK]) and ordered intercept terms for the ordinal outcomes ($\mathbf{c}$ in [VK]) are both set to be cluster-specific, for the former can account for different variability, while the latter accounts for different distribution of ordinal levels (intercepts are cluster-specific for all outcomes anyway).
To estimate the suitable number of clusters we adopted the sparse finite mixture approach used by Frühwirth-Schnatter and Malsiner Walli (2019). Hence, the maximal number of underlying clusters has been set to $G_{\mathsf{max}} = 20$ and sparsity was induced by a Dirichlet prior favouring emptying the mixture components.

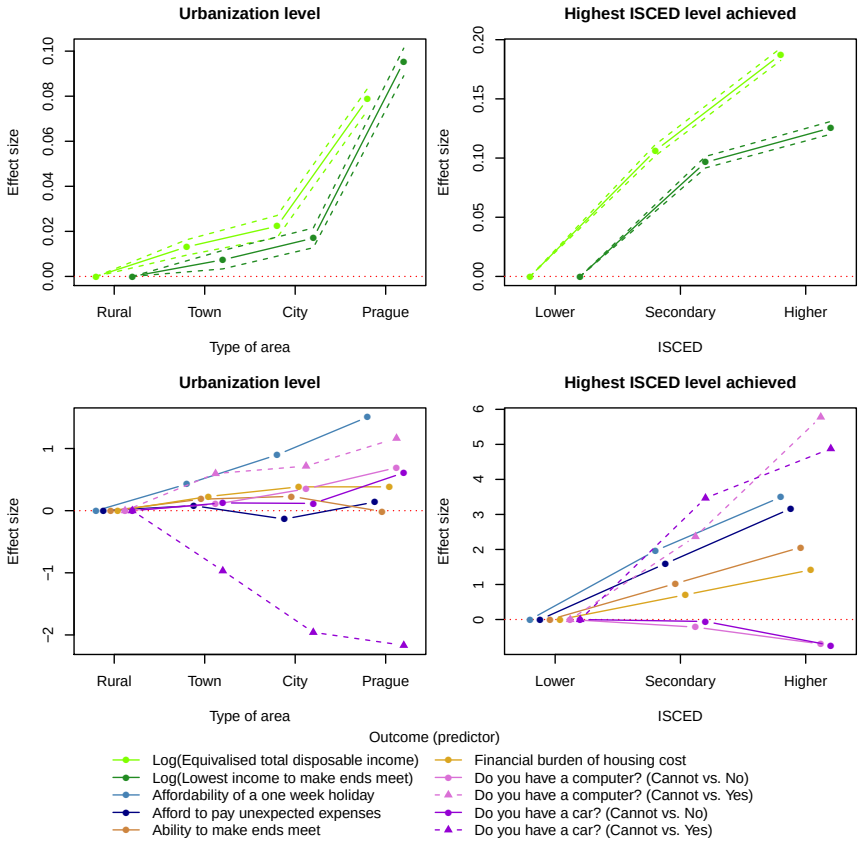


FIGURE 2. Estimated effects of additional covariates on categorical outcomes.

3 Results

Under settings above, the sampled Markov chains converged to a four-cluster solution. Households were assigned to the cluster of the highest frequency of sampled allocation indicators provided it was higher than 50%, otherwise the household remained unclassified (0.55% of households). The clusters will be referred to by the assigned colours in Figure 1.
First, the irregular curves of the red cluster (0.12%) have to be addressed. These households have experienced unusually low values of the *Equivalised total disposable income* and, hence, should be considered as outliers. The remaining three clusters: the blue (3.65%), the yellow (17.11%) and the green (78.57%) share very similar shape evolution curves of both types of income - an increasing trend till 2009, when the stagnation (or even a slight decrease) begins. Nevertheless, it is obvious that it is rather the inner variance (controlled by $\boldsymbol{\tau}$) than the trend that distinguishes the discovered groups.
However, looking at the evolution of proportions of levels of categorical outcomes the clusters acquire yet another interpretation. The thin blue cluster, which experiences the highest volatility of the income, achieves the highest proportions of positive categories, hence could be considered as the cluster of wealthy households. On the other hand, the yellow cluster of medium income volatility has higher proportions of negative categories. The green cluster represents the vast majority of households of steady income evolution and proportions slightly worse than the wealthy cluster.
Figure 2 presents the estimated effects of the urbanization level and the highest achieved ISCED level. Clearly the capital city of Prague exhibits the highest effect compared to the rural area. The probability of possession of a car is the only one of decreasing trend with increasing population density. As expected, the higher education level achieved within the household the higher income is expected, which also relates to the increment in predictors for categorical outcomes.

Acknowledgments: This research was supported by the Czech Science Foundation (GAČR) grant 19-00015S. The first author was additionally supported by Charles University, project GA UK No. 298120.

References

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, **13**(1), 33–64.

Vávra, J. and Komárek A. (2022) GLMM Based Clustering of Multivariate Mixed Type Longitudinal Data. *Proceedings of IWSM 2022*.

Nearest Neighbours Gaussian Process Model for Time-Frequency Data: An Application in Bio-acoustic Analysis

Hiu Ching Yip[1], Gianluca Mastrantonio[1], Enrico Bibbona[1], Marco Gamba[2], Daria Valente[2]

[1] Politecnico Di Torino, Italy
[2] Universita degli Studi di Torino, Italy

E-mail for correspondence: hiu.yip@polito.it

Abstract: Bio-acoustic signal analysis often reduces to feature analysis on the frequency structure in a lower dimensional space. This approach usually treats the time-frequency bins of spectrograms as independent features or extracts common statistics from waveforms. It is known to entail human perceptual bias that is induced by the neglect of the relative relationship between the spectral shape of vocalization and time as well as the dependence on domain knowledge of animals' behaviours. In light of this, we propose a Nearest Neighbour Gaussian Process (NNGP) model to account for the time varying components in the latent spectral structure of bio-acoustic data.

Keywords: NNGP; Time-frequency data; Latent spectral structure; Time-varying effects; Bio-acoustics

1 Motivation & Data

In comparative bio-acoustic studies, one area of interest is to understand the acoustic structures of non-human primates in order to provide insights on the evolution of the communication mechanism of our closest relatives. The most common practices are feature engineering methods, which involves selecting a set of basis-features for quantitative comparison. The identification of meaningful features in the vocal repertoire relies on biologists to observe and interpret the behavioural contexts in which the animals emit the signals. These interpretations are costly to acquire, inaccurate due to human subjectivity and difficult to generalize for cross-species comparison. Furthermore, feature selection always ignores the time-varying effect

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

of observed vocalizations on the latent acoustic structure. The aim of this project is to propose a NNGP model that accounts for time in bio-acoustic analysis. The dataset that will be available for model implementation are vocal signals of lemurs that were recorded in Madagascar.

The data format is equivalent to the published data in (Valente, D. et al. 2019). Each recorded signal is represented by a spectrogram and lasts for a unique duration of time that is measured in seconds. We refer to Figure 1 for a time-frequency representation of 3 signals of different durations. Furthermore, each signal is categorized by a call-type label and a species label, which are characterized by the behaviour of the lemur during emission and the species to which the lemur belongs to, respectively. As an example, Table 1 lists the number of recorded signals of 3 different species/call-type groups of signals. The group labels are given by biologists.

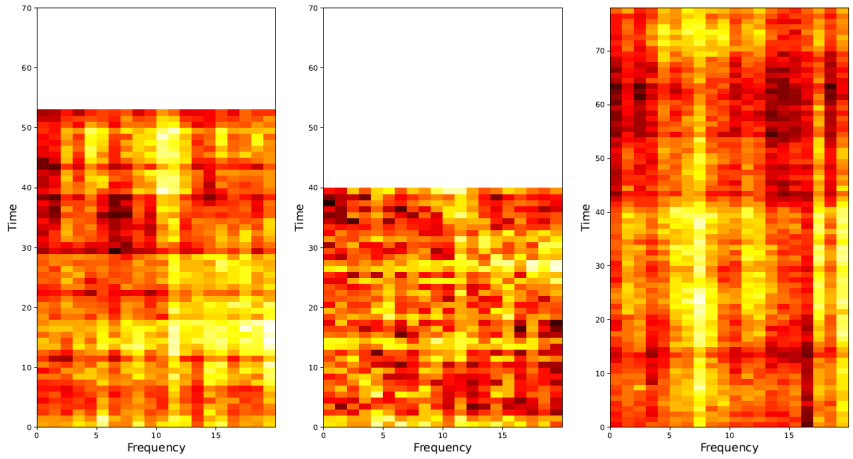


FIGURE 1. Spectrograms of 3 observed signals with different durations

TABLE 1. Number of recorded signals of 3 different species/call-type categories

species	**call-type**	**# signals**
Indri indri (II)	Clacson (CL)	622
Indri indri (II)	Grunt (GR)	1145
Indri indri (II)	Hum (HU)	418

2 NNGP Hierarchical Model

Let G denote the total number of observed signals of one species/call-type category. Denote each g-th observed signal by $\boldsymbol{z}_g$ where $g \in \{1, 2, ..., G\}$. Let t_g^* be its respective duration. Each $\boldsymbol{z}_g$ is a Gaussian field : $\boldsymbol{z}_g = \{ z_g(t,h) : (t,h) \in \mathcal{U}_g \}$ where (t,h) is a location in the observed spatial domain $\mathcal{U}_g = \mathcal{T}_g \times \mathcal{H}_g$, $\mathcal{T}_g = [0, t_g^*]$ is the time-axis and $\mathcal{H}_g$ is the frequency-axis. Each domain $\mathcal{U}_g$ is unique. Let $\boldsymbol{w}_\mathcal{R} = \{ w(t,h) : (t,h) \in \mathcal{R} \}$ be a latent Gaussian field of zero mean over $\mathcal{R} = \mathcal{T} \times \mathcal{H} : \mathcal{T} = [0,1]$ that needs to be inferred from the data $\boldsymbol{z}_g$. This latent field $\boldsymbol{w}_\mathcal{R}$ is the inherent acoustic structure of a given set of signals of the same species/call-type category that has factored in the effects of each unique $\mathcal{U}_g$. The model is :

$$\begin{aligned} z_g(t,h) &= \mu_g + y_g(t,h) + \epsilon_g(t,h) \\ &= \mu_g + w(\alpha_g + \beta_g\ t, h) + \epsilon_g(t,h) \end{aligned}$$

where $y_g(t,h) = w(\alpha_g + \beta_g\ t, h)$ is a point value evaluated at the location $(\alpha_g + \beta_g\ t, h) \in \mathcal{R}$; $\alpha_g,\ \beta_g$ are the time-distortion parameters $i.i.d.\ \forall\ g$: $\alpha_g + \beta_g\ t \in \mathcal{T} = [0,1]\ \forall\ t \in \mathcal{T}_g = [0, t_g^*]$; μ_g is the scalar mean $i.i.d\ \forall\ g$; and; $\epsilon_g(t,h) \sim N(0, \tau_g^2)$ is the random noise $i.i.d.\ \forall\ g, t, h$.

Let $C(\cdot)$ be the covariance kernel of that is specified by :

$$\begin{aligned} C((t_1,h_1),(t_2,h_2)) = \sigma^2 e^{-(\ \psi_t|t_1-t_2|\ +\ \psi_h|h_1-h_2|\ )} \\ +\ \sigma_c^2 e^{-\psi_c d_c(t_1,t_2,\gamma)} \end{aligned}$$

$\forall\ (t_1,h_1),(t_2,h_2) \in \mathcal{R}$. The first component of $C(\cdot)$ describes how the acoustic structure changes across the time-frequency grid $\mathcal{R}$. The second component addresses the circular nature of time-frequency data. The distance function $d_c(t_1,t_2,\gamma)$ is the periodic distance between two time points on $\mathcal{T}$ such that $d_c(t_1,t_2,\gamma) \in [0, \gamma/2]\ \forall\ t_1, t_2 \in \mathcal{T}$. The parameters of $C(\cdot)$ that need to be inferred are the time and frequency decay : $\psi_t,\ \psi_h$; the periodicity and its decay : $\gamma,\ \psi_c$; and ; the variances : $\sigma,\ \sigma_c$. Write $\boldsymbol{\Sigma}$ as the exact covariance matrix given by the kernel $C(\cdot)$ and $\boldsymbol{\theta} = \{\psi_t,\ \psi_h,\ \psi_c,\ \gamma,\ \sigma,\ \sigma_c\}$. The hierarchical model is :

$$\begin{aligned} \boldsymbol{z}_g \mid \mu_g,\ \boldsymbol{y}_g,\ \tau_g^2 &\sim \text{GP}(\ \mu_g + \boldsymbol{y}_g\ ,\ \tau_g^2\ ) \\ \boldsymbol{y}_g \mid \boldsymbol{\theta},\ \alpha_g,\ \beta_g &\sim \text{GP}(0, \boldsymbol{\Sigma}) \\ \boldsymbol{w}_\mathcal{R} \mid \boldsymbol{\theta} &\sim \text{GP}(0, \boldsymbol{\Sigma}) \end{aligned}$$

We refer to Figure 2 for a graphical representation of the relationship between $\boldsymbol{w}_\mathcal{R},\ \boldsymbol{y}_g$ and $\boldsymbol{z}_g$. The relative relationship between the times given by data and the spectral shape of $\boldsymbol{w}_\mathcal{R}$ is described by the time-distortion

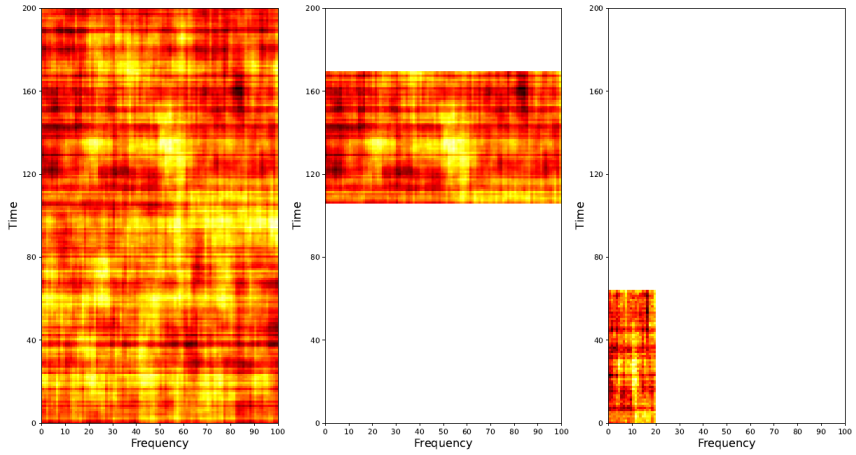


FIGURE 2. (From left to right) Spectrogram of $\boldsymbol{w}_{\mathcal{R}}$, $\boldsymbol{y}_g$ and $\boldsymbol{z}_g$ respectively

parameters : $\alpha_g,\ \beta_g$, which map the data points in $\mathcal{U}_g$ onto the latent domain $\mathcal{R}$. Since $\boldsymbol{y}_g$ evaluates spatial locations in $\mathcal{R}$, it is thus specified by the same distribution of $\boldsymbol{w}_{\mathcal{R}}$. The data $\boldsymbol{z}_g$ can be marginalized over $\boldsymbol{y}_g = \{\ w(\alpha_g + \beta_g\ t, h)\ \ :\ \ (t,h) \in \mathcal{U}_g\ \}$. The marginal distribution of $\boldsymbol{z}_g$ over $\boldsymbol{y}_g$ is completely specified by the scalar mean μ_g, the noise τ_g^2, the time-distortion parameters $\alpha_g,\ \beta_g$ and the kernel parameters $\boldsymbol{\theta}$. Let k_g be the number of data points $z_g(t,h) \in \boldsymbol{z}_g$, the g-th observation. Define D_g as the diagonal matrix of dimension $k_g \times k_g$ with τ_g^2 as the diagonal entries. The marginal distribution is :

$$\begin{matrix}\boldsymbol{z}_1 \\ \boldsymbol{z}_2 \\ \vdots \\ \boldsymbol{z}_G\end{matrix} \sim \mathrm{GP}\left(\begin{matrix}\mu_1 \\ \mu_2 \\ \vdots \\ \mu_G\end{matrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{1,1} + D_1 & \boldsymbol{\Sigma}_{1,2} & \cdots & \boldsymbol{\Sigma}_{1,G} \\ \boldsymbol{\Sigma}_{2,1} & \boldsymbol{\Sigma}_{2,2} + D_2 & \cdots & \boldsymbol{\Sigma}_{2,G} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{G,1} & \boldsymbol{\Sigma}_{G,2} & \cdots & \boldsymbol{\Sigma}_{G,G} + D_G \end{bmatrix} \right)$$

Since inverting the high-dimensional exact covariance matrix $\boldsymbol{\Sigma}$ is too computationally expensive, we resort to the approximated NNGP instead of the exact GP. The idea of NNGP is that for Gaussian processes, if the covariance kernel is monotonic with respect to the distance between two spatial points, then only the data at neighbouring locations is needed for inference. Define the neighbour set $\mathcal{N}(t,h)$ as the set of m points that are "closest" to the point (t,h) such that the points in $\mathcal{N}(t,h)$ have the maximum correlation with point (t,h) given by $C(\cdot)$.

Let n denotes the total number of points in $w_{\mathcal{R}}$ and $(t_i, h_i) \in \mathcal{R}\ \forall\ i =$

$1, 2, ..., n$. The density of $w_{\mathcal{R}}$ that is expressed in terms of the full conditional densities can then be approximated in terms of the neighbour sets $\mathcal{N}(t, h)$. Write $\boldsymbol{z} = \{ \boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_g \}$. The marginal distribution of $\boldsymbol{z}$ specified above can also be approximated similarly.

$$\begin{aligned}
\mathbb{P}(\boldsymbol{w}_{\mathcal{R}}) = \; & w(t_1, h_1) \prod_{i=2}^{n} \mathbb{P}\big(\, w(t_i, h_i) \mid \{w(t_j, h_j) \; : \; t_j \leq t_i \; , \; h_j \leq h_i\} \, \big) \\
\approx \; & w(t_1, h_1) \prod_{i=2}^{n} \mathbb{P}\big(\, w(t_i, h_i) \mid \boldsymbol{w}_{\mathcal{N}(t_i, h_i)} \, \big) \\
\mathbb{P}(\boldsymbol{z}) = \; & z_1(t_1, h_1) \prod_{g=1}^{G} \prod_{i=1}^{k_g} \mathbb{P}\big(\, z_g(t_i, h_i) \mid \{z_{g'}(t_j, h_j) \; : \; g' \leq g \; , \; \alpha_{g'} + \beta_{g'} \, t_j \leq \alpha_g + \beta_g \, t_i \; , \; h_j \leq h_i\} \, \big) \\
\approx \; & z_1(t_1, h_1) \prod_{g=1}^{G} \prod_{i=1}^{k_g} \mathbb{P}\big(\, z_g(t_i, h_i) \mid \boldsymbol{z}_{\mathcal{N}(\alpha_g + \beta_g t_i, h_i)} \, \big)
\end{aligned}$$

References

Datta, A. et al. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *J Am Stat Assoc*, **111(514)**, 800-812.

Sainburg, T., Thielk, M., and Gentner, T.Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PloS Comput Biol*, **16(10)**, e1008228.

Valente, D. et al. (2019). Finding Meanings in Low Dimensional Structures: Stochastic Neighbor Embedding Applied to the Analysis of Indri indri Vocal Repertoire. *Animals : an open access journal from MDPI*, **9(5)**, 243.

Bayesian multiscale mixtures of multivariate Gaussian kernels for density estimation

Daniele Zago[1], Antonio Canale[1], Marco Stefanucci[2]

[1] Dipartimento di Scienze Statistiche, Università di Padova
[2] Dipartimento di Scienze Statistiche, Sapienza Università di Roma

E-mail for correspondence: daniele.zago.1@phd.unipd.it

Abstract: In this article we discuss some preliminary results related to a multivariate extension of the Bayesian nonparametric multiscale model introduced in Stefanucci and Canale (2021). This model provides a flexible alternative to classical Bayesian nonparametric methods for density estimation in $\mathbb{R}^d$ and can be applied to both continuous and mixed data through the specification of appropriate kernels. By construction, the proposed method is able to model densities characterized by either smoothness or localized abrupt changes. Posterior inference is possible via a Gibbs sampler.

Keywords: Bayesian nonparametrics; Multiscale models; Stick-breaking.

1 Introduction

Nonparametric density estimation in the Bayesian setting is dominated by single-scale methods such as the Dirichlet process mixture of kernels Escobar and West (1995). However, in many cases these methods fail to adequately represent the underlying true density. One such case is when the density displays a high degree of variation in its smoothness, for instance when a broad density shows several local abrupt changes. In these settings it would be beneficial to adopt a multiscale approach, in which the density is naturally modeled with an increasing degree of resolution. One such proposal is the multiscale mixture of kernels introduced in Stefanucci and Canale (2021), which offers a flexible alternative to the smoothed Pólya tree model (Cipolli and Hanson, 2017). However, the application of the multiscale mixture of kernels is bounded to univariate densities on $\mathcal{X} \subseteq \mathbb{R}$, and it is not clear how a generalization to the multivariate setting might be derived. Inspired by that construction, we show some preliminary re-

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

sults related to the application of the multiscale mixture of kernels in the multivariate context.

2 Multivariate mixture of kernels

Assume that $Y \in \mathcal{Y} \subseteq \mathbb{R}^d$ is a d-dimensional random variable, and that we can represent its density f using a multiscale structure,

$$f(y) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} \text{MVN}_d(y; \mu_{s,h}, \Sigma_{s,h}), \tag{1}$$

where $\text{MVN}_d(\cdot; \mu, \Sigma)$ denotes the multivariate Gaussian density with mean $\mu \in \Theta_\mu = \mathbb{R}^d$ and variance-covariance matrix $\Sigma \in \Theta_\Sigma = S^+_{d\times d}$, where $S^+_{d\times d}$ is the space of positive-definite symmetric matrices of dimension $d \times d$. $\{\pi_{s,h}\}$ is a sequence of random weights such that $\sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} = 1$. Since we want the multiscale model to adapt to the smoothness of the data, we place a prior distribution on the parameters in (1). The prior process for the weights $\{\pi_{s,h}\}$ is described in Canale and Dunson (2016) and Stefanucci and Canale (2021), and it is summarized below. For each s and h, we define a pair of random variables

$$S_{s,h} \sim \text{Beta}\big(1-\delta, \alpha + \delta(s+1)\big), \quad R_{s,h} \sim \text{Beta}(\beta, \beta), \tag{2}$$

which represent the probability of stopping on the (s,h)-th node and taking the right path at the (s,h)-th node of the binary tree, respectively. Then, with the sequence of auxiliary random variables defined in (2), it is possible to show that the weights $\{\pi_{s,h}\}$ are generated according to

$$\pi_{s,h} = S_{s,h} \prod_{r<s} \big(1 - S_{r,\lceil h2^{r-s} \rceil}\big) T_{shr}, \tag{3}$$

Using the above construction, it is possible to develop an efficient slice sampler for performing posterior inference conditionally on the cluster allocations via Markov Chain Monte Carlo methods, as shown in Stefanucci and Canale (2021). In order to complete the specification of the Gaussian multiscale model (1), we need to define a prior distribution for $\{\mu_{s,h}\}$ and $\Sigma_{s,h}$. We thus define a prior distribution on $\{\mu_{s,h}\}$ in such a way that the location parameter space $\Theta_\mu \subseteq \mathbb{R}^d$ is fully explored at each scale s of the binary tree. This entails splitting the parameter space Θ_μ at scale s into 2^s disjoint rectangles $\Theta_{\mu;s,h}$ and sample the location parameters from a multivariate Gaussian distribution truncated to each of these sets. In order to traverse the multivariate partition $\mathcal{P}_{s,h} = \{\Theta_{\mu;s,h} : s = 1, 2, \ldots, h = 1, \ldots, 2^s\}$ using a binary tree, we first employ the Hilbert curve (Hilbert, 1891) to index the 2^s centers of a partition $\tilde{\mathcal{P}}_{s,h}$ of $[0,1]^d$. Then, this partition is transformed into the required set of rectangles $\mathcal{P}_{s,h}$ at each scale s. With this

approach we are able to maintain the binary tree structure for the weights described in (3) and apply the multiscale model in the multivariate setting (1). The prior distribution on $\{\Sigma_{s,h}\}$ is specified so that deeper scales are associated on average to more concentrated kernels, which reflects the increase in locality of the modelled features as s increases. This property can be obtained by appropriately rescaling a fixed prior distribution on $\Sigma_{s,h}$ via a diagonal matrix $C(s) = \text{diag}\big(c_1(s), c_2(s), \ldots, c_d(s)\big)$ such that each $c_i(s)$ is decreasing as $s \to \infty$ for $i = 1, \ldots, d$. By choosing the fixed prior distribution as a conjugate prior, we can take advantage of an efficient Gibbs sampler for performing posterior inference on the model parameters, conditionally on cluster allocations.

3 Simulation

We apply the model under simulated data by setting $d = 2$ in order to demonstrate its performance both graphically and quantitatively. Specifically, for $i = 1, \ldots, 50$ we repeatedly generate a sample of size $n = 250$ from a mixture of normal and skew-normal distributions, which has been constructed in order to display a clear multiscale structure in its components. This can be seen from its contour plot in Figure 1, which also shows the average of the mean predictive densities for the multiscale model over the 50 simulations. We can see that the proposed multiscale approach can naturally adapt to different degrees of smoothness of the underlying density, and correctly represents both global and local features.

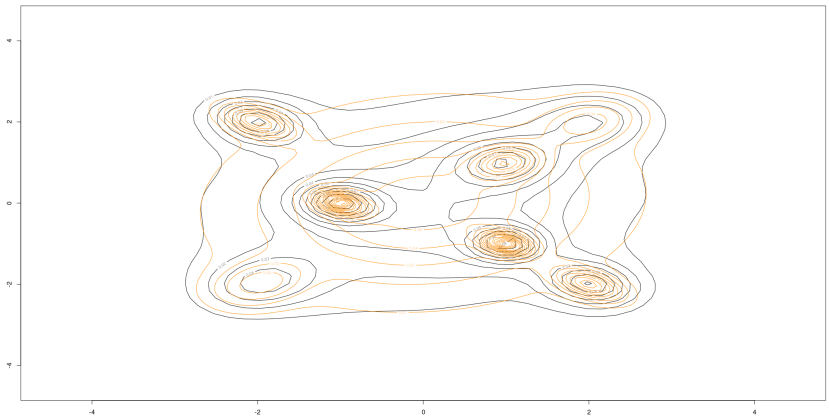


FIGURE 1. Contour plot of the true underlying density (*black*) and mean posterior predictive densities using the multiscale model (*orange*), averaged over 50 simulations.

We compare the mean posterior predictive density using our multiscale approach with the Pitman-Yor mixture of Gaussian kernels, using a generalization of the default hyperparameters recommended in Cipolli and Hanson (2017). A quantitative comparison of the two models can be done for instance by using the LPML criterion. Table 1 shows that the multiscale approach results in a higher average LPML over the 50 replications when compared to the Pitman-Yor mixture model. This suggests that the multiscale model is able to offer an advantage over single-scale methods when the underlying density shows concentrated local features, such as the one in Figure 1. In light of these preliminary results, we believe that an in-depth study of model performance using higher-dimensional simulated and real datasets is warranted.

TABLE 1. Average LPML and related standard error over 50 simulations of the data-generating process.

Model	ave(LPML)	sd(ave(LPML))
Multiscale	-3.34	0.11
Pitman-Yor	-3.58	0.10

References

Canale, A. and Dunson, D. B. (2016). Multiscale Bernstein Polynomials for Densities. In: Statistica Sinica 26.3

Cipolli, W. and Hanson, T. (2017). Computationally Tractable Approximate and Smoothed Polya Trees. In: Statistics and Computing 27.1.

Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. In: Journal of the American Statistical Association 90.430, pp. 577–588.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. In: Journal of the Royal Statistical Society. Series B 56.3, pp. 501–514.

Hilbert, D. (1891). Über die stetige Abbildung einer Line auf ein Flächenstück. In: Mathematische Annalen 38.3, pp. 459–460.

Pitman, J. and Yor, M. (1997). The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. In: The Annals of Probability 25.2, pp. 855–900

Stefanucci, M. and Canale, A. (2021). Multiscale Stick-Breaking Mixture Models. In: Statistics and Computing 31.2. ISSN : 1573-1375. DOI : 10.1007/ s11222-020-09991-1.

Simultaneous linear dimension reduction and clustering with flexible variance matrices

Yingjuan Zhang[1], Jochen Einbeck[1,2]

[1] Department of Mathematical Sciences, Durham University, UK
[2] Durham Research Methods Centre, UK

E-mail for correspondence: yingjuan.zhang@durham.ac.uk

Abstract: This paper revisits a modelling technique previously introduced as the 'generative linear mixture model'. An identifiability problem which was inherent to the original work is solved, and a simulation is conducted to test the accuracy of the methodology. The framework given in this paper allows for flexible variance specifications around the cluster (mixture) centres which are linearly spanning the data space. A real data application is provided, demonstrating how the resulting latent variable can be employed as an efficient competitor to one-dimensional principal component regression.

Keywords: Random effect; Mixture; Dimension reduction; EM algorithm.

1 Introduction

This paper complements previous work by Lawson and Einbeck (2012), in which random effect methodology was considered for the dimension reduction of high-dimensional data, $x_i \in \mathbb{R}^m$. This was achieved by projecting the original data onto the estimated low-dimensional latent space, $\alpha + \beta z$, where $\alpha, \beta \in \mathbb{R}^m$ and z is a one-dimensional random effect represented by a discrete mixture with mass points z_1,...,z_k and masses π_1,...,π_k, $k = 1, ..., K$. The observed data are assumed to be generated from the 'generative linear mixture model'

$$x_i = \alpha + \beta z_k + \varepsilon_i \tag{1}$$

where $\alpha + \beta z_k$ are the cluster centers on the straight line, and $\varepsilon_i \sim N(0, \Sigma)$ is the Gaussian noise added to the cluster centers. Under the original approach, the variance matrix $\Sigma \in \mathbb{R}^{m \times m}$ is assumed to be a diagonal matrix, $\mathrm{diag}(\sigma_j^2)_{\{1 \le j \le m\}}$ and to be the same for all K components of the mixture.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

The previous assumption on the variance disregards other geometric features that clusters might have, such as clusters with different sizes, shapes or orientations determined by the covariance. So, we consider several types of variance matrix parametrizations. First, we can use the same diagonal variance matrix for each component, as in Lawson and Einbeck (2012). Secondly, we can use different diagonal variance matrices for different components, $\Sigma_k \in \mathbb{R}^{m\times m}$, $\text{diag}(\sigma_{jk}^2)_{\{1\leq j\leq m\}}$, $k = 1, \ldots, K$. This type of variance yields an improvement for estimating data that has clusters of different sizes. Third, since the shape of clusters may not always be ball-shaped, we consider a full variance-covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$ to be the same for all components, $\Sigma = \Sigma_1 = \ldots = \Sigma_k$. Fourth, we consider different full variance-covariance matrices for different components, $\Sigma_k \in \mathbb{R}^{m\times m}$, $k = 1, \ldots, K$, which can be used on data that have clusters that differ by shape and size. The EM algorithm will be used to estimate the parameters mentioned above.

2 Methodology

2.1 EM algorithm

By using the posterior probability that x_i belongs to component k,

$$w_{ik} = \frac{\pi_k f_{ik}}{\sum_{l=1}^{K} \pi_l f_{il}} \tag{2}$$

where for the (original) generative linear mixture model

$$f_{ik} = \frac{1}{(2\pi)^{m/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \alpha - \beta z_k)^T \Sigma^{-1}(x_i - \alpha - \beta z_k)\right), \tag{3}$$

one obtains the corresponding (expected) complete log-likelihood,

$$l = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log \pi_k + w_{ik} \log f_{ik}.$$

For the Maximization step, the equations for parameters, $\hat{\alpha}, \hat{\beta}, \hat{z}_k, \hat{\pi}_k$ and $\hat{\sigma}_j$ are obtained by Lawson and Einbeck (2012) by taking partial derivatives of l with respect to each of the parameters. The following are the estimators when using the variance parametrizations described in section 1, with (ii) to (iv) being new contributions of this work,

(i) $\Sigma \in \mathbb{R}^{m\times m}$, $\text{diag}(\sigma_j^2)_{\{1\leq j\leq m\}}$,

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k)^2$$

(ii) $\Sigma_k \in \mathbb{R}^{m\times m}$, $\text{diag}(\sigma_{jk}^2)_{\{1\le j\le m\}}, k = 1, ..., K$,

$$\hat{\sigma}_{jk}^2 = \frac{\sum_{i=1}^n w_{ik}(x_{ij} - \hat{\alpha}_j - \hat{\beta}_j \hat{z}_k)^2}{\sum_{i=1}^n w_{ik}}$$

(iii) $\Sigma \in \mathbb{R}^{m\times m}$, $\Sigma = \Sigma_1 =, ..., = \Sigma_k$, $k = 1, ..., K$,

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)^T$$

(iv) $\Sigma_k \in \mathbb{R}^{m\times m}$, $k = 1, ..., K$,

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^n w_{ik}(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)(x_i - \hat{\alpha} - \hat{\beta}\hat{z}_k)^T}{\sum_{i=1}^n w_{ik}}$$

2.2 Identifiability

There is a product term of βz_k in (3), which makes the parameters β, z_k unidentifiable. Furthermore, also α is unidentifiable as the same model could be attained by translating all z_k's along the line. In order to fix this identifiability problem, we standardize z_k, by letting

$$\sum_{k=1}^{K} \pi_k z_k = 0, \qquad \sum_{k=1}^{K} \pi_k z_k^2 - (\pi_k z_k)^2 = 1, \tag{4}$$

where $\text{Var}[z_k] = \sum_{k=1}^K \pi_k z_k^2 - (\pi_k z_k)^2$ (Marques da Silva Júnior et al. 2018). The first equation fixes the position of z_k's on the one-dimensional coordinate system, solving the problem for α, and the second equation solves the scale problem for β. Additionally, to identify the direction of the latent variable, we enforce $\hat{\beta}_1 \ge 0$.

3 Simulation

A simulation is set up to test the correctness of the methodology, after implementing the identifiability fixes, under variance parametrization (i). We use 2-dimensional data with three individual sample sizes $n = 100$, $n = 300$, and $n = 500$, and generate 1000 data sets for each sample size. Then we compare the average estimated values to the true values of the parameters used to generate these data, the results are shown in Table 1 and Table 2. Overall, most biases are around 0.005, and no biases greater than 0.05 were found. The estimated parameters are getting closer to the true values as the sample size gets larger (Figure 1).

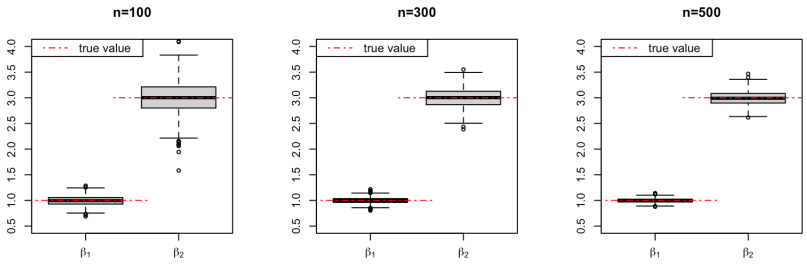


FIGURE 1. Estimations of key parameter β with different sample sizes.

TABLE 1. Simulation Results for β, α and z_k.

	β_{true}	α_{true}	z_{true}
	(1.0000, 3.0000)	(-1.0000, 1.0000)	(2.8023, 1.1675, -0.6171)
n	$\bar{\beta}_{est.}$	$\bar{\alpha}_{est.}$	$\bar{z}_{est.}$
100	(0.9915, 2.9974)	(-0.9936, 1.0235)	(2.8547, 1.2262, -0.6186)
300	(0.9986, 2.9982)	(-0.9985, 1.0036)	(2.8130, 1.1693, -0.6193)
500	(0.9966, 2.9899)	(-0.9985, 0.9983)	(2.8119, 1.1708, -0.6191)

4 Application

The data used here is the Soils data set (Fox et al. 2020). We construct a data frame of six variables: Nitrogen, Phosphorous (in ppm), Calcium, Magnesium, Phosphorous (in me/100 gm) and Sodium from the Soils data set. The features in this data frame are on wildly different scales and in different units. We apply the methodology with variance parametrization (ii). Fitting a model with $k = 3$ mass points leads to an AIC value of 828.4529. When adding one mass point and refitting the model with $k = 4$, the AIC value drops to 823.3411, and does not drop significantly when increasing k further. Then we obtain the projected data points by $x_i' = \sum_{k=1}^{K} w_{ik}\hat{z}_k$ (Aitkin, 1996). We fit a linear regression model with the variable Density (in gm/cm^3) as the response variable and the projected data as the predictor. For fair comparison, we construct the first principal component scores

TABLE 2. Simulation Results for π_k and σ_j.

	π_{true}	σ_{true}
	(0.0500, 0.2500, 0.7000)	(0.5000, 2.0000)
n	$\bar{\pi}_{est.}$	$\bar{\sigma}_{est.}$
100	(0.0463, 0.2518, 0.7019)	(0.5043, 1.9866)
300	(0.0507, 0.2504, 0.6988)	(0.4966, 1.9892)
500	(0.0498, 0.2512, 0.6990)	(0.4985, 1.9912)

by projecting all data points onto the 1-dimensional space and use these scores as the predictor. Table 3 shows the statistical measures that evaluate the performance of these two regression models. We find that our approach performs better for the non-scaled data, and that both approaches perform similarly for the scaled data.

5 Conclusion

In the presented approach to dimension reduction, the original data are linearly approximated, and represented by a single latent variable of which we conceptually think as a random effect. The distribution of the random effect is described by a discrete mixture, whose parameters are estimated through the EM algorithm along with the other model parameters. The mixture centres can be thought of as cluster centres positioned along a straight line spanned through the original data space, where several parametrizations are possible to describe the shape of the clusters around those centres. After solving an identifiability problem with this methodology, a simulation study has evidenced that parameters are accurately estimated. The real data application demonstrated that the approach is competitive to Principal Component Regression (PCR) in the special case of a one-dimensional approximation of the space of predictors, that it is not unduly affected by scales or units, and in particular is robust to scaling.
Another important application of the proposed methodology would be the joint ranking of multiple continuous variables (via the posterior random effect) with view to the construction of league tables. This would however require a multi-level version of this methodology, which allows for at least two levels as well as the inclusion of covariates on both. This work is currently in progress.

TABLE 3. Statistical measures of fit for the two regression models.

Regression Model	Non-scaled Data	Scaled Data
Mixture-based Approach	R^2 : 0.7534	R^2 : 0.7457
	$RMSE$: 0.1084	$RMSE$: 0.1096
Principal Component Regression	R^2 : 0.6226	R^2 : 0.7435
	$RMSE$: 0.1378	$RMSE$: 0.1099

References

Aitkin, M. (1996) Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In: *Proceedings of the 11th International Workshop on Statistical Modelling*, pp 87–94, Orvieto, Italy.

Lawson, A. and Einbeck, J. (2012) Generative linear mixture modelling. In: *Proceedings of the 27th International Workshop on Statistical Modelling*, pp 595–600, Prague, Czech Republic.

Marques da Silva Júnior, A.H., Einbeck, J. and Craig, P.S. (2018) Fisher information under Gaussian quadrature models, *Statistica Neerlandica*, **72**, 74–89.

Fox, J., Weisberg, S. and Price, B. (2020) carData: Companion to Applied Regression Data Sets. R package version 3.0-4. https://CRAN.R-project.org/package=carData

Dependent Dirichlet Mixture Processes for Causal Inference

Dafne Zorzetto[1], Falco J. Bargagli-Stoffi[2], Antonio Canale[1], Francesca Dominici[2]

[1] Department of Statistics, University of Padova, Italy.
[2] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Massachusetts, USA.

E-mail for correspondence: dafne.zorzetto@phd.unipd.it

Abstract: In this work we propose a Bayesian nonparametric model that exploits probit stick-breaking processes for the estimation of causal effects in observational studies. The proposed model leverages the flexibility of the nonparametric specification to overtake the imputation of the missing potential outcomes in the context of causal inference, under the standard Rubin Causal Model. The proposed model allows us to: (i) estimate the individual treatment effects, (ii) identify the subgroups defined by similar conditional treatment effects, and (iii) characterize the heterogeneity in the effects in a precise and interpretable manner.

Keywords: Bayesian Nonparametric; Rubin Causal Model; Heterogeneous effects; Clustering; Observational studies.

1 Introduction

Bayesian nonparametric (BNP) offers an interesting new perspective for causal inference. The high flexibility of BNP methods is well-known in many contexts, but their application to causality is quite recent. BNP is seldom applied in causal inference due to the complex relationships between outcomes and confounders, that require computationally intensive modeling of the joint distribution of all of the observed data (Oganisian, 2021). However, Roy et al. (2018) underline that recent developments in computing capacity allow for new powerful Bayesian approaches in causal inference.
Linero and Antonelli (2021) review BNP applications in causal inference. In particular, Hahn et al. (2020) develop a rework of the Bayesian Addi-

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

tive Regression Trees (BART) model (introduced in Chipman et al., 2010) that re-parameterizes the conditional average treatment effect directly incorporating an estimate of the propensity function. The usage of infinite mixture models has been proposed in the causal inference literature as well. Schwartz et al. (2011) apply the Dirichlet process (DP) (Escobar and West, 1995) mixtures in the context of principal stratification, Roy et al. (2018) in missing at random confounders context, and Oganisian et al. (2021) use the DP mixture for the zero-inflated regressions. However, the DP has never been used in causal inference with the aim to group the observations based on the heterogeneity in the effects.
To account for this shortcoming of the literature on BNP in causal inference, we propose to use the probit stick-breaking processes (Rodriguez and Dunson, 2011) to allow the observed confounders to influence the probability of each subject to be part of subgroups with higher/lower heterogeneity in the causal effects. The proposed method allows us to group subjects with similar treatment effects and, thanks to the posterior predictive distributions, to impute subgroup-specific (and individual-specific) treatment effects. The proposed approach provides a flexible, yet computationally scalable, algorithm that incorporates the confounders in the weights of the infinite mixture of the outcome probability distribution.

2 Probit Stick-breaking Process for Causal Inference with Observational Data

We observe n units (read, individuals), each of which can potentially be assigned a treatment. For each unit i we observe the treatment level ($T_i = 1$ for treatment and $T_i = 0$ for control), a p-dimensional vector, $\mathbf{X}_i$, of background characteristics (*confounders*), and the outcome Y_i. Accordingly to the Rubin Causal Model (Rubin, 1974), we postulate the existence of two potential outcomes: $Y_i(0)$ which is the potential outcome when unit i is assigned to the control, and $Y_i(1)$ when the same unit is assigned to the treatment. In causal inference, one wants to measure the causal effect on the outcome caused by the treatment, namely the difference between the potential outcome under treatment and under control for each unit: $\tau_i = Y_i(1) - Y_i(0)$. Unfortunately, one can never observe both potential outcomes as any unit can not be simultaneously assigned to both treatment and control. Rubin (1974) refers to the missing potential outcomes as counterfactual outcomes. In this context of partial exchangeability, the nonparametric mixtures provide effective and powerful tools to impute the missing potential outcomes via the estimation of the marginal distributions of $Y(0)$ and $Y(1)$, the imputation of the counterfactual outcomes and the estimation of the treatment effect for each unit ($\hat{\tau}_i$). Furthermore, the probit stick-breaking (Rodriguez and Dunson, 2011) permits to incorporate the confounders in the weights of the nonparametric mixture that defines the distribution of the potential

outcome and allows for the grouping of units into subgroups defined by similar imputed $\hat{\tau}_i$.
In particular, the model is defined as

$$Y_i(0)|\mathbf{X}_i \sim \sum_{l\geq 1} \omega_{0l}(\mathbf{X}_i)\mathcal{N}(\eta_{0l}, \sigma_0^2), \quad Y_i(1)|\mathbf{X}_i \sim \sum_{l\geq 1} \omega_{1l}(\mathbf{X}_i)\mathcal{N}(\eta_{1l}, \sigma_1^2),$$

$$\omega_{tl}(\mathbf{X}_i) = \Phi(\alpha_{tl}(\mathbf{X}_i)) \prod_{r<l} [1 - \Phi(\alpha_{tr}(\mathbf{X}_i))], \quad \text{for } t = \{0, 1\}.$$

Where, for $t = \{0, 1\}$, $\{\omega_{tl}\}_{l\geq 1}$ is an infinite sequence of weights such that $\omega_{tl} \geq 0$ and $\sum_{l=1}^{+\infty} \omega_{tl} = 1$; $(\{\eta_{tl})\}_{l\geq 1}, \sigma_t^2) \overset{iid}{\sim} H$ with H called the base distribution; $\Phi(\cdot)$ is the Gaussian cumulative distribution function; and $\alpha(\cdot)$ is a parametric function.

2.1 Simulation study

The ability of the proposed model to identify the data heterogeneity and precisely estimate the causal effects is evaluated in a simulation study. In particular, we focused in two scenarios. Both scenarios involve a binary treatment, confounding variables, and continuous outcomes. In the first scenario (reported in the left panel of Figure 1) the outcome decreases due to the treatment, with different intensity within different subgroups defined by the confounders. In the second scenario (reported in the right panel of Figure 1) the outcome under control is different in each subgroup, while the outcome under treatment being identically distributef in each subgoup. In observational studies, we can observe only the marginal distribution of the observed potential outcomes (the histograms in Figure 1), but we want to reconstruct the join distribution (the scatter-plots in Figure 1), in order to have both the potential outcomes for each of n units. Clearly, it is impossible to observe both the outcomes for the units, so the information about the correlation between the two marginal distribution can not be learned directly from the data. However this information can be acquired by confounders.
Essentially, the observed outcome are used to estimated the proposed model, and the distribution of missing outcomes are compute successively. In particular, in this second step, the allocation of each unit to the groups is driven by the observed counfounders $\mathbf{X}$, that are included in the the weights of infinite mixture model, while the value of the missing outcome is simulated from the assigned kernels of the mixture model.
Table 1 reports the estimated causal effect τ, obtained as the posterior mean of the difference between the outcomes under treatment and under control, over the units. Where for each unit, the posterior distribution of τ_i is a function the observed outcome and the posterior distribution of the missing outcome. Notably the results of the proposed model are competitive

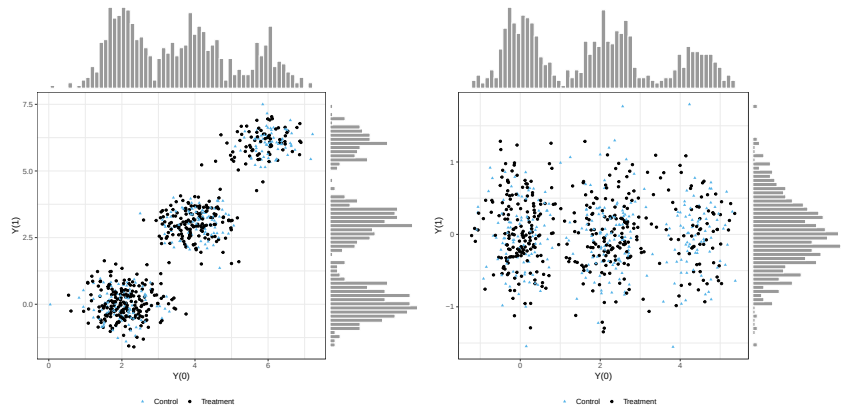


FIGURE 1. Simulation studies: first scenario (left) and second scenario (right). The scatter-plots report both the potential outcomes. The black points are the units with $Y(0)$ observed and $Y(1)$ missing (with marginal distribution reported in the x-axis histogram) and the blue triangles are the units with $Y(1)$ observed and $Y(0)$ missing (with marginal distribution reported in the y-axis histogram).

TABLE 1. Bias and mean square error (MSE) of causal effect τ, computed with the proposed model (PSB) and BART, for the two scenarios in Figure 1. Mean and standard deviation (in the brackets) are reported.

	Bias		**MSE**	
	PSB	**BART**	**PSB**	**BART**
1st scenario	-0.0068	-0.0073	0.3019	0.3027
	(0.0289)	(0.0285)	(0.0165)	(0.0168)
2nd scenario	0.0019	0.00133	0.2005	0.2015
	(0.0214)	(0.0218)	(0.0106)	(0.0110)

with BART model (Chipman et al., 2010), for both the described scenarios, with values for bias and mean square error close to zero.
Moreover, the peculiarity of the proposed model is the ability to identify groups of homogeneous units. Specifically, an estimate of the partition is obtained by minimizing the variation of information loss function as described in Wade and Ghahramani (2018). The estimated partition is compared with the true paritition assumed when simulating the data through the Rand adjusted index (Rand, 1971), a cluster comparison criterion. The range of Rand index is from 1, when the estimated partition is the same of real partition, and 0, when the estimated and real partition do not match. With the proposed model we obtain values close to 1 for Rand index.
Conditionally on the estimated partition, we can compute a subpopulation-specific causal effect τ. In Figure 2 the causal effect, measured within each

estimated group with a significant (more than 0.1%) percentage of observations, are reported, with bias and mean square error (MSE). The small values of bias and MSE show how the proposed model can identify successfully the the heterogeneity between units in the causal effect.
Within groups, we can compute measure that characterize the units, as the percentage of observations reported in Figure 2, or the distributions of the obsevred covariates $\mathbf{X}$, or various treatment effects.

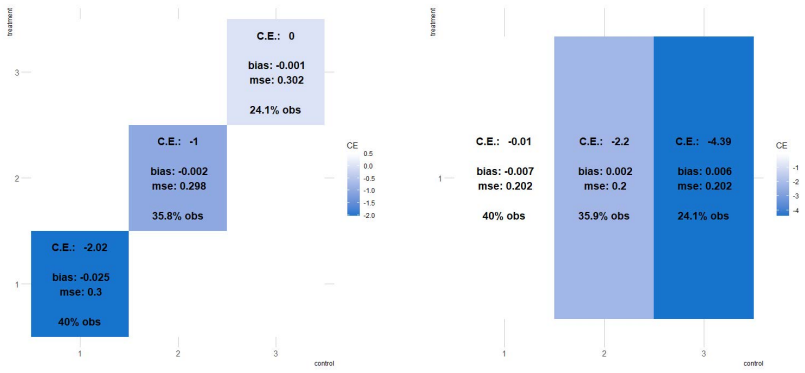


FIGURE 2. Causal effect (CE) τ measured within significant groups in the simulated scenarios. Bias, mean square error (MSE), and percentage of observations are reported, for each group. A group is considered significant if it contains more than 0.1% of the units.

References

Chipman, H.A., George, E.I., and McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*,**4(1)**. 266 – 298.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577 – 588.

Hahn, P.R., Murray, J.S., and Carvalho, C.M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, **15(3)**, 965 – 1056.

Linero, A.R. and Antonelli, J.L. (2021). The how and why of Bayesian nonparametric causal inference. *arXiv preprint arXiv:2111.03897*.

Oganisian, A. (2021). Bayesian nonparametric models for causal inference and clustering under Dirichlet process priors. In *Doctoral dissertation, University of Pennsylvania*.

Oganisian, A,, Mitra, N., and Roy, J.A. (2021). A Bayesian nonparametric model for zero-inflated outcomes: Prediction, clustering, and causal estimation. *Biometrics*, **77(1)**, 125 – 135.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66 (336)**, 846 -– 850.

Rodriguez, A. and Dunson, D. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, **6**, 145 – 178.

Roy, J., Lum, K.J., Zeldow, B., Dworkin, J.D., Re III, V.L., and Daniels, M.J. (2018). Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics*, **74(4)**, 1193 – 1202.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688 – 1202.

Schwartz, S.L., Li, F., and Mealli, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association*, **106(496)**, 1331 – 1344.

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, **13(2)**, 559 – 626.

Modelling the recurrence of injuries in football players using piece-wise exponential additive mixed models

Lore Zumeta-Olaskoaga[1,2], Andreas Bender[3], Helmut Küchenhoff[3], Dae-Jin Lee[1]

[1] BCAM - Basque Center for Applied Mathematics, Spain
[2] Departamento de Matemáticas, Universidad del País Vasco UPV/EHU, Spain
[3] Statistical Consulting Unit StaBLab, Ludwig-Maximilians Universität München, Munich, Germany

E-mail for correspondence: lzumeta@bcamath.org

Abstract: Sports injury research has gained increased interest in professional sports, including professional football. Injuries are common and modelling and understanding their occurrence would assist in developing tailored prevention programs. This work aims at modelling the recurrence of injuries in an elite male football team participating in LaLiga. We propose the use of piece-wise exponential additive mixed models for modelling such data and to study the correlation between injuries and within-player variability.

Keywords: piece-wise exponential additive mixed models; survival analysis; sports analytics; football injuries; recurrent events.

1 Introduction

Statistical modelling of football injuries has become an increasing area of interest in the field of sports injury research, as it has in professional football clubs. Injury data poses many challenges for statistical modelling. An overview of existing strategies to monitor and predict occurrence and duration of sports injuries comprising classical statistical and machine learning models is given by Ruddy *et al.* (2019). In this work, we study the risk and timing of football related injuries in an elite male football team participating in LaLiga, taking into account that injuries are of time varying and recurrent nature. Given that players sustain more than one injury over time, we aim at studying whether a player's risk of further injury is the same as for the first injury, and also whether these risks vary across players. In this regard, piece-wise exponential additive mixed models (PAMMs) offer a flexible and useful methodological toolbox for recurrent time-to-event

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

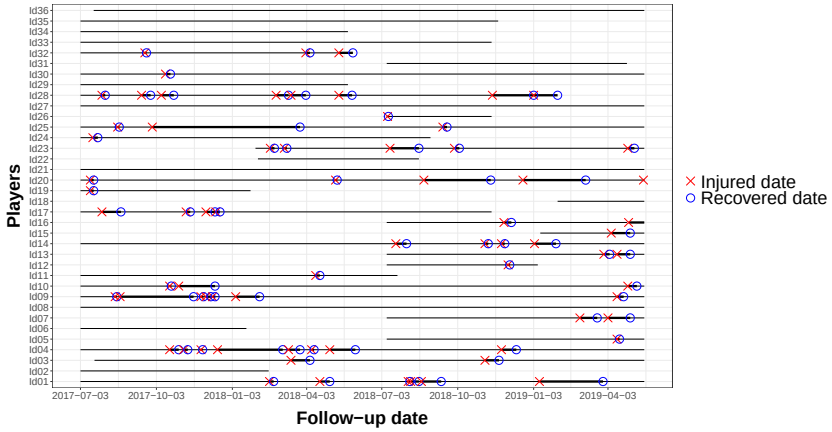


FIGURE 1. Overview of injuries that occurred during the 17-18 and 18-19 seasons. Timeline of the football players is depicted horizontally: red cross indicates the exact injury date, blue circle the recovery date and bold black line indicates the duration of the injury (time-loss).

data, that takes advantage of advanced inferential and algorithmic methods that have been developed for generalized additive mixed models (Bender *et al.* 2018, Ramjith *et al.* 2021).

2 Data

Our application is based on a cohort of an elite male football club. Injury data was recorded during the seasons 2017-2018 and 2018-2019 for 36 players. The data, collected by the club's medical staff, includes the time spent training and time competing in games (in minutes) and the time loss incurred by non-contact injuries. Figure 1 shows an overview of the injuries sustained by each player, with their dates of injury and recovery indicating their onset, together with the player's follow-up period. The median exposure time per player was 30828 minutes. A total of 72 non-contact time-loss injuries resulting in 1595 days of absence were recorded. This is equivalent to an incidence of 3.9 injuries and an injury burden of 86.2 days lost, per 1000 hours of exposure, respectively.

3 Modelling approach

In essence, the follow-up period $(0, t_{\max}]$ is partitioned into J intervals with $J+1$ cut points, i.e. $0 = \kappa_0 < \kappa_1 < \ldots < \kappa_J = t_{\max}$ and the hazard is assumed to be constant in each interval. The formulation of the hazard rate of the i-th injury (event) of the l-th player is given by:

$$\lambda_{l,i}(t|\boldsymbol{x}_l(t), l) := \exp\left(f(\boldsymbol{x}_{lj}(t), t_j, l)\right) = \lambda_{l_i,j} \tag{1}$$

for all $t \in (\kappa_{j-1}, \kappa_J]$, $t > 0$ and $l = 1, \ldots, L$; $i = 1, \ldots, n_l$, where t is the time of interest, t_j a fixed timepoint in the j-th interval (e.g. $t_j := \kappa_j$), $\boldsymbol{x}(t) \in \mathbb{R}^P$ potentially time-varying covariates and $f(\cdot)$ the effect of (time-dependent) covariates on the hazard, that can be potentially (non-linearly) time-varying and injury-specific. This general representation allows to study different dependence structures arising in football injury data.
Here, relaxing some of the terms in (1), we investigate the following models:

1. **Stratified hazards model**: assuming different baseline hazards for each of the injury events and thus considering the dependence induced by the previous injuries.

$$\begin{aligned}\lambda_{l,i}(t|\boldsymbol{x}_l) &:= \lambda(t|\boldsymbol{x}_l, l, i) = \lambda_{0,i}(t)\exp(\boldsymbol{x}_l^\top\boldsymbol{\beta}) = \\ &= \exp\left(\beta_{0,i} + f_{0,i}(t_j) + \boldsymbol{x}_l^\top\boldsymbol{\beta}\right), \quad \forall t \in (\kappa_{j-1}, \kappa_j].\end{aligned}$$

2. **Shared frailty model**: assuming a common baseline hazard for all events ($\lambda_{0,i} = \lambda_0$), but accounting for within-player correlation by introducing a frailty term.

$$\begin{aligned}\lambda_{l,i}(t|\boldsymbol{x}_l) &:= \lambda(t|\boldsymbol{x}_l, b_l, i) = \lambda_0(t)\exp(\boldsymbol{x}_l^\top\boldsymbol{\beta} + b_l) = \\ &= \exp\left(\beta_0 + f_0(t_j) + \boldsymbol{x}_l^\top\boldsymbol{\beta} + b_l\right), \quad \forall t \in (\kappa_{j-1}, \kappa_j],\end{aligned}$$

   where $b \sim N(\boldsymbol{0}, \boldsymbol{D})$ is a Gaussian random effect.

The key idea of PAMMs is that by using penalized splines for fitting the baseline hazard (e.g. by P-splines; Eilers and Marx, 1996), the problem of the arbitrary choice of the cut-points defining the intervals is overcome, i.e. the choice of the number and placement of knots for the construction of basis functions.

4 Application

We fitted stratified hazards and shared frailty models to study whether the risk of injury varies (i) across injuries, i.e. is the hazard of further injuries different to the one for initial injury, and/or (ii) across players, i.e. do some players inherently have higher or lower hazard to get injured (frailty). We used `pammtools` **R** package (Bender *et al.* 2018) for model fitting and visualisation.
We found that stratifying hazards by injury recurrence, a two level categorical variable (first or recurrent injury, as categories), mostly captured within-player variability and that the frailty term was not needed. The best

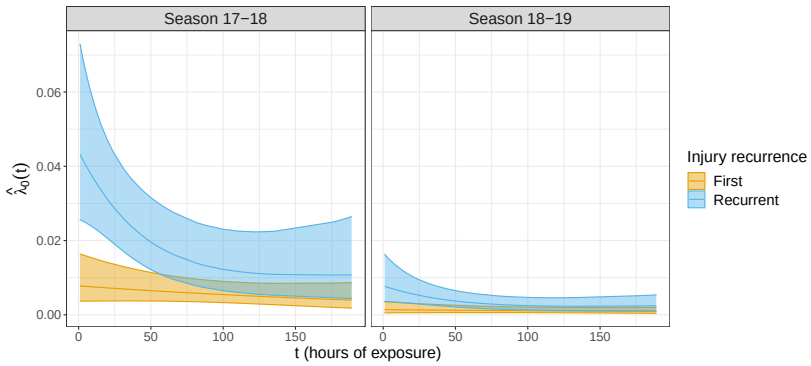


FIGURE 2. Estimated baseline hazards by a stratified PAM by two factor variables: injury recurrence (whether first or recurrent) and season (17-18 or 18-19)

fit, based on AIC criterion, was found using two additive factor-smooth interaction terms: stratifying by injury recurrence and season.
Figure 2 shows estimated baseline hazards of the aforementioned model. The risk of a recurrent injury found to be higher than the risk of sustaining a first injury in both seasons. Besides, 17-18 season injury rates were higher than the 18-19 rates, and in 17-18 season hazards rate of being injured of a first or recurrent injury were significantly different in the first 50 hours of exposure.

5 Discussion

We illustrated how to account for dependencies induced by recurrent injuries and within-person variability through a case study. The rate of injury occurrence varied across injuries and time, being higher for subsequent ones. No significant within-player variability was found. PAMM framework showed to be an adequate and comprehensive modelling approach for football injury data, which allowed to study recurrent events very flexibly. Further extensions of the model could provide more insights into the data, such as the inclusion of performance related (e.g. workload) time-dependent covariates effect along different cycles of a training program.

Acknowledgments: We thank Medical Services of Athletic Club for data support. We acknowledge the support of the Basque Government through the BERC 2022-2025 program; of the Ministry of Science, Innovation and Universities through BCAM Severo Ochoa accreditation and through SEV-2017-0718 PRE2018-084007 funding; and of AEI/FEDER, UE through the PID2020-115882RB-I00 and acronym “S3M1P4R”.

References

Bender, A., Groll, A. and Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, **18**(3-4), 299 – 321.

Bender, A. and Scheipl, F. (2018). Pammtools: Piece-wise exponential additive mixed modeling tools. *arXiv preprint arXiv:1806.01042*.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, **11**(2), 89 – 121.

Ramjith, J., Bender, A., Roes, K. C. and Jonker, M. A. (2021). Recurrent Events Analysis with Piece-wise exponential Additive Mixed Models. Preprint (Version 1) available at *Research Square* https://doi.org/10.- 21203/rs.3.rs-563303/v1.

Ruddy, J. D., Cormack, S. J., Whiteley, R., Williams, M. D., Timmins, R. G. and Opar, D. A. (2019). Modeling the risk of team sport injuries: a narrative review of different statistical approaches. *Frontiers in physiology*, **10**, 829.

Ullah, S., Gabbett, T. J. and Finch, C. F. (2014). Statistical modelling for recurrent events: an application to sports injuries. *British journal of sports medicine*, **48**(17), 1287 – 1293.

UNIVERSITÀ
DEGLI STUDI
DI TRIESTE