Proceedings of the

38th International Workshop on Statistical Modelling

Durham, UK
14-19 July, 2024

Editors: Jochen Einbeck, Reza Drikvandi, Georgios Karagiannis, Konstantinos Perrakis, Qing Zhang

Proceedings of the 38th International Workshop on Statistical Modelling Durham, July 14-19, 2024, Jochen Einbeck, Reza Drikvandi, Georgios Karagiannis, Konstantinos Perrakis, Qing Zhang (Editors),
Durham, 2024.
ISBN 978-0-907552-44-4

Editors:

Jochen Einbeck, jochen.einbeck@durham.ac.uk
Reza Drikvandi, reza.drikvandi@durham.ac.uk
Georgios Karagiannis, georgios.karagiannis@durham.ac.uk
Konstantinos Perrakis, konstantinos.perrakis@durham.ac.uk
Qing Zhang, qing.zhang@durham.ac.uk

Durham University
Mathematical Sciences & Computer Science Building
Durham University
Upper Mountjoy Campus
Stockton Road
Durham University
DH1 3LE

Preface

Dear Delegates,

It is a huge pleasure to host the 38th International Workshop on Statistical Modelling in Durham. The conference is attended by more than 130 delegates from about 20 countries, who, between them, will present about 100 papers to the audience, in form of oral or poster presentations.

This is the 4th time that this conference resides in the UK, following Exeter (1994), Glasgow (2010), and Bristol (2018). As you will have realized on your way to the conference venue, Durham is pretty far up in the North of England, and indeed there have been frequent interactions with Scotland over the centuries — some of which you will have opportunity to get some insight into when we visit Auckland Castle in one of the excursions on Wednesday.

The community which annually finds its way to the IWSM conferences is a quite extraordinary one — it is a friendly, welcoming, open community (albeit still driven by a rather specific vision of contemporary statistical science), which puts particular emphasis on the integration and support of postgraduate students. Indeed, it is quite remarkable that, among the 56 contributed oral presentations at this conference, 31 are delivered by postgraduate students. As some way to honour their contributions, the workshop will once more provide prizes for the best student oral presentation, poster, and paper. The Statistical Modelling Society has furthermore contributed two overseas travel awards for presenting PhD students, and the Durham Research Methods Centre two Durham summer grants to facilitate attendance of the short course.

A particular thank goes to the Scientific Committee, who took up their roles soon after the Trieste meeting in 2022, and have closely worked with the Organizers since then. The work of the SC included suggesting and selecting invited speakers, reviewing submissions to the conference (all contributions got reviewed in the same way, whether oral or poster, and whether student or non-student), and, even during this meeting, the scoring of the student prizes.

We are also particularly looking forward to the presentations by the invited speakers, Dimitris Rizopoulos (Erasmus MC Rotterdam), Robin Henderson (Newcastle University), Fiona Steele (University College London), Maria Kateri (RWTH Aachen), and Ernst Wit (Lugano), who will cover a range of exciting topics in their talks, touching the boundaries of the state of current knowledge in their fields. In addition, Emanuele Giorgi (Lancaster) is delivering a short course on model-based geostatistics for Public Health.

Arriving into Durham on the eve of the workshop, one can't help noting that the sentiment appears familiar – a long summer night with omnipresent screens displaying the football finals. Surely this will result in some more, and some less, smiling faces in the audience in the early sessions of the conference on Monday! But certainly, the science will take over soon, and we we will be engaged in a week of stimulating exchanges on new developments and advances in statistical modelling.

Durham, July 2024

Jochen Einbeck[1,2], Reza Drikvandi[1,2], Hyeyoung Maeng[1], Konstantinos Perrakis[1,2], Georgios Karagiannis[2], and Qing Zhang[2]

[1] Hosts of the IWSM 2024; [2] Editors of this volume.

Scientific Committee

Elisabeth Bergherr, University of Göttingen (Germany)
Kevin Burke, University of Limerick (Ireland)
Enrico Colosimo, Federal University of Minas Gerais (Brazil)
Riccardo de Bin, University of Oslo (Norway)
Reza Drikvandi (Co-Chair), Durham University (UK)
Jochen Einbeck (Chair), Durham University (UK)
Andreas Groll, TU Dortmund University (Germany)
Marco Grzegorczyk, University of Groningen (The Netherlands)
Kenan Matawie, Western University Sydney (Australia)
Helen Ogden, University of Southampton (UK)
Konstantinos Perrakis (Co-Chair), University of Durham (UK)
Pere Puig, Universitat Autònoma de Barcelona (Spain)
Rosalba Radice, Bayes Business School (UK)
Javier Rubio, University College London (UK)
Cibele M Russo, Universidade de São Paulo (Brazil)
Jeff Simonoff, Leonard N. Stern School of Business (US)
Claudia Tarantola, University of Pavia (Italy)
Veronica Vinciotti, University of Trento (Italy)

Organising Committee

Hosts: Jochen Einbeck, Reza Drikvandi, Hyeyoung Maeng, Konstantinos Perrakis

Local organizing committe members: Tahani Coolen-Maturi, Jonathan Cumming, Andy Golightly, Sam Jackson, Georgios Karagiannis, Zhaocheng Li, James Liley, Emmanuel Ogundimu, Rachel Oughton, Adam Stone, Germaine Uwimpuhwe, Qing Zhang, Yingjuan Zhang

External organizing committe members: Kevin Burke (Limerick, Ireland), Giampiero Marra (UCL, UK), Pete Philipson (Newcastle, UK), Paul Wilson (Wolverhampton, UK), Bruce Worton (Edinburgh, UK)

Contents

Part I – Invited Papers

Robin Henderson: Statistical modelling for big and little data 3

Maria Kateri: Statistical modelling through the lens of divergence measures 4

Dimitris Rizopoulos, Jeremy M.G. Taylor: Dynamic predictions from joint models using super learning 12

Fiona Steele, Siliang Zhang, Paul Clarke: Modelling correlations among grouped random effects in multilevel models with an application to the estimation of household effects on longitudinal health outcomes 19

Ernst C. Wit: Advances in modelling dynamic networks 25

Part II – Contributed Papers

Timo Adam, Mads Peter Heide-Jørgensen, Susanne Ditlevsen: Modelling narwhal diving behaviour and responses to sound exposure using stochastic differential equations with state-switching coefficients 35

Pedro Miranda Afonso, Dimitris Rizopoulos, Rhonda D. Szczesniak, Eleni-Rosalina Andrinopoulou: A joint model for multiple (un)bounded longitudinal markers, competing risks, and recurrent events 39

Víctor Alonso-Lara, Amanda Fernández-Fontelo, Pere Toran, Meritxell Gómez-Maldonado, Gemma Falguera, David Moriña: Statistical detection of barriers in primary care access for young victims of gender-based violence 43

João Flávio Andrade Silva, Rafael Izbicki, Leonardo S. Bastos, Guilherme P. Soares: Monitoring viral infections in severe acute respiratory syndrome patients in Brazil 48

Sébastien Appleby, George Iosifidis, Arthur White: Empirical study of the Sardex network 49

Robert Bajons, Kurt Hornik: Adaptive generalized logistic lasso and its application to rankings in sports 53

Tathagata Basu: Regression analysis with missing data using interval imputation 54

Ina-Marie Berendes, Alexander Gerharz, Andreas Groll, Mathias Kolodziej: Modeling and predicting injuries in soccer with machine learning and conventional statistical approaches 55

Moritz Berger, Maria Iannario: Penalized mixed cumulative regression for modelling varying dispersion and cluster heterogeneity 60

Herlisson Bezerra, Luis F. B. de Messis, Cibele Russo, Thomas Peron: A comparison of methods for oil production forecasting of the Santos Basin 65

Anabel Blasco-Moreno, Pere Puig: Estimating a lower bound of the population size in capture-recapture experiments with right censored data 69

Martin P. Boer: REML for two-dimensional P-splines 70

Martina Boschi, Ernst C. Wit: Relational event additive modeling of alien plant and insect species invasions 71

Guillermo Briseño Sanchez, Nadja Klein, Andreas Groll, Andreas Mayr : Boosting distributional copula regression for bivariate time-to-event data 75

Kevin Burke, James P. Gleeson, Mike Quayle: Modelling social interaction data 81

Yilun Cai, Jochen Einbeck, Stephen Barnard, Elizabeth Ainsbury: Estimating dose and time of exposure from a protein-based radiation biomarker 85

Angela Carollo, Nicole Hiekel, Valeria Ferraretto: A boosted multistate model of partnership trajectories in Germany 86

Miguel de Carvalho, Vianey Palacios Ramirez: Nonparametric Bayesian modeling of nonstationary joint extremes 91

Joaquin Cavieres, Cole C. Monnahan, David Bolin, Elisabeth Bergherr: Approximated Gaussian random field under different parameterizations for MCMC 95

Claudia Collarin, Matteo Fasiolo: Integrating single index effects in generalized additive models 97

Tahani Coolen-Maturi: Combining biomarkers through PCA 98

Ilse Cuevas Andrade, Ardo van den Hout, Nora Pashayan: Nonparametric frailty model for the natural history of prostate cancer; using data from a screening trial 102

Davide Di Cecco, Andrea Tancredi: Species sampling with misidentification: a Bayesian parametric approach 103

Ultán P. Doherty, Rachel McLoughlin, Tracey Claxton, Arthur White: gateTree: A user-informed tree algorithm for population identification in flow cytometry 108

Paul H.C. Eilers, Martin P. Boer: Derivatives of the log of a determinant 114

Dalila Failli, Bruno Arpino, Maria Francesca Marino, Francesca Martella: A biclustering approach via mixture of latent trait analyzers for the analysis of digital divide in Italy 115

Luisa Ferrari, Massimo Ventrucci: Variance partitioning-based priors for species distribution models 116

Rita Fici, Luigi Augugliaro, Ernst C. Wit: Functional copula graphical regression model for analysing brain–body rhythm 122

Matteo Framba, Veronica Vinciotti, Ernst C. Wit: Inference for quasi-reaction models with covariate-dependent rates 123

Cristina Galán-Arcicollar', Danilo Alvares, Josu Najera-Zuloaga, Dae-Jin Lee: A joint modelling approach for longitudinal patient-reported outcomes and survival analysis 124

Vincenzo Gioia, Gioia Di Credico, Francesco Pauli: A Bayesian Markov-switching for smooth modelling of extreme value distributions 129

Colin Griesbach, Elisabeth Bergherr: Additive mixed models for location, scale and shape via gradient boosting techniques 130

Jacob Grytzka, Andreas Groll: Statistical modeling of UEFA EURO soccer matches with focus on player market value and other hybrid variables 131

Marco Grzegorczyk: Learning Bayesian networks from ordinal data - The Bayesian way 137

Gillian Heller, Andrew Forbes, Stephane Heritier : Statistical models for patient-centered outcomes in clinical studies 138

John Hinde, Alberto Alvarez-Iglesias, John Ferguson, Clarice G.B. Demétrio, John Crown, Bryan T. Hennessy, Vicky Donachie: Modelling of overdispersed count rates 139

Vanda Inácio, Richard A. Parker: Induced robust inference for the concordance correlation coefficient 140

Amaia Iparragirre, Irantzu Barrio: Optimism correction of the AUC with complex survey data 144

Darshana Jayakumari, Jochen Einbeck, John Hinde, Rafael A. Moral: A distance-based statistic for goodness-of-fit assessment 145

Rūta Juozaitienė, Ernst C. Wit: Misleading effects in relational event models 146

Lea Kaufmann, Maria Kateri: Likelihood ratio tests in penalized logistic regression with categorical covariates 150

Oliver Kemp, Ioannis Kosmidis: Bias-reducing adjustments for generalised additive models 154

Hannah Klinkhammer, Andreas Mayr, Carlo Maj, Peter M. Krawitz, Christian Staerk: Anchor regression to enhance transferability of genetic prediction models 159

Lars Knieper , Thomas Kneib , Elisabeth Bergherr: Spatial confounding in gradient boosting 165

Alexey A. Koloydenko, Ioan Notingher, Radu Boitor, Jüri Lember: Shape analysis of AF segments for rapid assessment of Mohs layers for BCC presence by AF-Raman microscopy 166

Jan-Ole Koslik, Carlina C. Feldmann, Sina Mews, Rouven Michels, Roland Langrock: Inferential tools for hidden Markov models with periodic components 167

Michele Lambardi di San Miniato, Michela Battauz, Ruggero Bellio, Paolo Vidoni: Bayesian ordinal regression for crowd-sourced fact-checking 171

Oskar Laverny, Nathalie Grafféo, Roch Giorgi: Non-parametric estimation of net survival under dependence between death causes ... 176

Melania Lembo, Rūta Juozaitienė, Veronica Vinciotti, Ernst C. Wit: Context matters: including global covariates in relational event models 180

Yuan Liu, Mu Niu, Claire Miller: Sparse intrinsic Gaussian processes for prediction on manifolds: extending applications to environmental contexts 184

Ivana Malá: Number of jobs during working life based on SHARE data 185

Jacob Martin, Carlo Giovanni Camarda: Modelling age-space mortality dynamics in small areas 190

Nirian Martín, J. Miguel Marín: Uncorrelatedness in high-dimension: hypothesis testing of independence in presence of outliers 194

Ana María Martínez-Rodríguez, Antonio Conde-Sánchez, María José Olmo-Jiménez, José Rodríguez-Avi: DGLMExtPois package: regression models for under-dispersed count data 200

Rui Martins: An underrated prior distribution for proportions. The logistic–normal for dynamical football predictions 206

Fernando Miguelez, Josu Doncel, Maria Dolores Ugarte: Multivariate forecasting of operational times using hidden Markov models . 207

Shirin Moghaddam, John Newell, John Hinde: Parametric and non-parametric Bayesian imputation for right censored survival data...... 211

Mabel Morales-Otero, Vicente Núñez-Antón: Bayesian approaches to model overdispersion in spatio-temporal binomial data 212

Lorenzo Mori, Maria Rosaria Ferrante: Addressing covariate lack in unit-level small area models using GAMLSS 213

Vito M.R. Muggeo, Gianluca Sottile: Segmented quantile regression process with τ-varying breakpoints 214

Eamonn Organ, James Sweeney: Spatio-temporal models for high resolution wind speed maps 219

Léa Orsini', Emmanuel Lesaffre, Guosheng Yin, Caroline Brard, David Dejardin, Gwénaël Le Teuff: Bayesian analysis of restricted mean survival time adjusted on covariates using pseudo-observations . 225

David J.P. O'Sullivan, Caroline Pena , Alina Dubovskaya: A mathematically tractable model for information diffusion between communities 230

Samyajoy Pal, Christian Heumann : Gene coexpression analysis with Dirichlet mixture model: accelerating model evaluation through closed-form KL divergence approximation using variational techniques 234

Andrea Panarotto, Manuela Cattelan, Ruggero Bellio: State-space models for clustering of compositional trajectories.............. 235

Roman Parzer, Peter Filzmoser, Laura Vana-Gür: Random projections for classification with high-dimensional data 236

Roman Pfeiler, Helga Wagner: Shrinkage in a Bayesian panel data model with time-varying coefficients 240

Johannes Piller, Helmut Küchenhoff, Andreas Bender: Flexible additive models for multi-event survival analysis 241

Luigi Cao Pinna, Claire Miller, Marian Scott: Latent Dirichlet allocation and hidden Markov models to identify public perception of sustainability in social media data 247

Nicola Piras, Silvia Columbu, Jeroen K. Vermunt: Model selection procedure in multilevel cross-classified latent class models 248

Xavier Piulachs, Klaus Langohr, Guadalupe Gómez: Semi-Markov multistate model with interval-censored transition times.............. 252

Sophie Potts, Anja Rappl, Karin Kurz, Elisabeth Bergherr: Joint models for longitudinal and time-to-event data in social science research 253

Daniel Racek, Paul Thurner, Göran Kauermann: Non-parametric smoothing for the diffusion of armed conflict 259

Sumal Randeni, Kenan M. Matawie, Laurence A. F. Park: Residual analysis in information retrieval models 265

Martje Rave, Göran Kauermann: Bridging the data gap - Estimate the exit rates, entries, and exits using only data on occupancy 271

Dayasri Ravi, Andreas Groll: Optimizing variable selection in multi-omics datasets: a focus on exclusive lasso 275

Michael Renfrew, Bruce J. Worton: A comparison of extreme gradient and Gaussian process boosting for a spatial logistic regression on satellite data 276

Juan M. Rodríguez-Díaz, M. Teresa Santos-Martín, Irene Mariñas-Collado: Design optimality for a general alcohol model for the human body 277

Ullrika Sahlin, Zheng Zhou: A suggestion for a test if a calibrated quantitative adverse outcome pathway is chemical agnostic based on between chemical heterogeneity 281

Kristína Sakmárová, Arnošt Komárek, Martin Otava: Employing random effects in variance components modelling 285

Mohammad Sayari, James Durrand, Christopher Taylor, Jochen Einbeck, Ehsan Kharatikoopaei, Joshua Craig, Nathan Griffiths: Using linear mixed models to compare a self-assessed frailty score with clinician assessed scores in patients approaching major surgery 291

Quentin Edward Seifert, Elisabeth Bergherr, Tobias Hepp: Function-on-scalar regression via first-order gradient-based optimization 297

Nikolai Spuck, Matthias Schmid, Moritz Berger: Confidence intervals for tree-structured varying coefficients based on parametric bootstrap 301

Annika Strömer, Nadja Klein, Andreas Mayr: A model-based boosting approach to deal with dependent censoring 307

Daniele Tancini, Francesco Bartolucci, Silvia Pandolfi: Bayesian hidden Markov models for early warning 313

Arantxa Urdangarin, Tomás Goicoa, María Dolores Ugarte: Mitigating spatial confounding: a unified approach integrating simplified spatial+ and restricted regression 314

Paul Wilson: An updated Wilcoxon–Mann–Whitney test 318

Qing Zhang, Germaine Uwimpuhwe, Dimitris Vallis, Akansha Singh, Tahani Coolen-Maturi, Jochen Einbeck: Elicitation of priors for intervention effects in educational trial data.................. 319

YINGJUAN ZHANG, JOCHEN EINBECK: R package mult.latent.reg for multivariate response scenarios with latent structures 320

ZHAOXI ZHANG, VANDA INÁCIO, MIGUEL DE CARVALHO, SARA WADE: The underlap coefficient as a measure of a biomarker's discriminatory ability and covariate dependence in cluster analysis 324

WEIYUE ZHENG, ANDREW ELLIOTT, CLAIRE MILLER, MARIAN SCOTT: Spatial regression with misaligned covariates for soil moisture mapping 329

ZHAOYUAN ZOU, RUTH O'DONNELL, CLAIRE MILLER, DUNCAN LEE, CRAIG WILKIE: A computationally efficient spatio-temporal fusion model for reflectance data... 334

Part I – Invited Papers

Statistical modelling for big and little data

Robin Henderson[1]

[1] Newcastle University, UK

E-mail for correspondence: Robin.Henderson@ncl.ac.uk

Abstract: While the difference between "Data Science" and "Statistics" disciplines is, at best, blurred, many people associate machine learning methods and big data with the former, and modelling and inference for small samples (little data) with the latter. We present a big data application where no sophisticated method at all is needed, a small data application where a partial modelling approach seems useful, and a big-and-little data application where we can borrow strength from limited information in a large sample, to improve estimation based on more detailed data in a small sample.

Keywords: Data science; Extrapolation; Inference; Smoothing; Two cultures.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Statistical modelling through the lens of divergence measures

Maria Kateri[1]

[1] Institute of Statistics, RWTH Aachen University, Aachen, Germany

E-mail for correspondence: maria.kateri@rwth-aachen.de

Abstract: Standard statistical models such as models for contingency tables, logistic regression, and models for rank data are revisited and redefined in a statistical information theoretical context, connecting them to divergences. This fact, on the one hand, reveals new properties for these models that lead to deeper understanding of their nature and new interpretation options and, on the other hand, offers the possibility of generalising them into flexible families of models. Choosing as divergence the Cressie-Read divergence, which is a parametric family, flexible parametric families of models are derived, controlling the scale by a single parameter that can be fixed or estimated by the data.

Keywords: Association models; Correspondence analysis; Mallows model; Kullback-Leibler divergence; ϕ–divergence.

1 Introduction

Traditional topics of categorical data analysis that remain popular in diverse fields of applications (e.g., medical sciences, psychology, education and social sciences, economics and machine learning) and simultaneously retain their research interest in new frameworks, include contingency table analysis, logistic regression and modelling of rank data (cf. Agresti, 2013; Kateri, 2014; Marden, 1995). We revisit these models and point out a structural property they share in a statistical information theoretic framework. In particular, it can be proved that each of them is, under certain conditions, the closest to a parsimonious reference model, when the closeness is measured in terms of the Kullback-Leibler (KL) divergence. Keeping the conditions and the reference model but changing the divergence, different models are derived as closest to the reference model on different scales than that imposed by the KL divergence. Thus, based on the ϕ–divergence, a generalized family of divergences that includes KL, (cf. Pardo, 2006), we can define a class of models, all of the same structural nature, but measuring the divergence from a common reference model on different scales. Thus, the fit of models of a class applied on the same data may vary significantly, depending on the scale. This provides a powerful tool that allows for flexible parsimonious modelling. Section 2 deepens in the understanding and modelling of the association structure of a contingency table, while Section 3 discusses illustrative examples. Section 4 refers briefly to generalised models for binary regression and for rank data. The final Section 5 provides concluding remarks.

2 Generalized association models

Consider an $I \times J$ contingency table $\mathbf{n} = (n_{ij})$ with n_{ij} being the observed frequency in cell (i, j). Let further the sample size $n = \sum_{i,j} n_{ij}$ be fixed and the random table $\mathbf{N}$ is multinomial distributed $\mathbf{N} \sim \mathcal{M}(n, \boldsymbol{\pi})$, with probability table $\boldsymbol{\pi} \in \Delta_{IJ}$, where Δ_{IJ} is the simplex $\Delta_{IJ} = \{\boldsymbol{\pi} = (\pi_{ij}) : \pi_{ij} > 0, \ \sum_{i,j} \pi_{ij} = 1\}$, and corresponding table of expected cell frequencies $\mathbf{m} = \mathrm{E}(\mathbf{N}) = n\boldsymbol{\pi}$.
The log-linear independence model (IM) of the underlying row and column classification variables X and Y is given by

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y \ , \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \tag{1}$$

with identifiability constraints applying on the row and column main effects (e.g., $\lambda_1^X = \lambda_1^Y = 0$). Upon rejection of model (1), the only alternative in the standard log-linear models setup is the saturated model

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \ , \quad i = 1, \ldots, I, \ j = 1, \ldots, J. \tag{2}$$

Notice that the $(I-1) \times (J-1)$ table of *local odds ratios* (LOR) $\boldsymbol{\theta}^L = (\theta_{ij}^L)$, where

$$\theta_{ij}^L = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}} \ , \quad i = 1, \ldots, I-1, \ j = 1, \ldots, J-1, \tag{3}$$

along with the row and column marginal probabilities $\boldsymbol{\pi}_r = (\pi_{1+}, \ldots, \pi_{I+})^T$ and $\boldsymbol{\pi}_c = (\pi_{+1}, \ldots, \pi_{+J})^T$, specify uniquely the corresponding $I \times J$ probability table $\boldsymbol{\pi}$. Thus, given $\boldsymbol{\pi}_r$ and $\boldsymbol{\pi}_c$, model (1) can equivalently be expressed in terms of $\boldsymbol{\theta}^L$, as $\log \theta_{ij}^L = 0$, for all $i = 1, \ldots, I-1$ and $j = 1, \ldots, J-1$, while (2) imposes no structure on $\boldsymbol{\theta}^L$. Hence, under IM, all LOR of the table are equal to zero in the log-scale. For ordinal X and Y, assuming that all LOR are equal but non-zero ($\log \theta_{ij}^L = c$, $c \neq 0$), we are lead to a highly structured dependence model, known as Uniform (U) association model, that has just one parameter more than IM. It is expressed in log-linear form as

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \varphi\mu_i\nu_j \ , \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \tag{4}$$

with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_I)$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_J)$ being known scores assigned to the rows and columns, equidistant for successive categories. From (4), it follows directly that $\log \theta_{ij}^L = \varphi(\mu_i - \mu_{i+1})(\nu_j - \nu_{j+1})$, which is constant in case of equidistant successive scores, and hence φ is an intrinsic association parameter. Relaxing the assumption of equidistant row and column scores, or considering one or both of them to be unknown parameters, further *association models* (AM) are defined. When $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are both unknown, the model is the multiplicative row-column (RC) AM and is not log-linear (s. Goodman (1985) and references therein). Whenever a set of scores is parametric, the scores need not to be ordered and the associated classification variable (X or/and Y) can also be nominal. The RC model can further be extended to AM of order M, $M \leq M^* = \min(I, J) - 1$,

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \sum_{m=1}^{M} \varphi_m \mu_{im} \nu_{jm} \ , \ i = 1, \ldots, I \ , \ j = 1, \ldots, J \ , \qquad (5)$$

with the row and column scores satisfying the following orthonormalising constraints

$$\begin{aligned} \sum_i w_{1i} \mu_{im} &= \sum_j w_{2j} \nu_{jm} = 0, \quad m = 1, \ldots, M \ , \qquad (6) \\ \sum_i w_{1i} \mu_{im} \mu_{i\ell} &= \sum_j w_{2j} \nu_{jm} \nu_{j\ell} = \delta_{m\ell}, \quad m, \ell = 1, \ldots, M, \end{aligned}$$

for some weights $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, with $\delta_{m\ell}$ being Kronecker's delta. Model (5) is denoted by RC(M) and its sum term corresponds to the generalized singular value decomposition of the matrix of interaction parameters of model (2). For $M = M^*$, RC(M^*) is saturated. Expression (5) resembles another score-based method for exploring association structures in contingency tables, namely the *correspondence analysis* (CA). The CA of order M is given by

$$p_{ij} = p_{i+} p_{+j} \left(1 + \sum_{m=1}^{M} \rho_m x_{im} y_{jm} \right), \ \ i = 1 \ldots, I, \ j = 1, \ldots, J \ , \qquad (7)$$

with $\mathbf{x}_m = (x_{1m}, \ldots, x_{Im})$ and $\mathbf{y}_m = (y_{1m}, \ldots, y_{Jm})$, $m = 1, \ldots, M$, being row and column scores, satisfying constraints analogue to (6) with marginal weights, i.e., $\boldsymbol{w}_1 = \boldsymbol{\pi}_r$ and $\boldsymbol{w}_2 = \boldsymbol{\pi}_c$ (s. Greenacre, 2007). CA is mainly a descriptive method, well-known for the graphical displays of its scores. Goodman (1981) developed inferential procedures for (7), analogue to (5), and called it the row-column *correlation model* of order M, while for $M = 1$ special correlation models for known scores have also been considered (s. Goodman, 1985).

Though association and correlation models were initially opposed to each other (s. Goodman, 1986), Gilula et al. (1988) emphasised in a pioneering work their similarity and linked them in a information theoretic setup. They proved that, for given marginal distributions ($\boldsymbol{\pi}_r$ and $\boldsymbol{\pi}_c$), given scores ($\boldsymbol{\mu}$ and $\boldsymbol{\nu}$) and fixed correlation $\rho = \text{corr}(\boldsymbol{\mu}, \boldsymbol{\nu})$, both have a common property. They are the closest model to independence (1), but are differentiated by the divergence used to measure their closeness. AM are the closest in terms of the KL divergence and correlation models in terms of the Pearson's divergence. Based on this result, Kateri and Papaioannou (1994) introduced a general class of dependence models, based on the ϕ–divergence, which is a family of divergences that includes the KL and Pearson divergences as special cases.

For two discrete finite bivariate probability distributions $\boldsymbol{\pi} = (\pi_{ij})$, $\mathbf{q} = (q_{ij}) \in \Delta_{IJ}$, the ϕ–divergence between $\mathbf{q}$ and $\boldsymbol{\pi}$ (also known as Csiszar's measure of information in $\mathbf{q}$ about $\boldsymbol{\pi}$), is given by

$$\mathrm{I}_{\phi}^{C}(\mathbf{q}, \boldsymbol{\pi}) = \sum_{i,j} \pi_{ij} \phi(q_{ij}/\pi_{ij}), \qquad (8)$$

where ϕ is a real–valued strictly convex function on $[0, \infty)$ with $\phi(1) = \phi'(1) = 0$, $0\phi(0/0) = 0$, $0\phi(y/0) = \lim_{x\to\infty} \phi(x)/x$ (cf. Pardo, 2006). For $\phi(x) = x \log x$

and $\phi(x) = (1-x)^2$, (8) becomes the KL and Pearson divergence, respectively. Then, under the conditions of Gilula et al. (1988), the joint distribution $\boldsymbol{\pi}$ that is closest to independence in terms of the ϕ–divergence is

$$\pi_{ij} = \pi_{i+}\pi_{+j}F^{-1}\left(\alpha_i + \beta_j + \varphi\mu_i\nu_j\right), \quad i = 1, \ldots, I, \; j = 1, \ldots, J \;, \tag{9}$$

where F^{-1} is the inverse function of $F(x) = \phi'(x)$ and the scores $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ satisfy (6) for $M = 1$ and with marginal weights. With the additional identifiability constraints on the main effect parameters

$$\sum_i \pi_{i+}\alpha_i = \sum_j \pi_{+j}\beta_j = 0 \;,$$

it is easily verified that φ is an intrinsic association parameter, since

$$\varphi = \varphi(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \sum_{i,j} \pi_{i+}\pi_{+j}\mu_i\nu_j F\left(\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}\right),$$

and $\varphi = 0$ if and only if the independence model (1) holds (Kateri and Papaioannou, 1994) .
If $\phi(x) = \frac{x^{\lambda+1}-x}{\lambda(\lambda+1)}$, $-\infty < \lambda < \infty$ and $\lambda \neq -1,\ 0$, (8) becomes the power divergence of Cressie and Read (1984) and model (9) leads to the parametric family of models

$$\pi_{ij} = \pi_{i+}\pi_{+j}\left[\frac{1}{\lambda+1} + \lambda(\alpha_i + \beta_j + \varphi\mu_i\nu_j)\right]^{1/\lambda}, \quad i = 1, \ldots, I, \; j = 1, \ldots, J \;, \tag{10}$$

denoted by U_λ^L. For $\lambda \to 0$ and $\lambda = 1$, (10) yields (4) and (7) for $M = 1$, respectively.
Furthermore, the RC(M) model (5) is generalized through the ϕ–divergence to the class of models RC$_\phi(M)$, given by

$$\pi_{ij} = \pi_{i+}\pi_{+j}F^{-1}\left(\alpha_i + \beta_j + \sum_{m=1}^{M}\varphi_m\mu_{im}\nu_{jm}\right), \quad i = 1, \ldots, I, \; j = 1, \ldots, J \;, \tag{11}$$

with $\boldsymbol{\mu}_m$ and $\boldsymbol{\nu}_m$ satisfying (6) with marginal weights. Models (5) and (7) are derived for $\phi(x) = x\log x$ and $\phi(x) = (1-x)^2$, respectively. For a discussion on ϕ–scaled more advanced AM and their properties we refer to Kateri (2014, 2018) and references cited therein. The LOR (3) generalises to the ϕ–scaled LOR

$$\theta_{ij}^{L(\phi)}(\boldsymbol{\pi}) = F(\tilde{\pi}_{ij}) + F(\tilde{\pi}_{i+1,j+1}) - F(\tilde{\pi}_{i+1,j}) - F(\tilde{\pi}_{i,j+1}), \tag{12}$$

with $\tilde{\pi}_{ij} = \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}$, a term that underlines the comparison of π_{ij} to the corresponding cell probability under the reference model of independence. For $\phi(x) = x\log x$, (12) simplifies to the well-known $\log(\theta_{ij}^L)$, and for the power divergence, to

$$\theta_{ij}^{L(\lambda)}(\boldsymbol{\pi}) = \frac{1}{\lambda}\left(\tilde{\pi}_{ij}^\lambda + \tilde{\pi}_{i+1,j+1}^\lambda - \tilde{\pi}_{i+1,j}^\lambda - \tilde{\pi}_{i,j+1}^\lambda\right). \tag{13}$$

Model (9) can then be expressed in terms of the ϕ–scaled LOR as

$$\theta_{ij}^{L(\phi)}(\boldsymbol{\pi}) = \varphi(\mu_i - \mu_{i+1})(\nu_j - \nu_{j+1}), \quad i = 1, \ldots, I-1, \; j = 1, \ldots, J-1 \;, \tag{14}$$

and analogously model (10) in terms of (13).
Beyond the LOR, there is a variety of alternative types of odds ratios, capturing other types of association than the local (s. Douglas et al., 1990), with probably the most representative being the global odds ratios (GOR)

$$\theta_{ij}^{G}(\boldsymbol{\pi}) = \frac{P(Y \leq j|X \leq i)/P(Y > j|X \leq i)}{P(Y \leq j|X > i)/P(Y > j|X > i)},$$

and the cumulative (with respect to the columns) odds ratios (COR)

$$\theta_{ij}^{C}(\boldsymbol{\pi}) = \frac{P(Y \leq j|X = i)/P(Y > j|X = i)}{P(Y \leq j|X = i+1)/P(Y > j|X = i+1)},$$

for $i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$. The families of AM discussed so far, they all model the LOR. However, their structure can apply to other types of odds ratios as well. Bartolucci and Forcina (2002) extended the RC model for generalized odds ratios while Forcina and Kateri (2021) for the general RC$_\phi(M)$ model (11), defining and modelling ϕ–scaled extensions of generalized odds ratios.
Here, we focus on the simple uniform association structure imposed on the generalised LOR (13) through the family of models (10), as well as the power divergence scaled generalisations of the global and cumulative odds ratios, i.e., considering model (14) for the corresponding generalised $\theta_{ij}^{G(\lambda)}$ and $\theta_{ij}^{C(\lambda)}$.

3 Examples

Here, we focus on the simple uniform association structure imposed on (13) through the family of models (10), as well as the power divergence scaled generalisations of GOR and COR (i.e., $\theta_{ij}^{G(\lambda)}$ and $\theta_{ij}^{C(\lambda)}$) and the corresponding uniform association models U_λ^G and U_λ^C. We implement these models on two contingency tables, provided in Tables 1 and 2. The values of the likelihood ratio statistic G^2 for varying λ are pictured in Figure 1. We observe that the type of odds ratio that best describes the underlying association is the LOR for Table 1 and the COR for Table 2. Though the best fit is achieved by models $\mathrm{U}_{-0.04}^L$ ($\lambda = -0.04$) and $\mathrm{U}_{-1.1}^C$ ($\lambda = -1.1$), since the corresponding standard models U^L and U^C (i.e., based on the KL divergence, $\lambda \to 0$) fit also the data very well, we finally propose them, due to their simplicity of interpretation. In particular we have $G^2(\mathrm{U}^L) = 1.469$ (p-value=0.917) for Table 1 and $G^2(\mathrm{U}^C) = 2.394$ (p-value=0.495) for Table 2. The maximum likelihood estimates (MLEs) under these models are provided in parentheses in the respective tables. The estimated common LOR value under U^L for Table 1 is 2.23 while the common COR value under U^C for Table 2 is estimated as 2.03. It is further worth to note the inadequacy of the correlation model ($\lambda = 1$) for the data in Table 1.

4 Other generalized classes of models

Beyond the association models discussed above, we shall see how other well-known models can be generalized to ϕ–scaled classes of models. In particular, the quasi symmetry (QS) model for square contingency tables is (under some conditions) the closest in terms of the KL divergence to the model of complete

TABLE 1. Students' survey about cannabis use at the University of Ioannina, Greece (1995). The MLEs of the expected cell frequencies under the uniform local AM (U^L) are given in parentheses.

Alcohol consumption	I tried cannabis			
	never	once or twice	more often	total
at most once/month	204 (204.4)	6 (5.7)	1 (0.9)	211
twice/month	211 (211.4)	13 (13.1)	5 (4.5)	229
twice/week	357 (352.8)	44 (48.8)	38 (37.4)	439
more often	92 (95.3)	34 (29.4)	49 (50.3)	175
total	864	97	93	1054

TABLE 2. Cross-classification of variables PARTYID (political party identification) and GRNEXAGG (opinion about whether environmental threats are exaggerated) from the 2010 General Social Survey (GSS). The MLEs of the expected cell frequencies under the uniform local AM (U^C) are given in parentheses.

Political	Environmental Threats are Exaggerated		
Party	Agree	Neutral	Disagree
Republican	172 (167.2)	57 (59.7)	82 (84.1)
Independent	178 (189.5)	115 (107.3)	227 (223.2)
Democrat	111 (104.0)	78 (82.9)	283 (285.2)

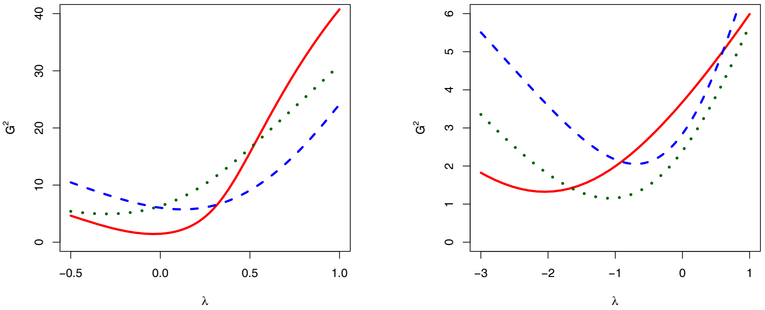


FIGURE 1. Likelihood ratio statistic (G^2) values for models U^L_λ (red line), U^G_λ (dashed blue line), and U^C_λ (dotted green line) as a function of λ, fitted on the data of Table 1 (left) and Table 2 (right).

symmetry. Based on this property, ϕ–scaled QS models have been defined (Kateri and Papaioannou, 1997; Kateri and Agresti, 2007). Analogously, the ϕ–scaled binary regression model is a flexible extension of the logistic regression (Kateri and Agresti, 2010). Finally, the ϕ–scaled Mallows model is discussed, highlighting the flexibility it offers in modelling rank data (Kateri and Nikolov, 2022).

5 Discussion

A statistical model can be characterized by the property of being the closest, under certain conditions, to a simple reference model. This property provides a new perspective for interpretation that allows a deeper understanding of the model's nature. Moreover, it enables its generalisation to a family of models by controlling the divergence used to measure this closeness, a fact that increases modelling options and may lead to parsimonious models by controlling the scale at which closeness is measured. Though we discuss scaled models for selected common setups, the concept is also applicable to other type of models.

Acknowledgments: I would like to express my sincerely thanks to the organisers of IWSM 2024 for inviting me to give this talk.

References

Agresti, A. (2013). *Categorical Data Analysis*, 3d ed. Hoboken: Wiley.

Bartolucci, F. and Forcina, A. (2002). Extended rc association models allowing for order restrictions and marginal modeling. *Journal of the American Statistical Association*, **97**, 1192 – 1199.

Cressie, N. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440 – 464.

Douglas, R., Fienberg, S.E. , Lee M.L.T., Sampson A.R. and Whitaker L.R. (1990). Positive dependence concepts for ordinal contingency tables, in: Positive Dependence Concepts for Ordinal Contingency Tables. In: *Lecture Notes - Monograph Series*, **16**, Institute of Mathematical Statistics, Hayward, CA, 189 – 202.

Forcina, A. and Kateri, M. (2021). A new general class of RC association models: Estimation and main properties. *Journal of Multivariate Analysis*, **184**, 104741 (1 – 16).

Gilula, Z., Krieger, A.M. and Ritov, Y. (1988). Ordinal association in contingency tables: some interpretive aspects. *Journal of the American Statistical Association*, **83**, 540 – 545.

Goodman, L. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association*, **76**, 320 – 334.

Goodman, L.A. (1985). Goodman, L.A. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics*, **13**, 10 – 69.

Goodman, L.A. (1986). Some useful extensions of the usual correspondence analysis and the usual log-linear models approach in the analysis of contingency tables with or without missing entries (with Discussion). *International Statistical Review*, **54**, 243 – 309.

Greenacre, M. (2007). *Correspondence Analysis in Practice*, 2nd ed. London: Chapman & Hall.

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

Kateri, M. (2018). ϕ–divergence in contingency table analysis. *Entropy*, **20**, 324 (1 – 12).

Kateri, M. and Agresti, A. (2007). A class of ordinal quasi symmetry models for square contingency tables. *Statistics & Probability Letters*, **77**, 598 – 603.

Kateri, M. and Agresti, A. (2010). A generalized regression model for a binary response. *Statistics & Probability Letters*, **80**, 89 – 95.

Kateri, M. and Nikolov, I.N. (2022). A generalized Mallows model based on ϕ–divergence measures. *Journal of Multivariate Analysis*, **190**, 104958 (1–14).

Kateri, M. and Papaioannou, T. (1994). f-divergence association models. *International Journal of Mathematical and Statistical Sciences*, **3**, 179 – 203.

Kateri, M. and Papaioannou, T. (1997). Asymmetry models for contingency tables. *Journal of the American Statistical Association*, **92**, 1124 – 1131.

Marden, J. I. (1995). *Analyzing and modeling rank data.* London: Chapman & Hall.

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures.* New York: Chapman & Hall.

Dynamic predictions from joint models using super learning

Dimitris Rizopoulos[1], Jeremy M.G. Taylor[2]

[1] Department of Biostatistics and Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, the Netherlands
[2] Department of Biostatistics, University of Michigan, USA

E-mail for correspondence: d.rizopoulos@erasmusmc.nl

Abstract: Joint models for longitudinal and time-to-event data are often employed to calculate dynamic individualized predictions used in numerous applications of precision medicine. In this work, we use the concept of super learning to specify a weighted combination of the dynamic predictions calculated from a library of joint models with different specifications. The weights are selected to optimize a predictive accuracy metric using V-fold cross-validation. We use as predictive accuracy measure the expected predictive cross-entropy. All proposed methodology is implemented in the freely available R package **JMbayes2**.

Keywords: Cross-entropy; Precision medicine; Time-varying covariates.

1 Introduction

Joint models for longitudinal and time-to-event data have been established as a versatile tool for calculating dynamic predictions for longitudinal and survival outcomes (Taylor et al., 2005; Rizopoulos, 2011; Taylor et al., 2013). The advantageous feature of these predictions is that they are updated over time as extra information becomes available. As a result, they have found numerous applications in precision medicine, including cancer and cardiovascular diseases.

The motivation for our research comes from prostate cancer patients who, after diagnosis, underwent surgical removal of the prostate gland (radical prostatectomy). The treating physicians closely monitor the prostate-specific antigen (PSA) levels of these patients to determine the risk of recurrence and metastasis. Increasing PSA values suggest the cancer may be regrowing, although it is generally not yet detectable on imaging. After the initial surgery, PSA levels drop to near zero; however, PSA may rise again for some patients, leading the treating physicians to recommend salvage therapy to reduce their risk of metastasis. After salvage therapy, PSA levels nearly always drop, sometimes substantially, but typically rise again if metastasis is going to occur.

Optimizing the accuracy of dynamic predictions from joint models is a difficult task. In particular, previous research has shown that two aspects are important. First, the time trend specification in the mixed-effects models, and second, the functional form that specify how the longitudinal history is linked to the hazard of the event. Previous applications of joint models have considered a single model for obtaining dynamic predictions. However, due to the aforementioned complexities, finding a well-specified model can be challenging, especially when multiple longitudinal outcomes are considered. Moreover, due to the dynamic nature of these predictions, different models may provide different levels of predictive accuracy at different follow-up times.

In this work we will use the concept of super learning (SL) to optimize the accuracy of dynamic predictions (Naimi and Balzer, 2018; Phillips et al., 2023). SL is an ensemble method that allows researchers to combine several different prediction algorithms into one where the candidate algorithms can be quite different. It uses V-fold cross-validation to build the optimally weighted combination of predictions from a library of candidate algorithms. Optimality is defined by a user-specified objective function, such as minimizing the mean squared error or maximizing the area under the receiver operating characteristic curve. We consider a library of joint models with different specifications and present appropriate objective functions for optimizing the accuracy of dynamic predictions. In particular, we focus on different formulations of the time effect for the longitudinal outcome and different functional forms to link this outcome with the event process. We measure the dynamic predictions' accuracy using the expected predictive cross-entropy and show how this is formulated under the SL framework.

2 Joint models

We start with a general definition of the joint modeling framework for longitudinal and time-to-event data (Rizopoulos, 2012). Let $\mathcal{D}_n = \{T_i, \delta_i, \mathbf{y}_i; i = 1, \ldots, n\}$ denote a sample from the target population, where T_i^* denotes the true event time for the i-th subject, C_i the censoring time, $T_i = \min(T_i^*, C_i)$ the corresponding observed event time, and $\delta_i = \mathbb{I}(T_i^* \leq C_i)$ the event indicator, with $\mathbb{I}(\cdot)$ being the indicator function that takes the value 1 when $T_i^* \leq C_i$, and 0 otherwise. In addition, we let $\mathbf{y}_i$ denote the $n_i \times 1$ longitudinal response vector for the i-th subject, with element y_{il} denoting the value of the longitudinal outcome taken at time point t_{il}, $l = 1, \ldots, n_i$.

To accommodate different types of longitudinal responses in a unified framework, we postulate that the response vector $\mathbf{y}_i$ conditional on the vector of unobserved random effects $\mathbf{b}_i$ has a distribution $\mathcal{F}_\psi$ parameterized by the vector ψ. This more general formulation allows for distributions not covered by the exponential family. The mean of the distribution of the longitudinal outcome conditional on the random effects has the form

$$g\big[E\{y_i(t) \mid \mathbf{b}_i\}\big] = \eta_i(t) = \mathbf{x}_i^{\mathrm{T}}(t)\beta + \mathbf{z}_i^{\mathrm{T}}(t)\mathbf{b}_i,$$

where $g(\cdot)$ denotes a known one-to-one monotonic link function, and $y_i(t)$ denotes the value of the longitudinal outcome for the i-th subject at time point t, $\mathbf{x}_i(t)$ and $\mathbf{z}_i(t)$ denote the time-dependent design vectors for the fixed-effects β and for the random effects $\mathbf{b}_i$, respectively. We let ϕ denote the scale parameter of $\mathcal{F}_\psi$, i.e., $\psi^{\mathrm{T}} = (\beta^{\mathrm{T}}, \phi)$. The random effects are assumed to follow a multivariate normal distribution with mean zero and variance-covariance matrix $\mathbf{D}$. For the survival process, we assume that the risk of an event depends on a function of the subject-specific linear predictor $\eta_i(t)$ and the random effects. More specifically, we have

$$\begin{aligned} h_i\{t \mid \mathcal{H}_i(t), \mathbf{w}_i\} &= \lim_{s\to 0} \Pr\{t \le T_i^* < t+s \mid T_i^* \ge t, \mathcal{H}_i(t), \mathbf{w}_i\}/s \\ &= h_0(t) \exp\bigl[\gamma^{\mathrm{T}} \mathbf{w}_i + f\{\eta_i(t), \mathbf{b}_i, \alpha\}\bigr], \quad t > 0, \end{aligned}$$

where $\mathcal{H}_i(t) = \{\eta_i(s), 0 \le s < t\}$ denotes the history of the underlying longitudinal process up to t, $h_0(\cdot)$ denotes the baseline hazard function, $\mathbf{w}_i$ is a vector of baseline covariates with corresponding regression coefficients γ. Finally, the baseline hazard function $h_0(\cdot)$ is modeled flexibly using a B-splines approach, i.e.,

$$\log h_0(t) = \sum_{p=1}^{P} \gamma_{h_0,p} B_p(t, \lambda),$$

where $B_p(t,\lambda)$ denotes the p-th basis function of a B-spline with knots $\lambda_1, \dots, \lambda_P$ and γ_{h_0} the vector of spline coefficients.
The function $f(\cdot)$, parameterized by vector α, specifies which features of the longitudinal outcome process are included in the linear predictor of the relative risk model. Some examples are:

$$f\{\mathcal{H}_i(t), \mathbf{b}_i, \alpha\} = \begin{cases} \alpha\eta_i(t), \\ \alpha\eta_i'(t), \text{ with } \eta_i'(t) = \frac{\mathrm{d}\eta_i(t)}{\mathrm{d}t}, \\ \alpha\eta_i''(t), \text{ with } \eta_i''(t) = \frac{\mathrm{d}^2\eta_i(t)}{\mathrm{d}t^2}, \\ \alpha\frac{1}{v}\int_{t-v}^{t} \eta_i(s)\,\mathrm{d}s, \quad 0 < v \le t, \\ \alpha^{\mathrm{T}}\mathbf{b}_i. \end{cases}$$

These formulations of $f(\cdot)$ postulate that the hazard of an event at time t is associated with the underlying level of the biomarker at the same time point, the slope/velocity of the biomarker at t, the acceleration of the biomarker at t, the average biomarker level in the period $(t-v, t)$, or the random effects alone. Combinations of these functional forms and their interactions with baseline covariates are also often considered.

3 Optimizing predictions via super learning

3.1 Model weights

The basic idea behind super learning is to derive model weights that optimize the cross-validated predictions. More specifically, we consider a library with L models denoted by $\mathcal{L} = \{M_1, \dots, M_L\}$. There are no restrictions to the models included in this library, and actually, it is recommended to consider a wide range of possible models. Among others, these joint models differ in the specification of the time trend in the longitudinal submodels (e.g., linear or nonlinear trajectories), the functional form for the longitudinal outcome in the event submodel, and the functional form of the other covariates (e.g., interactions and nonlinear terms) in both submodels.

We split the original dataset $\mathcal{D}_n$ in V folds. The choice of V will depend on the size and number of events in $\mathcal{D}_n$. In particular, for each fold, we need to have a sufficient number of events to robustly quantify the predictive performance. Using the cross-validation method, we fit the L models in the combined $v - 1$ folds, and we will calculate predictions for the v-th fold that was left out. Due to the dynamic nature of the predictions, we want to derive optimal weights at different follow-up times. More specifically, we consider the sequence of time points $t_1, \ldots, t_Q$. The number and placing of these time points should again consider the available event information in $\mathcal{D}_n$. For example, we should have at least 10-15 events per time interval (t_{q-1}, t_q), with $q = 1, \ldots, Q$, and $t_0 = 0$.
For any $t_q \in \{t_1, \ldots, t_Q\}$, we define $\mathcal{R}(t_q, v)$ to denote the subjects at risk at time t_q that belong to the v-th fold. For all subjects in $\mathcal{R}(t_q, v)$, we calculate the cross-validated predictions (conditioning on the covariates $\mathbf{w}_i$, $\mathbf{x}_i$ and $\mathbf{z}_i$ is assumed in the following expressions but omitted to simplify notation),

$$\hat{\pi}_i^{(v)}(t_q + \Delta t \mid t_q, M_l) = \Pr\{T_i^* < t_q + \Delta t \mid T_i^* > t_q, \mathcal{H}_i(t), M_l, \mathcal{D}_n^{(-v)}\}.$$

These predictions are calculated based on model M_l in library $\mathcal{L}$ that was fitted in the dataset $\mathcal{D}_n^{(-v)}$ that excludes the subjects in the v-th fold. The calculation is based on a Monte Carlo approach (Rizopoulos, 2011). We define $\hat{\tilde{\pi}}_i^v(t_q + \Delta t \mid t_q)$ to denote the convex combination of the L predictions, i.e.,

$$\hat{\tilde{\pi}}_i^v(t_q + \Delta t \mid t_q) = \sum_{l=1}^{L} \varpi_l(t_q) \hat{\pi}_i^{(v)}(t_q + \Delta t \mid t_q, M_l), \quad \text{for all } v \in 1, \ldots, V,$$

with $\varpi_l(t_q) > 0$, for $l = 1, \ldots, L$, and $\sum_l \varpi_l(t_q) = 1$. Note that the weights $\varpi_l(\cdot)$ are time-varying, i.e., at different follow-up times, different combinations of the L models may yield more accurate predictions. The weighted combination of the predictions from the L models is typically called the ensemble super learner (eSL). The model with the best cross-validated prediction metric is called the discrete super learner (dSL). Most often, but not always, this is the model with the largest weight $\varpi_l(\cdot)$.

3.2 Measuring predictive performance

For the ensemble super learner and for any time t, we will select the weights $\{\varpi_l(t); l = 1, \ldots, L\}$ that optimize the predictive performance of the combined cross-validated predictions. As a scoring rule in the interval $(t, t+\Delta t]$ we consider an adaptation of the expected predictive cross-entropy proposed by Commenges et al. (2012):

$$\text{EPCE}(t + \Delta t, t) = E\Big\{ -\log\Big[p\big\{T_i^* \mid t < T_i^* \leq t + \Delta t, \mathcal{Y}_i(t), \mathcal{D}_n\big\}\Big]\Big\},$$

where the expectation is taken with respect to $\{T_i^* \mid T_i^* > t, \mathcal{Y}_i(t)\}$ under the true model. An estimate of $\text{EPCE}(t + \Delta t, t)$ that accounts for censoring can be obtained using the sample at hand,

$$\widehat{\text{EPCE}}(t + \Delta t, t) = \frac{1}{n_t} \sum_{i:T_i > t} -\log\Big[p\big\{\tilde{T}_i, \tilde{\delta}_i \mid T_i > t, \mathcal{Y}_i(t), \mathcal{D}_n\big\}\Big],$$

where $\tilde{T}_i = \min(T_i, t + \Delta t)$ and $\tilde{\delta}_i = \delta_i \mathbb{I}(t < T_i \leq t + \Delta t)$. The $\text{EPCE}(t + \Delta t, t)$ assumes that $T_i^* \perp\!\!\!\perp C_i \mid \mathcal{Y}_i(t), \mathbf{w}_i$.

To use the EPCE for obtaining the super-learning model-specific weights, we need to formulate it as a function of the dynamic predictions from a joint model. It is convenient to redefine $\pi_i(u \mid t, M_l)$ as the dynamic subject-specific survival probabilities, i.e.,

$$\pi_i(u \mid t, M_l) = \Pr\{T_i^* > u \mid T_i^* > t, \mathcal{Y}_i(t), \mathcal{D}_n, M_l\}, \quad u > t.$$

Then we write the conditional predictive log-likelihood as (conditioning on M_l is assumed but is omitted from the following expressions for exposition):

$$\begin{aligned}
&\log\Big[p\big\{\tilde{T}_i, \tilde{\delta}_i \mid T_i > t, \mathcal{Y}_i(t), \mathcal{D}_n\big\}\Big] = \\
&\qquad \tilde{\delta}_i \log[h_i\{\tilde{T}_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\}] + \log \frac{\Pr\{T_i^* > \tilde{T}_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\}}{\Pr\{T_i^* > t \mid \mathcal{Y}_i(t), \mathcal{D}_n\}}.
\end{aligned}$$

The second term is $\log\{\pi_i(\tilde{T}_i \mid t)\}$. For the first term, we write the hazard function as

$$h_i\{\tilde{T}_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\} = -\frac{\frac{\mathrm{d}}{\mathrm{d}t} \Pr\{T_i^* > t \mid \mathcal{Y}_i(t), \mathcal{D}_n\}\Big|_{t=\tilde{T}_i}}{\Pr\{T_i^* > \tilde{T}_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\}},$$

and we approximate the derivative with a forward difference, i.e.,

$$\begin{aligned}
&h_i\{\tilde{T}_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\} \approx \\
&\qquad -\frac{\Pr\{T_i^* > \tilde{T}_i + \epsilon \mid \mathcal{Y}_i(t), \mathcal{D}_n\} - \Pr\{T_i^* > \tilde{T}_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\}}{\epsilon \Pr\{T_i^* > \tilde{T}_i \mid \mathcal{Y}_i(t), \mathcal{D}_n\}} \\
&= \quad \frac{1 - \pi_i(\tilde{T}_i + \epsilon \mid \tilde{T}_i)}{\epsilon}, \quad \epsilon \to 0.
\end{aligned}$$

Combining these two terms, we get the final expression:

$$\begin{aligned}
&\widehat{\text{EPCE}}(t + \Delta t, t) = \\
&\qquad -\frac{1}{n_t} \sum_{i: T_i > t} \tilde{\delta}_i \big[\log\{1 - \pi_i(\tilde{T}_i + \epsilon \mid \tilde{T}_i)\} - \log(\epsilon)\big] + \log\{\pi_i(\tilde{T}_i \mid t)\}.
\end{aligned}$$

In practice, we can compute $\widehat{\text{EPCE}}(t + \Delta t, t)$ using a small value for ϵ, e.g., $\epsilon = 0.001$. Numerical experiments we performed showed that the EPCE values are minimally affected by the value of ϵ.

In our context, $\text{EPCE}(t + \Delta t, t)$ is calculated using the convex combination of the cross-validated predictions $\tilde{\hat{\pi}}_i^v(t + \Delta t \mid t)$. In particular, using the super-learning procedure, we obtain the weights $\widehat{\varpi}_l(t)$ that maximize the $\text{EPCE}(t + \Delta t, t))$ of the cross-validated predictions,

$$\widehat{\varpi}_l(t) \quad = \quad \text{argmin}_{\varpi} \left[\sum_{v=1}^{V} \mathcal{S}\Big\{\sum_{l=1}^{L} \varpi_l \hat{\pi}_i^{(v)}(t + \Delta t \mid t, M_l), T_i, \delta_i\Big\}\right],$$

under the constraints $\varpi_l(t) > 0$, for $l = 1, \ldots, L$, and $\sum_l \varpi_l(t) = 1$. We can transform to an unconstrained optimization problem using the logistic transformation and use a general-purpose minimization algorithm (e.g., using functions `optim()` or `nlminb()` in R). The vignette *'Combined Dynamic Predictions via Super Learning'* (available at `https://drizopoulos.github.io/JMbayes2/`) describes how the super learning procedure is implemented in package **JMbayes2**.

4 University of Michigan prostatectomy data analysis

We return to our motivating University of Michigan Prostatectomy Data. We considered four versions of the linear mixed-effects model for the log(PSA + 1) longitudinal outcome. In the first model, we specified linear subject-specific time trends for log(PSA + 1) that change after the first salvage therapy. The second model considers the same specification as the previous model, but we additionally include the baseline covariates age at surgery, Charlson's index, Gleason score, and baseline PSA. These covariates are allowed to have a different effect after salvage. The third model considers nonlinear subject-specific time trends before salvage and linear subject-specific time trends after salvage. The final fourth model has the same specification for the time trends as the third one, but we again include the same covariates as in the second model. Using each of these linear mixed models, we fitted three joint models with different specifications of the hazard submodel for metastasis. In each hazard model, we include the covariates mentioned in the second and fourth linear mixed models above and a time-varying component. Each hazard submodel has a different specification of the time-varying component. In the first specification, we consider the current value of log(PSA + 1); in the second one, the current value and the velocity of log(PSA + 1); and in the third one, the mean log(PSA + 1) from the start of the follow-up. Each of these specifications has two branches, one before and one after salvage therapy. Hence, in total, we consider twelve joint models.

We split the UMP data into five folds and fitted the twelve joint models, holding out a fold each time. We calculated the cross-validated predictions from these models for the fold not used when fitting them. We aim to evaluate the predictive performance of the twelve models in two medically relevant time intervals $(t, t + \Delta t]$, namely, $(4, 7]$ and $(6, 9]$. At follow-up year 4, 2514 patients were still at risk, and 28 patients had metastasis in the interval $(4, 7]$. At follow-up year 6, 1914 patients were still at risk, and 16 patients had metastasis in the interval $(6, 9]$. We calculated the EPCE for each model using its cross-validated predictions. We also obtained the super learning weights that combined the dynamic predictions from these models to optimize the EPCE using the weighted predictions. Table 1 presents the results. The expected predictive cross-entropy seems sensitive in quantifying differences in the predictive performance of the different models. This also results in smaller super learning estimates of the EPCE than in each of the twelve models, and for both time intervals. In the $(4, 7]$ interval the nonlinear joint models with no covariates and the slope and mean functional form dominate the weights. In the $(6, 9]$ interval the weights are distributed among almost all models.

5 Discussion

In this paper, we presented an adaptation of the super learning framework for optimizing dynamic predictions from joint models for longitudinal and time-to-event data. We considered the expected predictive cross-entropy as a predictive accuracy metric and compared two super learning versions. Namely, the discrete super learner selects the model with the best cross-validated prediction metric among the candidate models, and the ensemble super learner specifies an optimal convex combination of the predictions from all candidate models. In the University of Michigan Prostatectomy Data, the ensemble super learner performed better than the discrete one.

TABLE 1. Expected predictive cross-entropy (EPCE) for the University of Michigan Prostatectomy Data under the twelve joint models, and their combination using super learning. Results are based on 5-fold cross-validation.

	$(t, t+\Delta t] = (4, 7]$		$(t, t+\Delta t] = (6, 9]$	
	EPCE	weights	EPCE	weights
SL	0.07208		0.05166	
linear-noCov-value	0.07347	0.00026	0.05543	0.03259
linear-noCov-slope	0.07299	0.00004	0.05471	0.15417
linear-noCov-mean	0.07476	0.00235	0.05365	0.01329
linear-Cov-value	0.07338	0.00000	0.05506	0.02167
linear-Cov-slope	0.07298	0.00006	0.05455	0.09639
linear-Cov-mean	0.07484	0.00274	0.05353	0.02836
nonlinear-noCov-value	0.07324	0.00000	0.05376	0.01562
nonlinear-noCov-slope	0.07242	0.79539	0.05436	0.00317
nonlinear-noCov-mean	0.07457	0.18346	0.05303	0.02524
nonlinear-Cov-value	0.07316	0.00000	0.05337	0.10840
nonlinear-Cov-slope	0.07265	0.00121	0.05283	0.08131
nonlinear-Cov-mean	0.07454	0.01448	0.05284	0.41979

References

Commenges, D., Liquet, B., and Proust-Lima, C. (2012). Choice of prognostic estimators in joint models by estimating differences of expected conditional Kullback-Leibler risks. *Biometrics*, **68**, 380 – 387.

Naimi, A., and Balzer L. (2018). Stacked generalization: An introduction to super learning. *European Journal of Epidemiology*, **33**, 459 – 464.

Phillips, R., van der Laan, M., Lee, H., and Gruber, S. (2023). Practical considerations for specifying a super learner. *International Journal of Epidemiology*, **52**, 1276 – 1285.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, **67**, 819 – 829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R*. Boca Raton: Chapman & Hall/CRC.

Taylor, J.M.G., Park, Y., Ankerst, D., Proust-Lima, C., et al. (2013). Real-time individual predictions of prostate cancer recurrence using joint models. *Biometrics*, **69**, 206 – 213.

Taylor, J.M.G., Yu, M., and Sandler, H. (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology*, **23**, 816 – 825.

Modelling correlations among grouped random effects in multilevel models with an application to the estimation of household effects on longitudinal health outcomes

Fiona Steele[1], Siliang Zhang[2], Paul Clarke[3]

[1] Department of Statistics, London School of Economics & Political Science, UK
[2] School of Statistics, East China Normal University, China
[3] Institute for Social and Economic Research, University of Essex, UK

E-mail for correspondence: `f.a.steele@lse.ac.uk`

Abstract: A standard assumption of multilevel models is that all the random effects at a given level in the data structure are independent for different units. We develop multilevel models for grouped data structures where correlations are allowed between pairs of random effects for units in the same group, and within-group random effect correlations may depend on covariates that characterise the relationship between pairs of units. Constrained MCMC estimation is used to ensure that the group-specific correlation matrices are positive definite. The research is motivated by the study of household effects in longitudinal studies where household membership may change over time. Household random effects are allowed to be correlated within clusters of households that share individuals over time, with correlations depending on covariates that describe the connections between household pairs. The proposed model is applied in analyses of household and area effects on self-rated health in the UK.

Keywords: Correlated random effects; Correlation model; Joint mean-covariance model.

1 Introduction

There is a long history of joint mean-covariance models that allow correlations among observations to depend on covariates. Much of this research has focused on longitudinal data where within-individual covariance (or correlation) matrices are modeled as functions of time (e.g. Pourahmadi 1999). Outside the longitudinal setting, models have been proposed for covariances in high-dimensional multivariate data where the number of correlated responses p is large and the number of observations n small (Zou et al. 2022). Zhang et al. (2023) also consider the multivariate case, but for small p and large n, where correlations among a set of individual-specific latent variables measured by multivariate binary items depend on covariates.

In this paper, we build on this previous research to relax standard assumptions about the covariance structure of random effects in multilevel models. While it is common practice to allow for correlation between *different* random effects for the *same* unit, for example in 'random slopes' multilevel models between the random intercept and the random coefficients of the predictor variables, the random effects for different units (e.g. their random intercepts) are assumed to be drawn independently. To address this, we develop models for grouped multilevel data structures in which correlations are allowed between the random effects for different units in the same group, and within-group random effect correlations may depend on covariates that characterise the relationship between pairs of units.

The proposed model is motivated by the study of household effects in longitudinal studies which is complicated by changes in household membership over time. We specify a multilevel model with time-varying household-specific random effects defined for each unique combination of coresidents. These household effects are allowed to be correlated within clusters of households that share individuals over time, with correlations depending on covariates that describe the connections between household pairs. The application of this 'grouped' random effects model is illustrated in analyses of household and area effects on physical and mental health in the UK. More generally, the proposed approach can be applied in situations where random effects are grouped and the dependence of their within-group correlations on covariates is of interest. One potential application is to longitudinal data on individuals who are clustered, for example in families, schools or workplaces which are fixed over time, and the within-cluster correlations between pairs of individual random effects depends on pair-specific covariates.

2 Motivating application

There is considerable interest among health researchers in the degree of correlation in coresidents' health-related attitudes, behaviours and outcomes. Household effects have been found in measures of general health, physical health functioning, mental health and wellbeing. However, the estimation of household effects with longitudinal data presents a major methodological challenge because, while a household can be defined cross-sectionally as a group of individuals who share accommodation, it is difficult to define households in a longitudinal sense because of changes in household composition over time. As a result, most longitudinal analyses have ignored household effects and considered only individual effects. Apart from missing a potentially important component of variation that is of substantive interest, omitting household effects may lead to underestimation of standard errors of coefficients of household-level covariates and misleading inferences regarding the relative contributions of omitted variables at the individual and higher levels of aggregation (e.g. areas).

Most previous studies have sidestepped the problem of changing household membership by examining household effects at a cross-section. The few studies that have attempted to estimate household effects in a longitudinal analysis have taken one of three approaches. The simplest of these, usually applied to couple data, is to restrict analysis to those who remain coresident throughout the observation period, resulting in a highly selective sample. The other two approaches explicitly allow for changes in household composition. The first is a multiple membership random effects model (Goldstein et al. 2000) which includes a time-varying household effect specified as a weighted sum of independent random effects for the households that an individual belongs to over the observation period, but this imposes an overly restrictive and unrealistic association structure among coresidents (Steele et al. 2019). Steele et al. instead propose a flexible marginal modelling approach that allows for correlation within clusters containing all observations from individuals who have lived together, or have mutual coresidents, during the observation period. They propose a joint model comprising a marginal model for the mean of each outcome and a model for the within-cluster correlations which may both depend on covariates. While the marginal model provides substantively interesting information about the within and between individual association structure, however, researchers are often interested in estimating the relative contributions of omitted individual and household characteristics. Moreover, its flexibility does not extend to allowing for area effects, or other forms of higher-level clustering.

To address these limitations, we propose a random effects model that allows for changes in household composition and in area of residence over time, without the restrictive assumptions on the covariance structure of the multiple membership model. The model is used to estimate the relative contributions of unmeasured individual, household and area characteristics to individuals' self-rated mental and physical health using annual household panel data.

3 Multilevel model with correlated household random effects

We describe the proposed model in terms of its application to the study of household and area effects on repeated measures of a continuous response. Denote by Y_{ti} the response for individual i at wave t and $\mathbf{x}_{ti}$ a vector of potentially time-varying individual and household covariates with coefficient vector $\boldsymbol{\beta}$. We consider a multilevel linear model of the form

$$Y_{ti} = \boldsymbol{\beta}'\mathbf{x}_{ti} + u_i + v_{h(ti)} + w_{a(ti)} + e_{ti} \tag{1}$$

where $w_{a(ti)} \sim N(0, \sigma_a^2)$ is an area effect associated with the area of residence of individual i at wave t, $u_i \sim N(0, \sigma_u^2)$ is an individual effect and $e_{ti} \sim N(0, \sigma_e^2)$ is a time-varying residual. The household effect is $v_{h(ti)}$ where $h(ti)$ denotes the household of individual i at wave t, and thus the effect of household changes whenever there is a change in an individual's coresidents.

To allow for possible overlap in the membership of pairs of households, we allow for correlation among $v_{h(ti)}$. A key component of the correlation model specification, also employed in the marginal modelling approach of Steele et al. (2019), is the "superhousehold" cluster constructed to contain outcomes from individuals who are connected through coresidence, either directly or indirectly through a mutual coresident.

Suppose there are K superhouseholds and let m_s be the number of households in superhousehold $s \in \{1, \ldots, K\}$. Labelling households consecutively within superhouseholds as $1, \ldots, m_s$, let $\mathbf{v}_s = (v_1, \ldots, v_{m_s})$. We assume $\mathbf{v}_s \sim N(\mathbf{0}, \boldsymbol{\Omega}_{vs})$ where $\boldsymbol{\Omega}_{vs} = \sigma_v^2 \mathbf{R}_{vs}$, $\text{var}(v_h) = \sigma_v^2$ for $h = 1, \ldots, m_s$ and $\mathbf{R}_{vs}$ is a correlation matrix. Denote by $s(h)$ the superhousehold of household h, $s(h) \in \{1, \ldots, K\}$. The elements of $\mathbf{R}_{vs}$ are modelled as linear functions of covariates $\mathbf{z}$

$$\text{cor}(v_h, v_{h'}) = \boldsymbol{\gamma}' \mathbf{z}_{h,h'} \quad \text{for } h, h' = 1, \ldots, m_s;\ h \neq h';\ s(h) = s(h') \qquad (2)$$

where $\mathbf{z}_{h,h'}$ includes variables that characterise the connection between households h and h', for example the number of individuals (if any) in both h and h' and their relationship, and $\boldsymbol{\gamma}$ is a vector of coefficients. We assume $\text{cor}(v_h, v_{h'}) = 0$ when $s(h) \neq s(h')$.

We propose a block-wise Gibbs sampling procedure in which individual, (grouped) household and area-level random effects are sampled in turn followed in the final stage by the fixed parameters of the multilevel model including those of the linear model for the household-group correlations. The sampling of the correlation parameters in the final step must be constrained to ensure the between-household correlation matrix for each superhousehold remains positive definite. This is achieved by extending the constrained correlation Metropolis-Hastings sampler developed by Zhang et al. (2023) for a fixed 4 × 4 correlation matrix (in their case for latent variables capturing the dimensions of reciprocal giving and receiving) to one with common parameters for each of K correlation matrices for the superhouseholds where the dimension of each varies between superhouseholds.

4 Data analysis

The joint mean-correlation model of eq. (1) and (2) was applied to annual data from the UK Household Longitudinal Survey for the period 2009–2020. The response variables are separate continuous scales for everyday physical and mental functioning, based on the widely-used SF-12 instrument. We focus on selected results from the correlation model where the covariates $\mathbf{z}_{h,h'}$ in (2) describe the connection between the members of households h and h'. As we expect that correlations between household effects on health would be highest for households that share closely related individuals, we focus on indicators of whether households share partners (past, current or future) and whether a member of one household is the parent of a member of the other. Such partnership and parent-child links account for 91% of all pairs, with the remainder contributed mainly by unrelated sharers. In addition we include the proportion of the total number of individuals across the two households who are in both households.

To aid interpretation, the parameter estimates from the correlation model are used to calculate predicted correlations for each pair of households. The mean predictions are then computed for each possible combination of the indicators of partnership and parent-child links. Table 1 shows the mean predicted correlations for the five most frequent combinations, conditional on individual and household covariates in the mean model. The predicted correlations provide insights into the demographic events leading to change in household membership that are associated with the largest changes in unmeasured household influences on health: a low predicted correlation between a pair of households in the same superhousehold suggests that the two households have largely distinct unmeasured characteristics. The most common type of household pair results from an adult child either leaving or entering their parental household (type 1). Pairs of this type have both a partnership link (as they share a couple) and a parent-child link, and on average 65% of their members are in both households. For both physical and mental health, the correlation is highest for these pairs, and higher than the correlation for pairs comprising a couple before and after the birth of a(nother) child (type 3). Even when the households have no one in common, as in the case of separate parent and child households (type 2), there is a moderate correlation between the household random effects, possibly due to shared lifestyle or genetic factors that persist after coresidence ends. Following a partnership breakdown (type 4), there is a moderate correlation between the household effects for the couple household and the household containing one of the ex-partners (and possibly a new partner or other coresidents). The majority of the remaining household pairs (type 5) contain unrelated adults; the two households have neither a partnership nor parent-child link, but nevertheless have individuals in common and therefore a positive correlation.

TABLE 1. Mean predicted between-household random effect correlations and proportion overlap of household members by type of connection between households h_1 and h_2. Results for most frequent patterns on indicators of partnership and parent-child links.

Connection between h_1 and h_2	$\text{Cor}(v_{h_1}, v_{h_2})$		Overlap	%
	Physical	Mental		
1. Child and parent-child hhs i.e. adult child leaves/enters parent hh	0.737	0.700	0.653	14.2
2. Separate parent and child hhs	0.439	0.439	0	12.6
3. Parent and parent-child hhs i.e. couple hh before and after birth	0.633	0.598	0.688	11.8
4. Couple in h_1, 1 partner after split in h_2	0.366	0.331	0.477	11.2
5. Adult sharers, $\geq$ 1 person moves in/out	0.413	0.447	0.249	8.9

References

Goldstein, H., Rasbash, J., Browne, W. J., Woodhouse, G. and Poulain, M. (2000). Multilevel models in the study of dynamic household structures. *European Journal of Population*, **16**, 373 – 387.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**, 677 – 690.

Steele, F., Clarke, P. and Kuha, J. (2019). Modeling within-household associations in household panel studies. *Annals of Applied Statistics*, **13**, 367 – 392.

Zhang, S., Kuha, J., and Steele, F. (2023). Modelling correlation matrices in multivariate data, with application to reciprocity and complementarity in child-parent exchanges of support.*arXiv preprint:2210.14751.*

Zou, T., Lan, W., Li, R. and Tsai, C.-L. (2022). Inference on covariance-mean regression. *Journal of Econometrics*, **230**, 318 – 338.

Advances in modelling dynamic networks

Ernst C. Wit[1]

[1] Università della Svizzera italiana, Switzerland

E-mail for correspondence: wite@usi.ch

Abstract: Many automatic monitoring systems generate big dynamic network data, also called relational data: from invasive species diffusion across the globe (10-100K), bike-sharing rides between bike stations (100K-1M) to patent citations of novel technologies (10M-100M). The aim in analysing these data is typically to discover what drives the interactions to find effective strategies, respectively, to control invasive species, to predict bike sharing at any location at any time, to develop technological innovation.

This short introduction explores the advancements in relational event modelling (REM) within the context of time-stamped relational data, commonly generated by email exchanges and social media interactions, but covering also the applications mentioned above. In a short introduction to REMs, we make the connection to generalized linear models. The GLM connection allows easy generalizations to mixed effect additive REMs, demonstrating how to integrate non-linear specifications and time-varying covariate influences. I will show how emergence effects, such as reciprocity and triadic effects, can be modelled via temporal counter-parts of traditional network statistics. Global covariates, previously challenging in traditional REMs, are addressed, allowing the inclusion of factors such as weather or time-of-day. We derive goodness-of-fit statistics and apply the framework to several interesting and challenging studies.

My overall aim is to show a simple, but rich framework for modelling dynamic networks using techniques that are highly familiar in the statistical modelling community.

Keywords: Relational events; Dynamic network; Event history model; Nested case control sampling; Generalized additive mixed modelling.

1 Introduction

Statistical models for social and other networks are receiving increased attention not only in specialized social science and methodological statistical journals, but also in prominent interdisciplinary science journals such as *Science* and *PNAS*. The increasing availability of time-stamped data resulting from innovation in

TABLE 1. Notation in Relational Event Model

Notation	Meaning
(t, s, r)	relational event: sender s interacts with receiver r at time t
$\lambda_{sr}(t)$	Rate/hazard at which sender s contacts receiver r at time t
$\mathcal{H}_t$	History of process up until time t
L, L_P	Likelihood and partial likelihood
$\mathcal{R}(t)$	Risk set at time t
$\widetilde{\mathcal{R}}(t)$	Sampled risk set at time t
$x_{sr}(t)$	Dyadic covariate(s) with corresponding effect(s) β
$z_{sr}(t)$	Dyadic covariate(s) with corresponding random effect(s) γ
a	Alter, i.e., an individual different from sender or receiver

data production, collection, storage and retrieval technologies has shown that network data samples collected at fixed time intervals are likely to miss fundamental differences in the time scales over which relational processes unfold. Computer-mediated communication, sociometric badges, electronic trading platforms, on-line interaction logs, and video recordings, are just some of the new data-generating technologies capable of producing large quantities of relational event data connecting sender and receiver units.

Since its introduction, the REM has been significantly refined and adapted to an ever-increasing diversity and sophistication of emerging empirical problems (Butts et al., 2023). This overview provides an introduction to relational event modeling in the broader context of statistical models for network science, and assess contemporary methodological, computational, and inferential developments in this class of statistical models for directed social interaction.

2 Specifications of relational event models

The units of analysis in the REM are the edges connecting individual pairs of senders and receivers. Those edges are typically stored in tuples (t, s, r), where s is the sender, r is the receiver, and t is the time of the relational event connecting s to r. At its core, the REM is defined as a point process for directed pairwise interactions that, in turn, are modeled through their rate function λ. The model assumes that λ may depend upon sender, receiver, past event history, and/or exogenous covariates.

2.1 Types of relational event models

We consider a fixed time interval $[0, T]$, with $0 < T < \infty$, in which *events* occur. Events are defined as time-stamped interactions between senders and receivers. Both the set of senders $\mathbf{S}$ and receivers $\mathbf{R}$ are assumed to be finite. For *one-mode networks* the set of senders and receivers overlap, $\mathbf{S} = \mathbf{R}$, whereas for *two-mode* or *bipartite networks* they are distinct. The relational event process is a marked point process for event history sequences $\{(t_i, s_i, r_i) : \; i \geq 1, s_i \subset \mathbf{S}, r_i \subset \mathbf{R}\}$, and defined on a probability space $(\Omega, \mathcal{F}, P)$ adapted to the filtration $\mathcal{H}_t$, consisting

of the history of proces. In principle, the marks (s, r) can be individuals or sets of senders and receivers.
Associated with this marked point process, we define a multivariate counting process N, whose components N_{sr} record the number of directed interactions between s and r,

$$N_{sr}(t) = \sum_{i \geq 1} 1_{\{t_i \leq t;\ s_i = s;\ r_i = r\}}.$$

According to the Doob–Meyer decomposition theorem, there exists a predictable process Λ_{sr} and a residual martingale process M_{sr}, such that the counting process can be written as $N_{sr}(t) = \Lambda_{sr}(t) + M_{sr}(t)$. The aim of the REM is to describe the structure of the predictable cumulative hazard process Λ_{sr}. By assuming that the counting process is an inhomogeneous Poisson process, we can write the cumulative hazard as

$$\Lambda_{sr}(t) = \int_0^t \lambda_{sr}(\tau)\ d\tau,$$

where λ_{sr} is the hazard function of the relational event (s, r). The general REM is defined as

$$\lambda_{sr}(t) = 1_{\{(s,r) \in \mathcal{R}(t)\}}\ \lambda_0(t)\ \exp\left\{\beta^\top x_{sr}(t) + \gamma^\top z_{sr}(t)\right\},$$

where $\lambda_{sr}(t)$ is only non-zero if the event (s, r) is contained in the *risk set* $\mathcal{R}(t)$ of possible events at time t, $\lambda_0(t)$ is the baseline hazard function unrelated to (s, r), $x_{sr}(t)$ and $z_{sr}(t)$ are the $\mathcal{H}_t$ measurable set of endogenous and exogenous (possibly) time-varying variables, and $\beta(t)$ are the effect sizes, whereas γ captures the inherent heterogeneity in the system.

Endogenous vs Exogenous Covariates In statistical models for networks, covariates are endogenous to the extent that they depend on past interaction. Covariates are exogenous when they depend on characteristic of single nodes (monadic covariates) or pairs or nodes (dyadic covariates). One example of endogenous covariate is reciprocity, while gender and geographical distance are exogenous covariates, representing monadic and dyadic characteristics, respectively. An additional consideration refers to the *hierarchy principle*, whereby lower-order interaction terms should always be included in the presence of higher-order interaction terms. In the REM, for example, failing to account for heterogeneity of the senders and receivers may result in incorrect detection of triadic effects (Juozaitiene and Wit, 2024).

Heterogeneity Many social processes possess a large amount of heterogeneity or latent extrinsic variation. This type of heterogeneity in the REM can be captured by means of random effects. Juozaitiene and Wit (2024) and Uzaheta et al. (2023) proposed mixed effect extensions of the REM. Estimation of the random effects variance can be done via Expectation Maximization or Laplace approximations of the likelihood, or in certain cases via a penalized zero order spline approach (Wood, 2017). More recently, in an analysis of a communication network Juozaitiene and Wit (2024) showed that incorporating random effects for both the sender and receiver enhances the model fit compared to model specifications that solely rely on endogenous degree-based statistics. Therefore, the inherent differences between individuals in the network drives part of the heterogeneity in the interactions.

Stratification Conceptually, stratification can be introduced either to model event streams in multiplex networks or to account for heterogeneity by specifying different baseline intensity functions for individual sets of dyads. Perry and Wolfe (2023) and Bianchi and Lomi (2022), for example, use sender-based stratification, effectively allowing each sender to have its own individual baseline hazard. Receiver-based stratification usually occurs when there is heterogeneity in those nodes that are repeatedly targeted as receivers. In citation networks, for example, groundbreaking articles or patents have very distinct, individual citation profiles, which makes a receiver-based baseline hazards an attractive option, given that they not have to be estimated individually (Filippi-Mazzola, 2023). Juozaitiene and Wit (2022) proposed a stratified version of the REM, in which distinct baseline hazards are associated with distinct families of temporal network effects, such as reciprocity in its direct and generalized forms. Subsequent baseline hazard estimation reveals the tendency of some endogenous covariates to have very specific temporal effect-profiles.

3 Estimation and computation

The fundamental information about a sequence of relational events $\{(t_i, s_i, r_i) : i = 1, \ldots, n\}$ is contained in its likelihood function. For a REM, this function can be expressed as the product of the conditional generalized exponential event time densities and their associated multinomial relational event probabilities. Estimating the parameters of REMs by maximizing the full likelihood poses several challenges. The likelihood function is indeed a complex object that involves explicit integration across the unknown hazard function and sums over large risk sets. In this section, we will explore computational alternatives proposed to overcome the complexity of the full likelihood approach.

3.1 Partial likelihood estimation

The Cox model offers an attractive alternative to fully parametric models due to its absence of distributional assumptions regarding activity rates, which are then treated as nuisance parameters. It offers an effective simplification of the full REM likelihood through the application of the partial likelihood L_P to counting processes on network edges, which only involves multinomial event probabilities.This eliminates the unknown baseline hazard, resulting in a more adaptive representation of the underlying network dynamics, while being able to estimate the parameters in a straightforward way by maximizing $L_P(\beta)$. The partial likelihood corresponds to the full likelihood when only the event orderings are known, but not the exact timings. However, the partial likelihood approach faces a limitation in large networks, as the risk set in its denominator tends to expand quadratically with the number of nodes.

3.2 Risk set sampling

The computational bottleneck in the partial likelihood is the sum over the risk set in the denominator. Vu et al. (2015) initially introduced a nested-case control sampling strategy (Borgan and Keogh, 2015) to mitigate the computational complexity involved in estimating the partial likelihood. Nested case-control sampling consists of sampling from the current risk set $\mathcal{R}(t)$ according to some probability $\pi\{\cdot \mid \mathcal{R}(t)\}$ a set of non-events, or controls, for each event, or case. The sampled non-events together with the events are called the sampled risk set $\widetilde{\mathcal{R}}(t)$. Borgan et al. (1995) show that the sampled partial likelihood $\widetilde{L}_P$, accounting for the sampling probabilities is a valid likelihood. When this probability is assumed to be random, i.e., $\pi\{\cdot \mid \mathcal{R}(t)\} = 1/|\mathcal{R}(t) - 1|$, $\widetilde{L}_P(\beta)$ reduces to the simplified form, i.e.,

$$\widetilde{L}_P(\beta) \quad = \quad \prod_{i=1}^{n} \left(\frac{\exp\{\beta^\top x_{s_i r_i}(t_i)\}}{\sum_{(s,r)\in\widetilde{\mathcal{R}}(t)} \exp\left\{\beta^\top x_{sr}(t_i)\right\}} \right),$$

where $\widetilde{\mathcal{R}}(t)$ is the sampled risk set. Lerner and Lomi (2020) employed nested case-control sampling to empirically showcase the efficiency of estimates on large networks, even when a limited number of non-events is sampled.

4 Conclusions

Relational event models are currently undergoing an explosion in development. From the re-definition of the effects, including both time-varying and non-linear formulations, as well as new inferential and computational techniques, REMs have become a flexible tool for analyzing all time of dynamic network processes.

References

Bianchi, F. and Lomi, A. (2023). From ties to events in the analysis of interorganizational exchange relations. *Organizational Research Methods*, **26**, 524–565.

Borgan, Ø. and Keogh R. (2015). Nested case-control studies: Should one break the matching? *Lifetime Data Analysis*, **21**, 517–541.

Borgan, Ø, Goldstein, L. and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the cox proportional hazards model. *The Annals of Statistics*, **23**, 1749–1778.

Butts, C.T., Lomi, A. Snijders, T.A.B. and Stadtfeld, C. (2023). Relational event models in network science. *Network Science*, **11**, 175–183.

Filippi-Mazzola, E. and Wit, E.C. (2023). A stochastic gradient relational event additive model for modelling US patent citations from 1976 until 2022. *arXiv preprint:2303.07961.*

Juozaitienė, R. and Wit, E.C. (2024). Non-parametric estimation of reciprocity and triadic effects in relational event networks. *Social Networks*, **68**, 296–305.

Juozaitienė, R. and Wit, E.C. (2024). Nodal heterogeneity can induce ghost triadic effects in relational event models. *Psychometrika*, **89**, 1–21.

Lerner, J. and Lomi, A. (2020). Reliability of relational event model estimates under sampling: How to fit a relational event model to 360 million dyadic events. *Network Science*, **8**, 97–135.

Perry, P.O. and Wolfe, P.J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **75**, 821–849.

Uzaheta, A., Amati, V. and Stadtfeld, C. (2023). Random effects in dynamic network actor models. *Network Science*, **11**, 249–266.

Vu, D.Q., Pattison, P. and Robins, G. (2015). Relational event models for social learning in moocs. *Social Networks*, **43**, 121–135.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R (2nd ed.)*. CRC Press, Boca Raton.

Part II – Contributed Papers

Modelling narwhal diving behaviour and responses to sound exposure using stochastic differential equations with state-switching coefficients

Timo Adam[1], Mads Peter Heide-Jørgensen[2], Susanne Ditlevsen[3]

[1] Bielefeld University, Germany
[2] Greenland Institute of Natural Resources, Greenland
[3] University of Copenhagen, Denmark

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

Abstract: We propose stochastic differential equations (SDEs) with state-switching coefficients to model narwhal diving behaviour using high-resolution tracking data. For each dive, a non-homogeneous, N-state Markov chain selects which of N possible SDE models determines the depth observations through that dive. Using the proposed model, we show that narwhals exhibit two distinct dive types, namely deep, wiggly and shallow, smooth dives. By modelling the transition probabilities of the Markov chain as smooth functions of sound exposure we further show that, when being exposed to noise, narwhals are less likely to exhibit foraging behaviour.

Keywords: Hidden Markov models; Statistical ecology; Stochastic differential equations, Time series modelling.

1 Introduction

Stochastic differential equations (SDEs) with varying coefficients are popular tools for uncovering mechanistic relationships underlying time series (Michelot et al., 2021). Here, we propose SDEs with state-switching coefficients to model narwhal diving behaviour. Narwhals where outfitted with GPS- and dive-loggers off the Greenlandic coast and tracked through 15,660 dives, which resulted in 3,431,335 1-second observations of their depth. To investigate how narwhals respond to human disturbances, they were exposed to noise generated by a research vessel equipped with an underwater airgun during some dives, where the distance between the narwhals and the research vessel at the beginning of each dive were

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

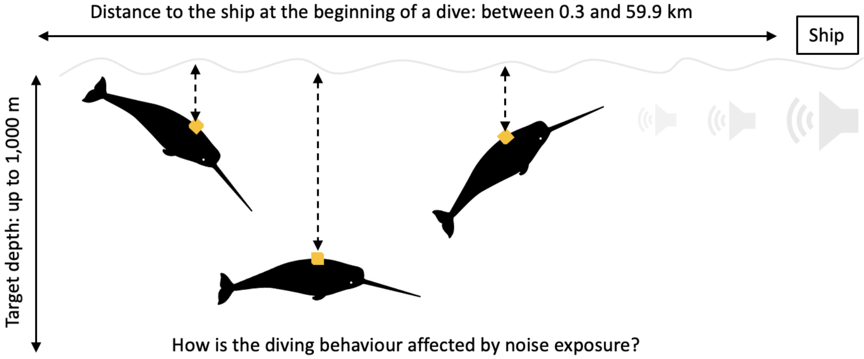

Distance to the ship at the beginning of a dive: between 0.3 and 59.9 km
Ship
Target depth: up to 1,000 m
How is the diving behaviour affected by noise exposure?

FIGURE 1. Illustration of the study design.

obtained from the GPS loggers. The study design, which is explained in detail in Heide-Jørgensen et al. (2021), is illustrated in Figure 1.

2 Methods

State-switching SDEs comprise two stochastic processes, one of which is hidden and the other is observed:

- a hidden state process $\{B_d\}_{d=1,\dots,D}$, which depends on covariates $\mathbf{z}_d$, where d is a dive index and D denotes the number of dives observed;
- an observed state-dependent process $\{Y_{d,t}\}_{d=1,\dots,D,t=1,\dots,T_d}$, which depends on covariates $\mathbf{x}_{d,t}$, where t is the t-th observation within the d-th dive and T_d denotes the number of observations.

The hidden process is modelled by a non-homogeneous, N-state Markov chain with initial distribution $\boldsymbol{\delta} = (\delta_i)$, $\delta_i = \Pr(B_1 = i), i = 1, \dots, N$, and covariate-dependent transition probability matrix (t.p.m.) $\boldsymbol{\Gamma}_d = (\gamma_{i,j}(\mathbf{z}_d))$, $\gamma_{i,j}(\mathbf{z}_d) = \Pr(B_{d+1} = j | B_d = i, \mathbf{z}_d), i, j = 1, \dots, N$. The observed process is modelled by Brownian motion with state- and covariate-dependent drift $r_{d,t}^{(b_d)}$ and diffusion $s_{d,t}^{(b_d)}$, where

$$r_{d,t}^{(b_d)} = \beta_{0,r}^{(b_d)} + f_r^{(b_d)}(\mathbf{x}_{d,t});$$
$$\log\bigl(s_{d,t}^{(b_d)}\bigr) = \beta_{0,s}^{(b_d)} + f_s^{(b_d)}(\mathbf{x}_{d,t}),$$

with $f_r^{(b_d)}$ and $f_s^{(b_d)}$ being (potentially) smooth functions of the covariates. The drift can be interpreted as the average change of the process over some small time interval, whereas the diffusion can be interpreted as its average variability. The dependence structure is illustrated in Figure 2. Model fitting is carried out by numerical likelihood maximisation using some Newton-Raphson-type optimisation routine, where the likelihood is evaluated using the forward algorithm (Zucchini et al., 2016), where

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

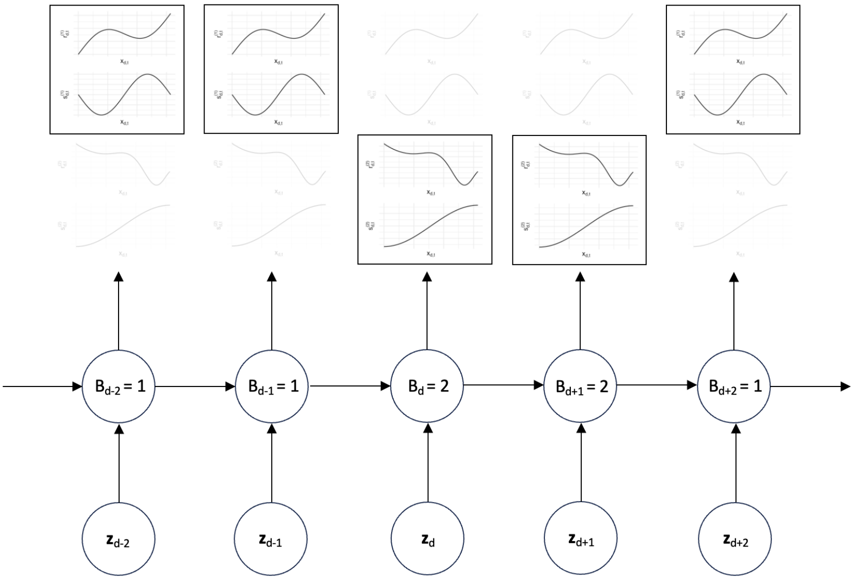


FIGURE 2. Dependence structure in state-switching SDEs.

with $\mathbf{P}(\mathbf{y}_d) = \text{diag}(f(\mathbf{y}_d|B_d = 1), \ldots, f(\mathbf{y}_d|B_d = N))$ and $f(\mathbf{y}_d|B_d = i)$ being the transition density of all depth observations within the d-th dive under the i-th SDE model. For Brownian motion, the transition density is given by

$$f(\mathbf{y}_d|B_d = i) = \prod_{t=2}^{T_d} \left(\frac{1}{s_{d,t}^{(i)}\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y_{d,t} - r_{d,t}^{(i)}}{s_{d,t}^{(i)}} \right)^2 \right) \right).$$

In principle, it would also be possible to consider other processes than Brownian motion (e.g., Ornstein-Uhlenbeck processes). Then, the above transition density has to be replaced by the corresponding transition density of that process.

3 Results

To model the narwhals' depth through a dive, we model drift and diffusion as smooth functions of the proportion of time through that dive (i.e., $\text{Dive\_prop}_{d,t}$). The transition probabilities of the Markov chain are modelled as smooth functions of the exposure level, which is defined as $\text{Exposure\_level}_d = 1/\text{Distance\_to\_noise}_d$ (i.e., the closer the research vessel, the higher the exposure level).

The estimated drift and diffusion for all individuals are displayed in Figure 3. State 1 (blue) is associated with deep, wiggly dives, which can be interpreted as foraging behaviour. State 2 (red) is associated with shallow, smooth dives, which is associated with resting or travelling behaviour.

The t.p.m.s for one individual and three different exposure levels (i.e., 0, 0.1, and

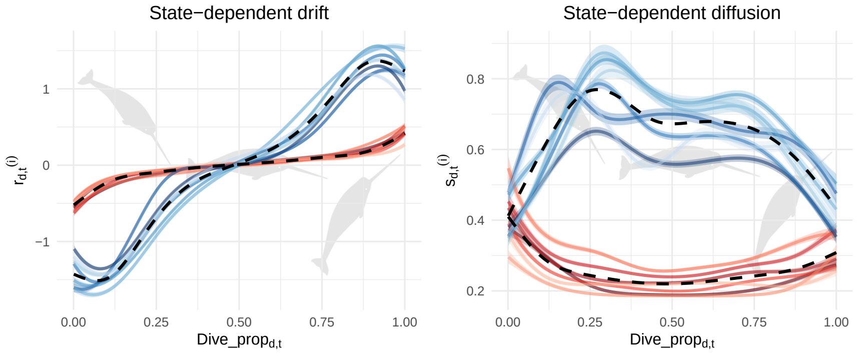


FIGURE 3. Estimated drift and diffusion through a dive.

0.2) were estimated as

$$\underbrace{\hat{\mathbf{\Gamma}}_d = \begin{pmatrix} 0.495 & 0.505 \\ 0.148 & 0.852 \end{pmatrix}}_{\text{Exposure\_level}_d = 0};\ \underbrace{\hat{\mathbf{\Gamma}}_d = \begin{pmatrix} 0.142 & 0.858 \\ 0.078 & 0.922 \end{pmatrix}}_{\text{Exposure\_level}_d = 0.1};\ \underbrace{\hat{\mathbf{\Gamma}}_d = \begin{pmatrix} 0.027 & 0.973 \\ 0.042 & 0.958 \end{pmatrix}}_{\text{Exposure\_level}_d = 0.2}.$$

The stationary distributions $(0.227, 0.743)$, $(0.084, 0.916)$, and $(0.041, 0.959)$ indicate that, without being exposed to noise, the narwhal spends, on average, 22.7 % of the dives in state 1 (i.e., foraging). When the research vessel is 10 km (5 km) away, then this figure drops to 8.4 % (4.1 %).

4 Discussion

We proposed SDEs with state-switching coefficients to model narwhal diving behaviour using high-resolution tracking data. Using the proposed mo- del, we showed that narwhals exhibit two distinct dive types, namely deep, wiggly and shallow, smooth dives, and that they are less likely to exhibit foraging behaviour when being exposed to noise, indicating that human disturbances can have severe ecological consequences.

References

Heide-Jørgensen, M.P., Blackwell, S.B., Tervo, O.M., Samson, A.L., Gar- de, E., Hansen, R.G., Ngô, M.C., Conrad, A.S., Trinhammer, P., Schmidt, H.C., Sinding, M.-H.S., Williams, T.M. and Ditlevsen, S. (2021). Behavioral response study on seismic airgun and vessel exposures in narwhals. *Frontiers in Marine Science*, **8**, 658173.

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**, 446 – 463.

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.

A joint model for multiple (un)bounded longitudinal markers, competing risks, and recurrent events

Pedro Miranda Afonso[1], Dimitris Rizopoulos[1], Rhonda D. Szczesniak[2], Eleni-Rosalina Andrinopoulou[1]

[1] Department of Biostatistics, Erasmus MC, the Netherlands
[2] Cincinnati Children's Hospital MC, Biostatistics & Epidemiology, USA

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

Abstract: Motivated by a clinical study on cystic fibrosis, we propose a Bayesian shared-parameter joint model that simultaneously accommodates multiple (possibly bounded) longitudinal markers, a recurrent event process, and competing risks. The model allows for various forms of association, discontinuous risk intervals, and both gap and calendar timescales. We analyse the US Cystic Fibrosis Foundation Patient Registry to study the associations between lung function decline (ppFEV$_1$), cumulative changes in BMI, and the risk of recurrent pulmonary exacerbations, while accounting for the competing risks of death and lung transplantation. Acknowledging ppFEV$_1$ as a bounded marker, we use the beta distribution to prevent biologically implausible values without sacrificing the interpretability of its associations, whereas BMI is modelled using the Gaussian distribution. Our efficient implementation allows fast fitting of the model despite its complexity and the large sample size. Our comprehensive approach provides new insights into cystic fibrosis progression by quantifying the relationship between the most important clinical markers and events more precisely than has been possible before. The model is available in the R package JMbayes2.

Keywords: Bounded outcomes; Competing risks; Joint model; Multivariate longitudinal data; Recurrent events.

1 Introduction

Joint models have become a popular framework in health research for studying longitudinal markers and their association with clinical events. However, integrating recurrent events and competing risks into a unified model remains a challenge, leading researchers to omit important information from their analyses. Additionally, most existing frameworks rely on Gaussian distributions to model

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

continuous markers. An important aspect of joint modelling is the appropriate parameterization of longitudinal submodels to ensure accurate extrapolation of unobserved marker evolution up to the event time. A Gaussian parameterization can be problematic for a bounded marker with many observations close to the boundaries, as it can cause the model to yield biologically implausible values, resulting in biased estimates of the marker evolution and its associations.

Cystic fibrosis (CF) is a severe genetic disorder that primarily affects the lungs and digestive system, and is characterized by recurrent pulmonary exacerbations (PEx) that can lead to permanent lung damage and increased risks of death or lung transplantation. The body mass index (BMI) and the percentage of predicted forced expiratory volume in one second (ppFEV$_1$) are routinely measured to monitor disease progression. CF care teams are interested in understanding the associations between ppFEV$_1$ decline, BMI changes, recurrent PEx, and the competing risks of death and lung transplantation using the US Cystic Fibrosis Foundation Patient Registry (CFFPR). The lack of an appropriate framework has hampered previous studies that aimed to investigate such associations using joint models. For example, Andrinopoulou et al. (2020) limited their analysis to the period up to the first PEx event, disregarding subsequent occurrences and informative censoring due to transplantation or death. Moreover, existing CF studies have modelled ppFEV$_1$ exclusively using a Gaussian distribution, resulting in predictions outside the feasible range (see Figure 1).

To overcome these challenges, we propose a comprehensive joint modelling framework capable of (i) accommodating competing risk and recurrent event processes alongside multiple longitudinal outcomes, and (ii) appropriately modelling bounded longitudinal markers using a beta distribution.

2 Methods

We propose the following shared-parameter joint model:

$$\begin{cases} \operatorname{logit}\{\mu_{1,i}(t)\} = \eta_{1,i}(t) = \boldsymbol{x}_{1,i}(t)^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_{1,i}(t)^{\mathrm{T}}\boldsymbol{b}_{1,i} & \text{ppFEV}_1 \\ y_{2,i}(t) = \mu_{2,i}(t) + \varepsilon_i(t) = \eta_{2,i}(t) + \varepsilon_i(t) = \boldsymbol{x}_{2,i}(t)^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_{2,i}(t)^{\mathrm{T}}\boldsymbol{b}_{2,i} + \varepsilon_i(t) & \text{BMI} \\ h_{1,i}(t) = h_{0,1}(t)\exp\left[\boldsymbol{w}_{1,i}^{\mathrm{T}}\boldsymbol{\gamma}_1 + \frac{\mathrm{d}}{\mathrm{d}t}\mu_{1,i}(t)\alpha_{1,1} + \frac{1}{t}\int_0^t \eta_{2,i}(s)\,\mathrm{d}s\ \alpha_{1,2} + \upsilon_i\right] & \text{PEx} \\ h_{2,i}(t) = h_{0,2}(t)\exp\left[\boldsymbol{w}_{2,i}^{\mathrm{T}}\boldsymbol{\gamma}_2 + \frac{\mathrm{d}}{\mathrm{d}t}\mu_{1,i}(t)\alpha_{2,1} + \frac{1}{t}\int_0^t \eta_{2,i}(s)\,\mathrm{d}s\ \alpha_{2,2} + \upsilon_i\,\alpha_{2,\upsilon}\right] & \text{Transp.} \\ h_{3,i}(t) = h_{0,3}(t)\exp\left[\boldsymbol{w}_{3,i}^{\mathrm{T}}\boldsymbol{\gamma}_3 + \frac{\mathrm{d}}{\mathrm{d}t}\mu_{1,i}(t)\alpha_{3,1} + \frac{1}{t}\int_0^t \eta_{2,i}(s)\,\mathrm{d}s\ \alpha_{3,2} + \upsilon_i\,\alpha_{3,\upsilon}\right] & \text{Death} \end{cases}$$

where $i = 1, \ldots, n$ represent individuals, and $(\boldsymbol{b}_{1,i}, \boldsymbol{b}_{2,i})$, ε_i, and υ_i are Gaussian random variables assumed independent of each other. To describe the individual-specific time evolution of ppFEV$_1$ and BMI, we specify two linear mixed-effects models with a beta and a Gaussian distribution, respectively. The beta distribution ensures that the ppFEV$_1$ value is bounded. The terms $\boldsymbol{x}_{\cdot,i}(t)$ and $\boldsymbol{z}_{j,i}(t)$ are the design vectors for the fixed effects $\boldsymbol{\beta}_{\cdot}$ and the random effects $\boldsymbol{b}_{j,i}$. We assume a non-linear evolution over time for both markers, modelled using natural cubic splines. The longitudinal outcomes are associated through the covariance matrix $\boldsymbol{D}$.

For the event processes, we rely on three proportional hazard risk models. We use penalized B-splines to define flexible baseline hazards $h_{0,k}\,(t)$. The design vector $\boldsymbol{w}_{k,i}$ is the parameter vector of measured characteristics with a corresponding vector of regression coefficients $\boldsymbol{\gamma}_k$.

Our model allows the specification of various functional forms to link the longitudinal and event processes. In particular, we include the rate of change of ppFEV$_1$ evaluated at its original scale (rather than the linear predictor scale), $\mathrm{d}\mu_{1,i}(t)/\mathrm{d}t$, where $\mu_{1,i}(t) = 1/\left[1 + \exp\left\{-\eta_{1,i}(t)\right\}\right]$, and the standardized cumulative effect of BMI's underlying value, $\frac{1}{t}\int_0^t \eta_{2,i}(s)\,\mathrm{d}s$. The recurrent and terminal processes are correlated through the common frailty term υ_i. The magnitude of the association between each pair of processes is quantified by $\alpha_{k,j}$ and $\alpha_{k,\upsilon}$. Regarding the recurrent event process, our model accommodates both gap and calendar timescales, as well as the non-risk period during the occurrence of a PEx.
The model has been made available in the CRAN R package JMbayes2 (Rizopoulos et al., 2023). The full implementation of the Markov chain Monte Carlo algorithms in C++ allows for fast fitting of the model despite its complexity and the CFFPR's large sample size. The computational expense of model fitting has been a major problem in previous analyses.

3 Results

Figure 1 shows the estimated evolution of BMI and ppFEV$_1$ with age. The estimates in Table 1 suggest that both ppFEV$_1$ and BMI are associated with the risks of experiencing PEx, transplantation, and death. A ten-unit increase in the rate of ppFEV$_1$ decline increases the hazard of PEx by 14.69% (95% CI 13.09–14.69%). A one-unit increase in the standardized cumulative effect of BMI increases the hazard of PEx by 13.80% (95% CI 13.05–14.41%). The incidence of PEx is positively associated with transplantation and death. Frailer individuals are at a higher risk of PEx and are more likely to receive a lung transplant or die. A one-SD frailty increase raises the hazard of transplantation and death by 290.74% (95% CI 264.96–317.43%) and 229.95% (95% CI 211.98–247.93%), respectively.

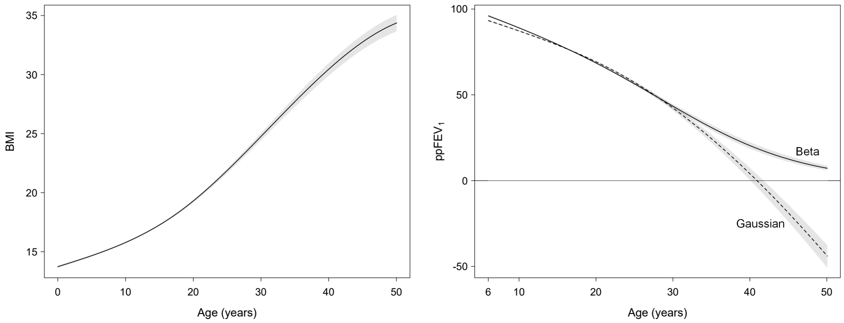


FIGURE 1. Left: estimated BMI evolution with age, with associated 95% credible interval. Right: estimated ppFEV$_1$ evolution with age, with associated 95% credible interval, when assuming either a beta or Gaussian distribution. For a Gaussian distribution, the model generates non-feasible negative values. The ppFEV$_1$ values begin at age six due to the difficulty of obtaining accurate measurements in young children.

TABLE 1. Estimated posterior means and 95% credible intervals for the association parameters in the proposed joint model.

Risk model	log HR	Mean	95% CI
Recurrent PEx	ppFEV$_1$, $\alpha_{1,1}$	-2.38	$(-2.66, -2.10)$
	BMI, $\alpha_{1,2}$	0.13	(0.12, 0.13)
Transplantation	ppFEV$_1$, $\alpha_{2,1}$	-1.94	$(-3.78, -0.06)$
	BMI, $\alpha_{2,2}$	0.04	(0.01, 0.07)
	PEx, $\alpha_{2,\upsilon}$	1.25	(1.18, 1.31)
Death	ppFEV$_1$, $\alpha_{3,1}$	-5.88	$(-7.20, -4.52)$
	BMI, $\alpha_{3,2}$	0.04	(0.01, 0.06)
	PEx, $\alpha_{3,\upsilon}$	1.09	(1.04, 1.14)

CI: credible interval; HR: hazard ratio; PEx: pulmonary exacerbation.

4 Conclusion

Our findings shed new light on the progression of CF, and we hope they will contribute to the effective management of PEx, reducing the frequency and severity of episodes. By making our model publicly available in JMbayes2, we hope to assist others in performing joint analyses of longitudinal and time-to-event data in other complex settings.

References

Andrinopoulou, E.-R., Clancy, J. P. and Szczesniak, R. (2020). Multivariate joint modeling to identify markers of growth and lung function decline that predict cystic fibrosis pulmonary exacerbation onset. *BMC Pulmonary Medicine*, **20**, 1–11.

Rizopoulos, D., Papageorgiou, G. and Miranda Afonso, P. (2023). JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data. R package version 0.4-5, https://CRAN.R-project.org/package=JMbayes2.

Statistical detection of barriers in primary care access for young victims of gender-based violence

Víctor Alonso-Lara[1], Amanda Fernández-Fontelo[2], Pere Toran[3], Meritxell Gómez-Maldonado[3], Gemma Falguera[3], David Moriña[1]

[1] Department of Econometrics, Statistics and Applied Economics, Universitat de Barcelona, Spain
[2] Departament de Matemàtiques, Universitat Autònoma de Barcelona, Spain
[3] Institut Universitari d'Investigació en Atenció Primària (IDIAP Jordi Gol), Institut Català de la Salut, Spain

E-mail for correspondence: `valonsolara@ub.edu`

Abstract: Primary health care system in Catalonia (Spain) aims to be the reference system for detecting gender-based violence cases. A remarkable effort has been done in the last years in order to train professionals to improve their proficiency in detecting cases. However, some specific subpopulations seem to be more difficult to reach for different reasons, one of them being the youngest women (under 20 years old) victims of gender-based violence, who are still reluctant to seeking professional health after suffering an assault. This work suggests that this is the case by comparing the different impact of age in the probability of suffering gender-based violence on the basis of data sources from primary care attention and from the general population.

Keywords: Gender-based violence; Primary care; Random forest; Support vector machines; Neural network.

1 Introduction

Gender-based violence (GBV), that is, violence directed against a person because of their gender or violence that affects persons of a particular gender disproportionately, can take various forms, including, most notoriously, domestic violence. According to the United Nations, GBV refers to harmful acts directed at an individual based on their gender. It is rooted in gender inequality, and might adopt different forms: physical, sexual, emotional, financial or structural, and

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

the victim can require medical assistance after an episode of GBV or not. Here, we are concerned specifically with physical or sexual violence directed against adult women, who are, jointly with girls, the main victims of GBV. According to the World Health Organization (WHO), 30% of women worldwide have suffered either physical and/or sexual violence at some point in their lives (WHO, 2018), and tackling this issue is one of the specific development goals of their 2030 agenda. It is clearly also a significant healthcare policy concern, as victims of sexual and physical GBV are more likely to require health services than the general population to address the physical, gynecological and psychological consequences of the aggressions suffered (in Catalonia, the average number of Primary Care visits of women who haven't suffered GBV is 8 for 13 among GBV victims as reported in Generalitat de Catalunya (2019)). In this work, we will focus on GBV cases in which the victims required assistance from the public health primary care system in one of the most populated areas in Catalonia, Spain, in the period 2009-2019.

The hypothesis is that the age distribution among GBV victims detected in the primary care system is different than the distribution of age among GBV victims identified through two surveys conducted by the Spanish government in 2015 and 2019. One of the most relevant differences is that the anonymous surveys reveal a relevant number of GBV victims between 16 and 25 years old, while a very small proportion of these victims seek medical attention. The reasons behind these differences are many, from the self perception of younger population about their maturity on the decision-making process regarding their health and well-being in general, to the feeling of lack of support from police, judicial statements and health professionals. These reasons are additional to the denial of sexual violence related to guilt often experienced by GBV victims (Toledo-Vásquez and Pineda-Lorenzo, 2016).

Combating GBV and promoting gender equality remain critical priorities on the policy agendas of both the Spanish and Catalan governments. Since the early 2000s, Spain has launched comprehensive initiatives to tackle GBV, which include conducting awareness campaigns, establishing a Ministry for Equality, and creating specialized courts for GBV cases as described in García-Hombrados and Martínez-Matute (2022). Furthermore, under Spanish Organic Law 1/2004 on Integrated Protection Measures against Gender Violence, healthcare professionals are mandated to be vigilant for signs of GBV during patient interactions (Otero-García et al., 2018). Primary healthcare providers are encouraged to handle these situations with due diligence and collaborate across multiple disciplines, ensuring a coordinated response with various agencies and sectors. The significant role of healthcare workers in combating GBV is underscored by their direct and frequent contact with patients.

2 Methods

Our study draws on three sources of data:

- A random sample of 6,556 women aged 16 or over assigned to *Àmbit Metropolità Nord* of the Barcelona health region, of which 3,484 had a diagnosis of GBV from the Primary Care system between January 2010 and December 2021. These data include information on age, nationality and number of children.

- Survey on violence against women conducted by the Spanish Ministry of Equality in 2015. The main objective of this survey was to determine the percentage of women aged 16 or over residing in Spain that have been the victims of any type of GBV. The interviews were conducted with a representative sample of 10,171 women.
- Survey on violence against women conducted by the Spanish Ministry of Equality in 2019. The interviews were conducted with a representative sample of 9,568 women.

Three supervised learning classification methods are considered in this work: Random forest, support vector machine and neural network. In each case, the binary outcome having suffered physical or sexual gender-based violence is modelled using age of the victim, nationality (Spanish or foreign) and children (having any or not) as features. The contribution of age in the estimated probability of suffering GBV in each data source is estimated by means of partial dependence (PD) plots (Friedman, 2001), which allow to visualize global feature effects by visualizing how model predictions change on average when varying the values of a given feature of interest. The results of the three considered methods were similar, so only results corresponding to random forests are reported here, as this was the method that showed a higher accuracy.

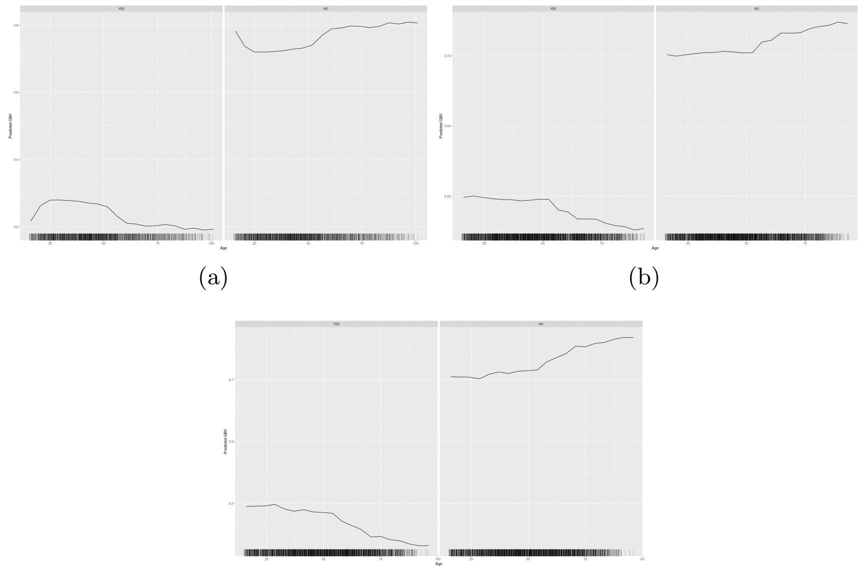
3 Results

As can be seen in Table 1, the characteristics of the samples from primary care and from the surveys are very similar except for an overrepresentation of foreign women in primary care data.

TABLE 1. Median (IQR) or percentage of features by data source (surveys 2015 and 2019 and primary care data (PC)).

Source	Feature	Global	Victim	No victim
Surv. 2015	Age	47 (29)	43 (24)	48 (30)
	Children (Yes)	73.8%	71.1%	74.5%
	Spanish	92.9%	88.7%	94.0%
	Women	9,952	2,128	7,824
Surv. 2019	Age	49 (48)	44 (26)	51 (50)
	Children (Yes)	72.2%	67.9%	73.7%
	Spanish	91.8%	89.2%	92.6%
	Women	9,568	2,412	7,156
PC	Age	45 (27)	43 (22)	50 (31)
	Children (Yes)	72.9%	73.6%	71.4%
	Spanish	71.7%	66.3%	80.0%
	Women	5,792	3,119	2,673

Both 2015 and 2019 surveys suggest that the higher probability of suffering GBV is between 16 and almost 50, as displayed in Figure 1. However, in the primary



(a)

(b)

(c)

FIGURE 1. Partial dependence plots for age impact on the probability of suffering GBV in each of the available data sources: (a) primary care data, (b) survey 2015 and (c) survey 2019.

care data, this probability is increasing from 16 to 25 and then nearly constant until 50. These different profiles suggest a remarkable underdetection of GBV cases in youngest women in the primary care data that could be even larger than among the general population where this issuee is indeed very perturbing, as discussed in Moriña et al (2024).

4 Discussion

The primary health care systems in developed countries should take a central role in identifying cases of GBV, as nearly all women interact with these systems at some point for sexual and reproductive health care. However, despite this, they are often not where GBV is most frequently detected (Muñoz-Sellés et al., 2023). This oversight is due to multiple factors, including personal issues such as shame, fear of retaliation, and economic dependence, as well as societal factors like gender power imbalances, the sanctity of family privacy, and attitudes that blame the victim (Gracia, 2004). These elements are deeply stigmatized and are often rooted in longstanding cultural and religious traditions. A particular concern is the rising incidence of GBV among younger women, with research indicating that the age at which women experience victimization is dropping, and that young people are increasingly subjected to violence in romantic relationships (Racionero-Plaza et al., 2021). The present study advocates for targeted measures to facilitate access for young GBV victims to primary health care services. Despite the age-related findings of various supervised learning methods showing a higher estimated GBV risk between ages 16 and 45, actual detection rates for those aged 16-25 in primary care are notably low. Therefore, there is a clear need to develop specific strategies to improve access and enhance detection capabilities within primary health care for the youngest victims of GBV.

Acknowledgments: This study has been funded by the Social Observatory of the "la Caixa" Foundation (LCF/PR/SR22/52570005). A.F-F acknowledges Agencia Estatal de Investigación for the financial support IJC2020-045188I/AEI/10.13039/501100011033 and a María Zambrano scholarship. This work was partially funded through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020–001084-M), by grant PID2022-137414OB-I00 fom the Spanish Government through the Ministerio de Ciencia e Innovación and by the Agència de Gestió d'Ajudes Universitàries i de Recerca (2021SGR00299).

References

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, **29**, 1189 – 1232.

García-Hombrados, J. and Martínez-Matute, M. (2021). Specialised courts and the reporting of intimate partner violence: Evidence from Spain. *IZA Discussion Papers*, No. 14936.

Generalitat de Catalunya (2019). Avaluació del protocol per a l'abordatge de la violència masclista en l'àmbit de la salut a Catalunya.

Gracia, E. (2004). Unreported cases of domestic violence against women: towards an epidemiology of social silence, tolerance, and inhibition. *Journal of Epidemiology & Community Health*, **58**, 536 – 537.

Moriña, D., Millán, I., Fernández-Fontelo, A., Puig, P., Toran, P., Gómez Maldonado, M. and Falguera, G. (2024). Exploring what lies beneath the tip of the gender-based violence iceberg. *medRxiv preprint.*

Muñoz-Sellés, E., Pujolar-Díaz, G., Fuster-Casanovas, A. and Miró Catalina, Q. (2023). Detection of gender-based violence in primary care in central Catalonia: a descriptive cross-sectional study *BMC Health Services*, **23**, 1 – 9.

Otero-García, L., Briones-Vozmediano, E., Vives-Cases, C., García-Quinto, M., Sanz-Barbero, B. and Goicolea, I. (2018). A qualitative study on primary health care responses to intimate partner violence during the economic crisis in Spain. *European Journal of Public Health*, **28**, 1000 – 1005.

Racionero-Plaza, S., Tellado, I., Aguilera, A. and Prados, M. (2021). Gender violence among youth: an effective program of preventive socialization to address a public health problem. *AIMS Public Health*, **8**, 66 – 80.

Toledo-Vásquez, P. and Pineda-Lorenzo, M. (2016). L'abordatge de les violències sexuals a Catalunya. Part 1. Marc conceptual sobre les violències sexuals. *Generalitat de Catalunya.*

Monitoring viral infections in severe acute respiratory syndrome patients in Brazil

João Flávio Andrade Silva[1], Rafael Izbicki[2], Leonardo S. Bastos[3], Guilherme P. Soares[4]

[1] Interinstitutional Graduate Program in Statistics UFSCar-USP (PIPGEs), Federal University of São Carlos and University of São Paulo, São Carlos, Brazil
[2] Department of Statistics, Federal University of São Carlos, São Carlos, Brazil
[3] Scientific Computing Program, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil
[4] Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil

E-mail for correspondence: joaoflavio@usp.br

Abstract: We introduce a novel methodology for estimating the distribution of viruses in Severe Acute Respiratory Syndrome (SARS) patients in Brazil, addressing significant challenges for data in that country, such as data delays and the absence of negative test results. By employing a probabilistic classifier, our approach offers precise, adaptable estimates across various demographic characteristics and regions of the country without the need for predefined groups. Comparative analyses demonstrate the effectiveness of the model. This methodology significantly contributes to public health by enhancing disease monitoring and supporting targeted prevention strategies.

Keywords: Severe acute respiratory syndrome; Epidemiological modeling; Machine learning; Virus distribution; Public health surveillance.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Empirical study of the Sardex network

Sébastien Appleby[1], George Iosifidis[2], Arthur White[1]

[1] Trinity College Dublin, Ireland
[2] Delft University of Technology, Netherlands

E-mail for correspondence: applebys@tcd.ie

Abstract: Sardex is a regional currency designed to facilitate economic development and social innovation. Businesses and companies that trade in Sardex form an economic network that favours the development of local activity and enables companies to use a credit line guaranteed by the network. Our aim is to explore how this network is structured and how it grows. We have collected a comprehensive 10-year dataset that gives us many opportunities to conduct an empirical study of the network. We will see how the network becomes more densely connected as it welcomes new members. Beyond providing us the ability to increase our understanding of Sardex, the study opens up new perspective to conduct future modelling analyses.

Keywords: Social network; Growth; Directed Graph, Undirected Graph.

1 Introduction

Regional currencies are perceived as an alternative way to promote local businesses. Our study will focus on a pioneering network named Sardex, initially started in Sardinia in 2010 but which rapidly seeded similar initiatives throughout Italy. Sardex is both a complementary currency and a network that was developed to encourage companies to trade with local businesses. The community currency concept is based on an idea originally proposed by Steve Gesell (1862-1930) which seeded several experiments in Europe after that time.
Sardex aims to enable businesses to exchange goods and services through a centrally managed compensation platform that allocates an initial balance to its members. This balance is then updated given all the transactions taking place between its members. By design, Sardex is restricted to Sardinia which makes it a local network, favouring local businesses. Among the main differences between Sardex and a cryptocurrency is that its conversion ratio is fixed and is pegged to the official currency (euro) and is centrally administered. Moreover, when opening an account with Sardex, the client company is awarded a credit line based on its revenue, allowing the company to spend some units even before having sold

any goods or services. If the negative balance of the Sardex account exceeds its limit, the company will not be able to buy more goods, but instead needs to sell in order to pay down the balance.

As such, Sardex plays a mixed role between a marketplace and a credit institution and since, contrary to cryptocurrencies, there is no exchange rate, it eliminates risk for business owners. Due to its special position, the Sardex network is of economic interest because of the role it plays within the local Sardinian economy, but it also motivates interesting questions in network theory. We can consider Sardex as a collection of nodes, represented by businesses that initiated at least one transaction, and edges, which are represented by the transactions between members. This network features longitudinal, weighted edges, with new nodes joining the network throughout.

From an economic standpoint, as Sardex is a zero sum network, the credit lines allowed to companies correspond to the total amount of cash accumulated by the companies with a positive balance. This is one of the main reasons why Sardex has been so successful as shown by the dataset. It also raises questions regarding how key characteristics of the network have changed as it grows over time.

2 Dataset and network growth

We analyse a full 10-year dataset of the Sardex network recording all interactions since the inception of the network in 2010 until 2020. Because all transactions are centrally managed and registered, the Sardex dataset offers a great perspective to study the relationships among its members and how the network evolves over time. Beginning in 2010, the network gathers a total of 5,373 nodes by 2020. Over the 10 year period, the number of transactions amounts to 650,242, representing a global exchanged amount close to 162 million Euros.

If we consider Sardex as a directed network where a transaction between two nodes represents an edge and a reverse transaction constitutes another edge, the transactions observed between 2010 and 2020 initiated 122,451 directed edges. On the contrary, if we consider Sardex as an undirected graph (a single transaction creates an edge, no matter which node is the buyer or the seller) the network totals 109,867 edges.

Figure 1 shows the increase in the number of nodes per month in the Sardex network (Fig 1.a) during this time frame and gives an indication of the significant growth that the network has witnessed over time.

3 Methodology

We describe the behaviour of the Sardex network using several metrics. Each metric describes specific aspects of the network, and collectively represent a broader understanding of network behaviour. Unless otherwise stated, larger values for these metrics indicate more highly connected networks. Selected metrics were calculated using custom python classes, based on the Networkx python package. We applied the following metrics to an undirected representation of the Sardex graphs: average directed path length, which represents the average length of the sequence of pair-wise directed ties between any two nodes, where larger values indicate less well connected networks; average degree, where the degree of a node

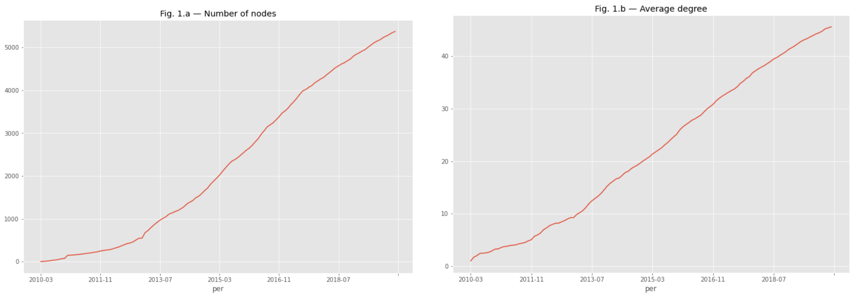


FIGURE 1. Nodes (1.a) and average degree (1.b) per month (2010 – 2020).

is defined as the total number of edges connected to it, and the average degree of the network is then simply the total number of edges in the network divided by its number of members; network diameter, which is the longest path existing between two members of the network, assuming that the network is connected; and clustering coefficient, which represents the number of existing triads (or three-way relationship) among all possible triads at the graph level.

4 Results

To give a few examples, as of Jan 2020, the average degree per Sardex member was equal to 40.76. The average directed path length was approximately 3.5 (median=3; s.d.=0.9) . The diameter of the network is at 10. In Jan 2020, the clustering coefficient is 0,19. The Average Degree Centrality represents the average of the fraction of the existing nodes a specific node is connected to and is 7.6 10^{-3}.
Following the study led by Iosifidis (2018) where Sardex was analysed between 2013 and 2014, we can make comparisons with the latest data we have for certain metrics as shown in Table 1.

TABLE 1. Compared metrics (2014 – 2020)

Type	2014	2020
Average directed path length	3.50	2.90
Average degree	18.60	40.76
Diameter	10	10
Clustering coefficient	0.14	0.19
Average degree centrality	8.5 10^{-3}	7.6 10^{-3}

These numbers show that, even though it is gaining a significant number of new members, the network is becoming more connected over time as shown on Fig1.b. This probably reflects the ability of the network to create bonds between its members.

5 Future work and conclusion

We have presented a descriptive analysis of the Sardex network. Due to its depth and the fact that it covers a 10 year period, this dataset motivates a number of interesting future analyses across several fields of research. Among many possible models which could be applied to this data, the stochastic blockmodel and is of considerable future interest. This model has the capacity to identify key features of study, such as communities of Sardex members with similar behaviour characteristics, key members who engage in the highest levels of trading, or with the widest number of network members. It would also be interesting to model how key parameters, such as relative community sizes, or network interactions, change as the network evolves.

References

Iosifidis, G., Charette, Y. and Airoldi, et al. (2018). *Cyclic Motifs in the Sardex Monetary Network.* Nature Publishing Group, Nature Human Behaviour.

Newman, M.E.J. (2010). *Networks: An Introduction.* Oxford University Press.

Networks analysis in Python : *https://networkx.org.*

Sardex official website : *https://www.sardexpay.net.*

Adaptive generalized logistic lasso and its application to rankings in sports

Robert Bajons[1], Kurt Hornik[1]

[1] Institute of Statistics and Mathematics, Vienna University of Economics and Business, Austria

E-mail for correspondence: robert.bajons@wu.ac.at

Abstract: The generalized lasso is a popular model for ranking competitors, as it allows for implicit grouping of estimated abilities. In this work, we present an implementation of an adaptive variant of the generalized lasso penalty for logistic regression using conic programming principles. This approach is flexible, robust, and fast, especially in a high-dimensional setting. The methodology is applied to sports data, with the aim of ranking soccer players based on their contribution to possession sequences.

Keywords: Generalized lasso; Logistic regression; Rankings.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

Regression analysis with missing data using interval imputation

Tathagata Basu[1]

[1] University of Strathclyde, Glasgow, UK

E-mail for correspondence: `tathagata.basu@strath.ac.uk`

Abstract: Regression analysis with missing data is a common problem in statistical modelling. Majority of the available methods use point imputation strategy to get rid of the missing entries. However, such methods rely on different observational assumptions. In this paper, we propose a novel approach based on interval imputation, that is, instead of a single value imputation, we replace the missing entries with the range of the variables obtained from the observational data. This way, we avoid any distributional assumption on the data and formulate our model based only on the information in hand. For estimation, we rely on the interval matrix algebra. We also introduce regularisation terms with Bayesian analysis which also allows us to incorporate our subjective belief and avoid singularity in the estimation process similar to ridge regression. We evaluate the maximum a posteriori estimates of these regression coefficients to obtain the approximate posterior bounds. Once we have these posterior bounds, we use cross validation to obtain mixing parameters between the lower and upper bounds of the posterior estimates for model fitting. Finally, we illustrate our method with real-life dataset and compare with other state of the art methods to showcase our methods applicability.

Keywords: Bayesian analysis, Linear regression, Missing data, Interval arithmetic.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Modeling and predicting injuries in soccer with machine learning and conventional statistical approaches

Ina-Marie Berendes[1], Alexander Gerharz[1,2], Andreas Groll[1], Mathias Kolodziej[2]

[1] Department of Statistics, TU Dortmund University, Germany
[2] Department of Sport Science, Borussia Dortmund, Germany

E-mail for correspondence: `ina-marie.berendes@tu-dortmund.de`

Abstract: To prevent injuries in professional soccer, the use of statistical approaches is on the rise. We compare conventional statistical methods and machine learning algorithms regarding their ability to predict the binary injury status of young professional soccer players. For modeling, we consider basic soccer-related features and physical covariates derived from tests of postural control, strength, and movement. Lasso-regularized logistic regression, naive Bayes, linear discriminant analysis, k-nearest neighbors, classification trees, random forests, XGBoost, and support vector machines are used for injury probability prediction and subsequent binary classification in a cross-validated procedure. Prediction results are assessed via several quality measures. The best results are obtained by a post Lasso logistic regression model with a reduced penalty.

Keywords: Injury prediction; Soccer; Lasso regularization; Machine learning; Binary classification.

1 Introduction

The injury incidence in professional soccer is high, for young players even more than for adults (Pfirrmann et al., 2016). Injured players may miss trainings or matches, and a connection between aspects like injury incidence or injury burden in teams and a negative performance in domestic and international competitions has been found (Hägglund et al., 2013). Injuries in a team can lead to worries and uncertainty in other members and negative emotions can also be transmitted between players by the social contagion phenomenon (Hurley, 2016). Moreover, the club loses substantial amounts of money when team members are unable to play (Ekstrand, 2013). There are increasing efforts to prevent injuries by understanding their cause and predicting them with conventional statistical methods and machine learning. Popular modeling approaches include logistic regression, linear discriminant analysis (LDA), k-nearest neighbors (KNN), naive Bayes classifiers, trees, random forests, extreme gradient boosting, and support vector machines (SVMs; Rossi et al., 2022).

Using the groundwork of Kolodziej et al. (2023) of different Lasso-penalized logistic regression models (Friedman et al., 2010) to predict the binary injury status of young professional soccer players, we compare additional machine learning and conventional statistical modeling approaches on the same data. All models are assessed fairly in a parallel cross-validation (CV) procedure via several prediction quality measures. We search for a model that uses basic soccer-related as well as physical covariates and achieves a high prediction quality even on unseen data. Moreover, we want to identify the essential covariates for injury prediction in the best models.

2 Data

We use data of 56 young professional male soccer players from three youth teams (under 16, under 17, under 19) from two German professional soccer clubs. Covariate data was collected via a questionnaire and physical testing before the season start. The players underwent 3D motion analysis and force plate measurements during the execution of a single-leg drop landing (SLDL) task and an unanticipated side-step cutting (USSC) task. Moreover, their postural control under different conditions and their lower body strength in different movements were tested. The details of the testing procedure and all obtained covariates are described in Kolodziej et al. (2021) and Kolodziej et al. (2022). After the testing, any time-loss non-contact lower body injuries of the players were documented for the remaining ten months of the season, where *time-loss* refers to at least one day of absence in training or matches after the occurrence of the injury.

3 Methods and implementation

We use the occurrence of an injury as our binary response variable with the encoding $0 = \textit{no injury}$ and $1 = \textit{injury}$. The response is then modeled using four different Lasso-regularized logistic regression models, the naive Bayes method, LDA, KNN, a classification tree, a random forest, the XGBoost method, an SVM, and two featureless constant benchmark learners. The Lasso models are either regularized with an optimal penalty λ_{opt}, or with a slightly smaller penalty λ_{1se} within one standard error of λ_{opt} (see also Kolodziej et al., 2023). This weaker regularization usually leads to a model with a slightly larger set of selected covariates. Moreover, the regression coefficients are either shrunk via the penalty or re-estimated without shrinkage using post Lasso (Meinshausen, 2007).

All methods except for the constant learners undergo a parallel leave-one-out CV procedure where the injury probability of the left out player is predicted using a model fitted on all other observations. The constant learners use the proportion of injured players in the whole data set as a predicted injury probability. Hyperparameters are tuned in an inner 15-fold CV with identical folds for all methods in the process to ensure comparability. After all 56 probabilities are predicted, an optimal threshold for converting them into a binary injury status prediction is chosen separately for each method, namely the respective one which maximizes Youden's index (Youden, 1950), the sum of the sensitivity and the specificity minus 1.

In the end, the prediction results of all methods are assessed calculating the prediction accuracy, the sensitivity, the specificity, Youden's index, the AUC, the predictive Bernoulli likelihood, and the Brier score.

4 Results

The resulting quality measure values from the CV procedure are shown in Table 1. The two thresholds $-\infty$ and ∞ lead to two best constant learners. One predicts the class $1 = injury$ for all players, while the other always predicts the class $0 = no\ injury$.

The largest accuracy of 0.661 is reached by the XGBoost model; it predicts the injury status of 37 out of 56 players correctly. Its sensitivity of 0.500 and specificity of 0.765 lead to the third largest value of Youden's index of 0.265. Moreover, the model shows an AUC value of 0.549, a predictive Bernoulli likelihood of 0.509, and a Brier score of 0.253. In the two latter measures, it is dominated by the constant benchmark learners. The probability threshold for the XGBoost model is 0.491.

The post Lasso model with the smaller penalty λ_{1se} has the largest value of 0.302 for Youden's index, resulting from a sensitivity of 0.773 and a specificity of 0.529. The model also shows the largest AUC of 0.672, the largest predictive Bernoulli likelihood of 0.593, and the smallest Brier score of 0.228. Its accuracy value is 0.625. The model dominates the two constant learners in all measures. Its best probability threshold with regard to Youden's index is 0.270.

TABLE 1. Quality measure results (prediction accuracy, sensitivity, specificity, Youden's index, AUC, predictive Bernoulli likelihood, Brier score) on external test data via a LOO CV approach of all models (best model in **bold font**)

	Acc	Sens	Spec	Youd	AUC	Pr L	Brier
Lasso λ_{opt}	0.625	0.727	0.559	0.286	0.586	0.532	0.238
Lasso λ_{1se}	0.482	0.864	0.235	0.099	0.508	0.527	0.257
Post Lasso λ_{opt}	0.554	**0.955**	0.294	0.249	0.638	0.577	0.239
Post Lasso λ_{1se}	0.625	0.773	0.529	**0.302**	**0.672**	**0.593**	**0.228**
Naive Bayes	0.643	0.273	0.882	0.155	0.515	0.536	0.420
LDA	0.607	0.136	**0.912**	0.048	0.370	0.440	0.510
KNN	0.643	0.364	0.824	0.187	0.549	0.546	0.252
Tree	0.625	0.364	0.794	0.158	0.473	0.508	0.424
Random Forest	0.571	0.455	0.647	0.102	0.449	0.509	0.258
XGBoost	**0.661**	0.500	0.765	0.265	0.549	0.509	0.253
SVM	0.518	**0.955**	0.235	0.190	0.434	0.542	0.232
Constant 1	0.393	——	——	——	0.500	0.523	0.239
Constant 0	0.607	——	——	——	0.500	0.523	0.239

We fit both the above best post Lasso model with λ_{1se} and the XGBoost model again on all observations with re-tuned hyperparameters. The resulting Lasso model contains three covariates. The one with the largest absolute coefficient value is the center of pressure (COP) sway of a force plate which was used in a test of the players' postural control under static conditions. The model associates a larger COP sway with a larger probability to be injured. The XGBoost model assigns the largest variable importance value to the COP sway as well.

Both Lasso and XGBoost model also share the same second most important covariate, namely the concentric knee extension torque. A larger value is associated with a lower injury probability in the Lasso model.

The third covariate in the Lasso model is the hip external and internal rotation moment in the single-leg drop landing task. The model links an increased value to a smaller injury probability. The XGBoost model, on the other hand, places the third largest importance on the hip adduction and abduction moment in the unanticipated side-step cutting task.

5 Discussion

We approach the challenge of binary injury status prediction for youth players in professional soccer. Based on the research by Kolodziej et al. (2023), we search for a model with a high prediction capacity that takes into account soccer-related as well as physical covariates.

No newly regarded model outperforms the previous ones by a substantial amount or in most quality measures. The resulting best model is again a post Lasso logistic regression model with a reduced penalty, reaching the best values for Youden's index, AUC, predictive likelihood, and Brier score. The XGBoost model obtains the largest accuracy. These two models agree on the two most important covariates for injury prediction.

Predicting soccer injuries with high precision still remains difficult. The investigated models provide a reasonable improvement in comparison to the benchmark models, but more benefit is desirable.

Acknowledgments: Special thanks to the participating youth academies, coaches, players, and the test team.

References

Ekstrand, J. (2013). Keeping your top players on the pitch: The key to football medicine at a professional level. *British Journal of Sports Medicine*, **47**, 723 – 724.

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

Hägglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson, H. and Ekstrand, J. (2013). Injuries affect team performance negatively in professional football: An 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, **47**, 738 – 742.

Hurley, O. (2016). Impact of player injuries on teams' mental states, and subsequent performances, at the Rugby World Cup 2015. *Frontiers in Psychology*, **7**, 807.

Kolodziej, M., Nolte, K., Schmidt, M., Alt, T. and Jaitner, T. (2021). Identification of neuromuscular performance parameters as risk factors of non-contact injuries in male elite youth soccer players: A preliminary study on 62 players with 25 non-contact injuries. *Frontiers in Sports and Active Living*, **3**, 615330.

Kolodziej, M., Willwacher, S., Nolte, K., Schmidt, M. and Jaitner, T. (2022). Biomechanical risk factors of injury-related single-leg movements in male elite youth soccer players. *Biomechanics*, **2**, 281 – 300.

Kolodziej, M., Groll, A., Nolte, K., Willwacher, S., Alt, T., Schmidt, M. and Jaitner, T. (2023). Predictive modeling of lower extremity injury risk in male elite youth soccer players using least absolute shrinkage and selection operator regression. *Scandinavian Journal of Medicine & Science in Sports*, **33**, 1021 – 1033.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis*, **52**, 374 – 393.

Pfirrmann, D., Herbst, M., Ingelfinger, P., Simon, P. and Tug, S. (2016). Analysis of injury incidences in male professional adult and elite youth soccer players: A systematic review. *Journal of Athletic Training*, **51**, 410 – 424.

Rossi, A., Pappalardo, L. and Cintia, P. (2022). A narrative review for a machine learning application in sports: An example based on injury forecasting in soccer. *Sports*, **10**, 5.

Youden, W. (1950). Index for rating diagnostic tests. *Cancer*, **3**, 32 – 35.

Penalized mixed cumulative regression for modelling varying dispersion and cluster heterogeneity

Moritz Berger[1], Maria Iannario[2]

[1] Institute of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Germany

[2] Department of Political Sciences, University of Naples Federico II, Italy

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

Abstract: Cumulative regression is the most frequently used tool for modelling ordinal outcomes. If data come in clusters the model needs to consider possible heterogeneity between measurement units. Additionally, differing variability within clusters might appear in the data. A cumulative model with random effects is introduced that accounts for both issues. Building on this, we propose a penalized maximum likelihood estimation procedure that allows for variable selection separately in the location and dispersion component of the model predictor. The new approach is illustrated using data of the Survey of Health, Ageing and Retirement in Europe.

Keywords: Clustered data; Ordinal regression models; Unobserved heterogeneity; Varying dispersion.

1 Mixed-effects models for ordinal data

Let the data be given by $(y_{ij}, \boldsymbol{x}_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, N_i$, where $y_{ij} \in \{1, \ldots, k\}$ denotes the ordinal outcome of observation j in cluster i and $\boldsymbol{x}_{ij}$ is a vector of covariates. There are n clusters for which the number of observations $N_1, \ldots, N_n$ may vary. The covariates $\boldsymbol{x}_{ij}$ may be constant or vary across measurements of one cluster. A popular tool for the analysis of ordinal outcome variables is the class of cumulative models (McCullagh, 1980; Tutz, 2012). Including a cluster-specific random intercept b_i, the basic form of the mixed-effects cumulative model (Hedeker and Gibbons, 1994) is given by

$$P(y_{ij} \leq r|\boldsymbol{x}_{ij}) = F(\eta_{ijr}) = F\left(\beta_{0r} - \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} - b_i\right), \quad r = 1, \ldots, k-1, \tag{1}$$

where $F(\cdot)$ is a distribution function, $\boldsymbol{\beta}^\top = (\beta_1, \ldots, \beta_p)$ is the vector of regression coefficients and $-\infty < \beta_{01} < \ldots < \beta_{0,k-1} < \infty$ are category-specific intercepts

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

that need to be strictly increasing. The random intercepts are assumed to be normally distributed with zero mean and variance σ^2, $b_i \sim N(0, \sigma^2)$. Presence of the random intercepts means that the category-specific thresholds are simultaneously shifted yielding the cluster-specific thresholds $\beta_{01} - b_i, \ldots, \beta_{0,k-1} - b_i$. A widely applied choice for $F(\cdot)$ is the logistic distribution function $F(\cdot) = \exp(\cdot)/(1 + exp(\cdot))$. The latter yields the logistic regression model for cumulative probabilities, namely *proportional odds model*, if the parallel assumption is taken into account. The use of the logistic distribution greatly simplifies the interpretation of the regression coefficients and offers interesting properties with regard to robustness (Scalera et al., 2021; Iannario, 2023).

2 Extended model for variance heterogeneity

If the fit of model (1) is unsatisfactory one frequently uses a more complex non-proportional odds model with category-specific parameters. A lack-of-fit, however, can also be caused by differing variability in subgroups of the population, by so-called dispersion effects. One way to introduce differing variability is to model it explicitly by a *scaling component* depending on covariates (McCullagh, 1980). An alternative cumulative type model that accounts for varying dispersion is the *location-shift model*, which was proposed by Tutz and Berger (2017) for ungrouped data and extended by Schauberger and Tutz (2022) for multivariate ordinal outcomes. The variant of the location-shift model with cluster-specific random intercepts, which is considered here, has the form

$$\eta_{ijr} = \beta_{0r} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} - b_{i,\ell} + (r - k/2) \left(\boldsymbol{z}_{ij}^\top \boldsymbol{\alpha} + b_{i,d} \right) , \tag{2}$$

where $\boldsymbol{z}_{ij}$ is an additional vector of covariates. The two random intercepts $b_{i,\ell}$ and $b_{i,d}$ are assumed to be normally distributed with zero mean and variance-covariance matrix $\boldsymbol{\Sigma}_b$. The predictor in (2) contains the familiar location term, which reflects the tendency to low or high categories, and a scaled dispersion term $\delta_{ij} = (r - k/2) \left(\boldsymbol{z}_{ij}^\top \boldsymbol{\alpha} + b_{i,d} \right)$ with random intercept $b_{i,d}$, which shifts the thresholds and reflects the tendency to the middle or extreme categories. As in the simple model (1), the random intercept in the location term represents between-cluster variability, while the dispersion term δ_{ij} represents within-cluster variability. Hence, the new predictor (2) allows for the possible *variance heterogeneity* of respondents to be taken into account. Furthermore, presence of random effects in the dispersion term means that the within-cluster variance is allowed to differ across clusters. In the case $k = 4$ the dispersion terms result to $-\boldsymbol{z}_{ij}^\top \boldsymbol{\alpha} - b_{i,d}$ $(r = 1)$, 0 $(r = 2)$ and $\boldsymbol{z}_{ij}^\top \boldsymbol{\alpha} + b_{i,d}$ $(r = 3)$, which means that the middle threshold remains fixed, but the lower and upper thresholds are shifted. This example illustrates that if $\delta_{ij} > 0$ the intervals defined by the thresholds are widened, indicating weaker dispersion and therefore more concentration in the middle. If $\delta_{ij} < 0$ the intervals are shrunk, indicating stronger dispersion and therefore more concentration in the extreme categories. Asymptotically, if $b_{i,d} \to \infty$ (all other model components fixed) one obtains $P(y_{ij} = 2 | \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) + P(y_{ij} = 3 | \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = 1$ with the whole probability mass in the middle categories.

3 Penalized marginal likelihood estimation

The predictor in (2) contains two different sets of covariates, the vector $\boldsymbol{x}_{ij}$ determining the location and the vector $\boldsymbol{z}_{ij}$ determining the dispersion (potential variance heterogeneity in clusters of respondents). In applications, however, there is only one set of covariates and it is typically not known which variables have an impact on which of the two components. This calls for a tailored variable selection strategy within the fitting procedure.

Estimates of the location-shift model, whose predictor is in (2), can be obtained by maximization of the marginal likelihood expressed as an integral over the likelihood of the form

$$L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_b) = \int \int \prod_{j=1}^{N_i} \prod_{r=1}^{k} P(y_{ij} = r | \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_b)^{\Delta_{ijr}} \\ \times \; f(b_{i,\ell}, b_{i,d}) \, db_{i,\ell} \, db_{i,d} \, , \qquad (3)$$

where $\Delta_{ijr} = 1$ if $y_{ij} = r$ and $y_{ijr} = 0$, otherwise, and $f(\cdot)$ denotes the two-dimensional normal density function. Gauß-Hermite quadrature can be applied to approximate the integral over the random-effects distribution (for details, see McCulloch and Searle, 2001). To obtain a sparse representation and in particular variable selection with regard to the location and the dispersion term, we consider a penalized marginal log-likelihood of the form

$$\ell_p(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_b) = \ell(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_b) - J_{\lambda_\ell, \lambda_d}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \, , \qquad (4)$$

where $\ell(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}_b)$ denotes the raw marginal log-likelihood and $J_{\lambda_\ell, \lambda_d}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ represents a penalty term that depends on the scalar tuning parameters λ_ℓ and λ_d. To allow for a different degree of regularization in the two model components, we propose to use two separate LASSO-type penalties, one on the location parameters $\boldsymbol{\beta}$ and one on the dispersion parameters $\boldsymbol{\alpha}$, which is given by

$$J_{\lambda_\ell, \lambda_d}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \lambda_\ell \, \|\boldsymbol{\beta}\|_1 + \lambda_d \, \|\boldsymbol{\alpha}\|_1 \, . \qquad (5)$$

The optimal tuning parameters λ_ℓ and λ_d can be chosen by subsampling, e.g. by cross-validation using the predictive log-likelihood as the criterion to be cross-validated. In our implementation, we make use of the SAS PROC NLMIXED, which offers a general framework for fitting nonlinear mixed-effects models. For optimization, PROC NLMIXED performs a quasi-newton algorithm, which also involves computing the first-order derivatives of the quadrature approximation.

4 Application to SHARE

We consider data from the seventh wave of the Survey of Health, Ageing and Retirement in Europe, in short SHARE, collected in 2017 (see Börsch-Supan et al., 2013, for methodological details). SHARE is a panel survey collecting detailed cross-national information on the health, socio-economic status and family networks of people aged 50 and over from a large group of European countries. The sample analysed consists of 3,430 respondents living in 27 countries. It is mainly characterised by female respondents (58%), with an average age of 67.9 years (SD = 9.7 years). In addition to standard demographics, data collection

TABLE 1. Analysis of the SHARE data. Coefficient estimates obtained from fitting the proposed penalized mixed-effects model with country-specific random intercepts. According to the predictive performance, the optimal tuning parameters were $\lambda_l = 0.0041$ and $\lambda_d = 0.0015$.

Covariate	**Location effect**		**Dispersion effect**		**Random effects**	
	$\hat{\beta}$	$\exp(\hat{\beta})$	$\hat{\alpha}$	$\exp(\hat{\alpha})$	$\mathbf{\Sigma}_b$	
Age	0.031	1.032	0.011	1.012	$\hat{\sigma}_l$	0.361
Gender	-0.442	0.643	0.165	1.179	$\hat{\sigma}_d$	0.074
BMI	0.069	1.071	0.011	1.011	$\hat{\sigma}_{l,d}$	0.091
Years of education	-0.040	0.960	—	—		
ADL index	0.324	1.383	—	—		
IADL index	0.277	1.319	0.011	1.011		
Life satisfaction	-0.317	0.728	-0.040	0.961		
Hand grip	-0.029	0.972	0.010	1.010		

comprised measurements on socio-economic information (e.g. employment status) and health-related aspects (e.g. chronic diseases). For the present analysis, we will focus on the perception of one's health status measured on an ordinal scale from 1 (excellent) to 5 (poor).

In our analysis we include the following eight covariates: age, gender (0: male, 1: female), body mass index (BMI), years of education, ADL index (number of limitations with activities of daily living), IADL index (number of limitations in instrumental activities of everyday life), life satisfaction (0: completely dissatisfied to 10: completely satisfied), and hand grip (which is a measure of physical health). The results from fitting the proposed model with predictor (2) are given in Table 1. The estimated variance-covariance matrix, with $\hat{\sigma}_l = 0.361$, shows that country-specific effects should not be ignored. While all eight covariates were selected in the location term, years of education and ADL index were excluded from the dispersion term by the penalized fitting procedure. From an interpretative point of view, the coefficient estimates for gender indicate that women perceive their state of health to be better than men ($\hat{\beta} = -0.442$) with a stronger tendency to the middle category ($\hat{\alpha} = 0.165$). Comparing women to men (all other model components fixed), the cumulative odds $P(y_{ij} \leq r|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})/P(y_{ij} > r|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$ increase by the factor 1.215 ($r = 1$), 1.433 ($r = 2$), 1.690 ($r = 3$) and 1.993 ($r = 4$), which reflects the difference in the location and in the dispersion. According to the estimates for IADL index, respondents with a higher number of limitations in instrumental activities report a poorer health status ($\hat{\beta} = 0.277$) with a tendency to the middle category ($\hat{\alpha} = 0.011$).

5 Outlook

When fitting the proposed model to the SHARE data, we treated all eight covariates as metrically scaled variables. For life satisfaction (measured on a 11-point scale) it might, however, be more appropriate to encode it as an ordinal variable, in particular, if the points on the scale can not be interpreted as equally spaced. In this case the LASSO-type penalty as defined in (5) should be replaced by a more complex group LASSO or fusion penalty (Tutz and Gertheiss, 2016), which enforces sparsity of groups and within each group. These extensions will be the subject of in-depth study and future work.

References

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B. and Stuck, S. (2013). Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, **42**, 992 – 1001.

Hedeker, D. and Gibbons, R.D. (2006). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, **50**, 933 – 944.

Iannaria, M. (2023). Robust regression modelling for ordinal categorical data. *Proceedings of the 37th International Workshop on Statistical Modelling*, Dortmund, Germany, Volume II, 28 – 38.

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

McCulloch, C. and Searle, S. (2001). *Generalized, Linear, and Mixed Models.* New York: Wiley.

Scalera, V., Iannario, M. and Monti, A.C. (2021). Robust link functions. *Statistics*, **55**, 963 – 977.

Schauberger, G. and Tutz, G. (2022). Multivariate ordinal random effects models including subject and group specific response style effects. *Statistical Modelling*, **22**, 409 – 429.

Tutz, G. and Berger, M. (2017). Separating location and dispersion in ordinal regression models. *Econometrics and Statistics*, **2**, 131 – 148.

Tutz, G. and Gertheiss, J. (2016). Regularized regression for categorical data. *Statistical Modelling*, **16**. 161 – 200.

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

A comparison of methods for oil production forecasting of the Santos Basin

Herlisson Bezerra[1], Luis F. B. de Messis[1], Cibele Russo[2], Thomas Peron[2]

[1] Interinstitutional Graduate Program in Statistics UFSCar-USP (PIPGEs), Federal University of São Carlos and University of São Paulo, Brazil
[2] Institute of Mathematical and Computer Sciences, University of São Paulo, Brazil

E-mail for correspondence: herlissonmaciel@hotmail.com

Abstract: Despite the current global eagerness to complete the transition of the energy matrix to clean and renewable sources, oil is undoubtedly one of the most important commodities in the global economy. In Brazil, the Santos Basin stands out as one of the largest oil producers in the country in recent years. Therefore, predicting the production of each field within it is of great importance for the development of business strategies. In this context, the goal of the presented study is to compare the performance of several methods of time series forecasting for the oil production in fields within the Santos Basin using open data from the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) on monthly production. The results showed that, among 20 fields, MiniRocket outperformed the other models in 7 fields, Theta model did it in 4 fields, Long-Short Term Memory in 3 fields, MultiRocket in 2 fields, Multilayer Perceptron in 2 fields, Rocket and Gated Recurrent Unit in 1 field.

Keywords: Oil production forecasting; Time series forecasting; Statistical Learning.

1 Introduction

The technological advancement of electric vehicles and renewable energy sources is undeniable. However, in the current global economic scenario, oil remains one of the most important commodity. In this context, Brazil is among the top ten oil producers worldwide. The Brazilian production is mainly derived from offshore extraction, while most other countries have onshore production. Brazilian offshore production is primarily extracted from the Santos Basin. It is located in the southeast region of the Brazilian coast and was responsible for producing approximately 74% of all Brazilian oil in 2023. Besides being the basin with

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

the highest production, it is also estimated to have a pre-salt region reserve of over 100 billion barrels of oil. Due to its importance, for decision-making and the adoption of business strategies, this work proposes the comparison of the predictions of time series for oil production, corresponding to the period of the series' onset between 2010 and 2020, taking into account the peculiarities of each analyzed production period. The predictions for each series of the fields within the Santos Basin were made using the proposed models: Random Convolutional Kernel Transform (Rocket), Minimally Random Convolutional Kernel Transform (MiniRocket), Multi Random Convolutional Kernel Transform (MultiRocket). They are explained in details in section 2. Additionally, for camparison, it is proposed the methods: Autoregressive Integrated Moving Average (ARIMA), Theta method, Artificial Neural Networks (ANNs) Multilayer Perceptron (MLP), ANN Long-Short Term Memory (LSTM) and ANN Gated Recurrent Unit (GRU). In section 3, the proposed models are compared with each other to observe which of them had the best fit to the studied data through the mean squared error metric.

2 Methodology

2.1 Dataset

The dataset was obtained from the Public Data Query portal of the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP). The website provides monthly hydrocarbon production data, categorized by monthly periods and by basin and production field.

2.2 Rocket, MiniRocket and MultiRocket

The Rocket method applies a transformation to time series data by utilizing numerous random convolutional kernels. These kernels have random characteristics such as length, weights, bias, dilation, and padding. The resulting features undergo training with a linear regressor. The MiniRocket method differs from Rocket by using a fixed-size set of 84 kernels that are applied as transformers to the time series resulting in a vector of extracted features that has a smaller computational cost. The set of kernels was proposed by a selection scheme based on combinatorial optimization. The Multirocket variant also uses the same set of kernel as MiniRocket but it has a step that increases the diversity of the time series features through adding multiple pooling operators.

Furthermore, in practical applications, especially for datasets of moderate size, like the studied, it is opted for a ridge regressor fed with the exctrated features with Rocket. This choice offers the advantage of quick cross-validation for the regularization hyperparameter without the need for other hyperparameters.

3 Results

For the analysis of model fitting, each of the 20 series was divided into 80% for training and 20% for testing, ensuring that the observations of the series were divided in a contiguous manner. The machine learning models were fitted only once with the data from the training set, and for each predicted point, only one-step-ahead forecasts were used. In contrast, the statistical models, ARIMA and Theta Method, utilized previous predicted points for forecasts with a horizon greater than one observation. The Mean Squared Error (MSE) metric was used to assess the quality of the fit for each trained model and the results of the tests are in Figure 1. The lower the MSE, the better is the fitted model.

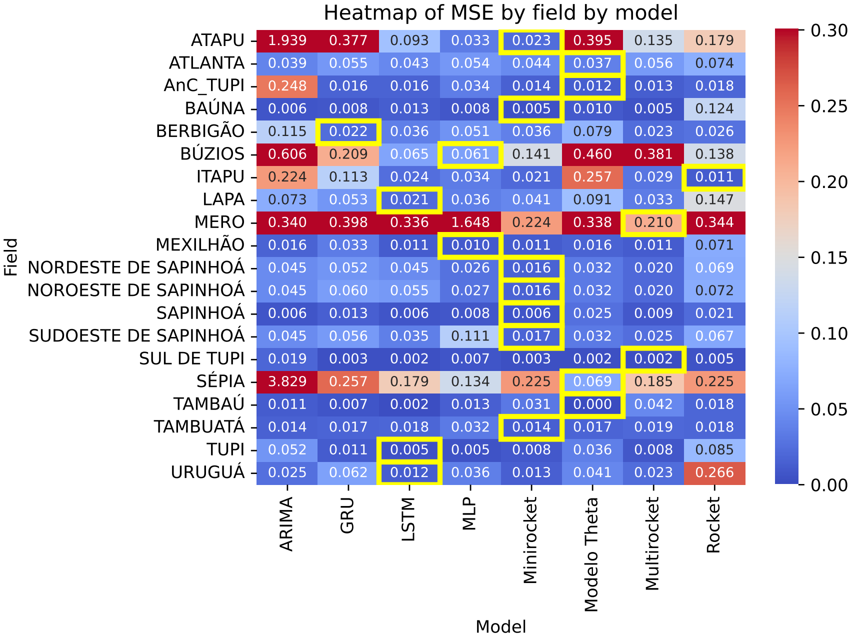


FIGURE 1. Metrics of the fitted models for each field by model. The yellow rectangles indicate the lowest MSE value in the row.

It is possible to notice in the Figure 1 that in the test dataset, the Minirocket outperformed the other models in 7 fields, Theta model did it in 4 fields, LSTM in 3 fields, Multirocket in 2 fields, MLP in 2 fields, Rocket and GRU in 1 field.

4 Conclusion

From the comparison among the methods used to predict the production of the 20 oil fields in the Santos Basin, it is concluded that: the MiniRocket method achieved the lowest MSE in the test set in 7 fields; the Theta method had the best performance in 4 fields; the LSTM method had the best performance in 3 fields; the MultiRocket method had the best performance in 2 fields; the MLP method had the best perfomance in 2 fields; the Rocket and GRU methods had the best performance in only 1 field. The ARIMA method did not achieve the lowest MSE in any field. It is important to emphasize that the main methods used are among the most current and tend to achieve good performance. However, for future work, other methods can be employed with the aim of discovering other approaches that yield the lowest possible prediction error.

Acknowledgments: Special Thanks to the Brazilian Foundation, CAPES, for financial supporting this study through process process number 88887.821 273/2023-00.

Appendix A: All scripts written to fit the models and the datasets are available in `github.com/HerlissonMB/iwsm2024`.

References

Assimakopoulos, V. and Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International journal of forecasting*, **16**, 521 – 530.

Box, G. E., Reinsel, G.C., Jenkins, G.M. and Ljung, G.M. (2015). *Time Series Analysis: Forecasting and Control.* John Wiley & Sons.

Dempster, A., Petitjean, F. and Webb, G. I. (2020). ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, **34**, 1454 – 1495.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning.* MIT press.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Tan, C. W., Dempster, A., Bergmeir, C. and Webb, G.I.(2022). MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. *Data Mining and Knowledge Discovery*, **36**, 1623 – 1646.

Public Query System of ANP. (2023). <https://cdp.anp.gov.br>.

Estimating a lower bound of the population size in capture-recapture experiments with right censored data

Anabel Blasco-Moreno[1], Pere Puig[2]

[1] Departament de Matemàtiques and Servei d'Estadística de la UAB, Spain
[2] Departament de Matemàtiques de la UAB, and CRM, Spain

E-mail for correspondence: `pere.puig@uab.cat`

Abstract: In this work, we present a new non-parametric approach for estimating a lower bound of the population size in capture-recapture experiments with right-censored observations, along with several applications in ecology and social sciences.

Keywords: Censored data, Chao's estimator, Mixed-binomial distributions.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

REML for two-dimensional P-splines

Martin P. Boer[1]

[1] Wageningen University and Research, The Netherlands

E-mail for correspondence: martin.boer@wur.nl

Abstract: We propose a new method based on residual (or restricted) maximum likelihood (REML) for P-splines. Existing methods use a transformation of P-splines to a mixed model; in the new model it is shown that a more direct method can be used keeping the sparse structure of P-splines. The method is illustrated with a two-dimensional example using the R-package LMMsolver on CRAN. We will show that for this example LMMsolver is several orders of magnitude faster than other methods, which use a transformation of P-splines to mixed models where the sparse structure of P-splines is lost.

Keywords: Mixed models; Precision matrices; Sparse matrix algebra.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Relational event additive modeling of alien plant and insect species invasions

Martina Boschi[1], Ernst C. Wit[1]

[1] Università della Svizzera italiana, Switzerland

E-mail for correspondence: martina.boschi@usi.ch

Abstract: The invasion of alien species into non-native environments is a critical issue due to its potential harm to biodiversity, economies, and human health. We propose an additive relational event model (REM) to analyze invasions of plants and insects that occurred between 1880 and 2005. Our proposed REM includes smooth and interpretable time-varying effects to explain the invasion rates. Our inference approach employs a case-control sampling technique, enabling efficient computation. Since the goodness of fit of REMs is an ongoing research challenge, especially for REMs that involve smooth terms, we propose to evaluate the adequacy of the model using cumulative weighted martingale residuals. We employ a Kolmogorov-Smirnov type test to determine if covariates are properly modeled. Implementation is performed through the R package `mgcv`.

Keywords: Relational event model; Alien species invasions; Generalized additive models; Goodness-of-fit.

1 Introduction

Each year, alien species disperse from their native habitats to new regions, often facilitated by, potentially human, vectors. While not all introduced species become invasive, the widespread nature of this phenomenon poses a significant threat due to environmental changes and associated costs. Various modeling approaches have been employed to elucidate these mechanisms, but they are either computationally complex or consider only one covariate at the time.

To address this, we present a smooth relational event model that incorporates a flexible time-varying specification of effects. This enhancement aims to improve the interpretability of the impact of various drivers on the risk of alien species invasions. Additionally, our proposed case-control partial likelihood inference technique allows for a significant reduction in computational costs. We consider 13094 invasions between 1880 to 2005 involving vascular plants and insects species across 275 areas as part of the alien species First Record (FR) Database (Seebens et al., 2017). Existing literature suggests that the spread of plants is notably influenced by certain insects (Richardson et al., 2000). Our aim is to unravel the various components and their relative importance in driving plant and insect dispersion. In particular, we will focus on geographic distance, climatic similarity, trade, land coverage and the presence of colonial ties.

2 Relational event additive model

A **relational event** is an interaction at time t starting from a sender s towards a receiver r. It will be denoted as a triplet (s, r, t), where t varies in $[0, \tau]$. In the species invasion context, we aim to model an event sequence $\mathcal{E}$ of $n = 13094$ FRs, where the invading species is the sender s and the invaded region is the receiver r. Relational events may be thought as the realizations of a marked point process, $\{(t_k, (s_k, r_k)); k \geq 1\}$, where at time t_k (point) when the interaction $s_k \to r_k$ (mark) occurs. A counting process $N_{sr}(t)$ may be associated to the process above, counting, for each mark (s, r), the number of occurred interactions $s \to r$ in $[0, t]$. Under standard assumptions, N_{sr} is a continuous time sub-martingale and, as such, we can decompose it according to the Doob-Meyer theorem,

$$N_{sr}(t) = M_{sr}(t) + \Lambda_{sr}(t) = M_{sr}(t) + \int_0^t \lambda_{sr}(u)\mathrm{d}u \tag{1}$$

where M_{sr} is a zero-mean martingale. Conditional on the filtration $\mathbb{H} = \{\mathcal{H}_t\}_{t \geq 1880}$, where $\mathcal{H}_t$ incorporates information on occurred events up to t, Λ_{sr} is the predictable part of the counting process N_{sr}. A **relational event model** (REM) is defined by modelling the intensity function $\lambda_{sr}(t) = \frac{d}{dt}\Lambda_{sr}(t)$. We consider a non-linear formulation,

$$\lambda_{sr}(t|\mathcal{H}_{t^-}; \boldsymbol{\beta}_0(t)) = \lambda_0(t) \exp\left[\boldsymbol{\beta}_0(t)^T \boldsymbol{x}_{sr}(t)\right] \tag{2}$$

where $\lambda_0(t)$ represents the baseline hazard, $\boldsymbol{x}$ are time-varying covariates, and $\boldsymbol{\beta}_0$ are non-linear, time-varying effects of linearly associated covariates.

2.1 Case-control partial likelihood via GAMs

The estimation procedure for REMs is normally based on partial likelihood (Perry & Wolfe, 2013). Due to the computational challenges associated with the denominator of the partial likelihood for large dynamic networks, nested case-control (NCC) sampling (Borgan et al., 1995) has been adapted for REMs (Vu et al., 2015). This involves evaluating, at each time point, a sample of individuals at risk, termed the sampled risk set SR. When at random a single non-event is sampled, then the partial likelihood reduces to the likelihood of an additive logistic regression, $\hat{\boldsymbol{\beta}}$ is found by maximizing:

$$\mathcal{L}^{PS}(\boldsymbol{\beta}|\mathcal{E}) = \prod_{k=1}^{n} \left\{1 + \exp\left[-\left(\boldsymbol{\beta}(t_k)^T \Delta_{s_k r_k} \boldsymbol{x}\right)\right]\right\}^{-1} \tag{3}$$

where $\Delta_{s_k r_k}\boldsymbol{x} = \boldsymbol{x}_{s_k r_k}(t_k) - \boldsymbol{x}_{s_k^* r_k^*}(t)$, where (s_k, r_k) is the dyad observed in event k, and (s_k^*, r_k^*) is the associated sampled non-event.

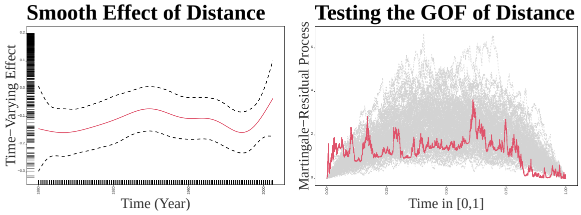


FIGURE 1. Smooth effect of distance on species invasions dynamics and GOF process for distance covariate.

2.2 Checking the goodness of fit

Evaluating the fit of REMs is challenging. While some methods have been proposed, they are either computationally expensive (Brandenberger, 2019) or are more exploratory (Juozaitienė et al., 2023). Our proposal consists of the evaluation of a martingale-residual type process (Marzec & Marzec, 1998). The process of interest, denoted as $G[\hat{\boldsymbol{\beta}}]$ follows,

$$G[\hat{\boldsymbol{\beta}}, u|\mathcal{E}] \quad = \quad \sum_{k \leq \lfloor nu \rfloor} \left[\phi_{s_k r_k}(t_k) - \sum_{(s,r)\in SR} \phi_{sr}(t_k) \cdot \pi_{sr}(\hat{\boldsymbol{\beta}}) \right]$$

defined on n equally spaced points $u \in [0,1]$. ϕ_{sr} is any statistic of interest and $\pi_{sr}(\hat{\boldsymbol{\beta}})$ is the fitted multinomial probability of the event.
We consider testing for the **goodness-of-fit** (GOF) of covariates, $\phi_{sr} = \boldsymbol{x}_{sr}$. In that case, the expression for G coincides with the scaled components of the score vector. The statistical test of the **Kolmogorov-Smirnov** (KS) type is defined as follows:

$$\mathrm{KS}^q = \sup_{u\in[0,1]} \|\hat{\boldsymbol{W}}[\hat{\boldsymbol{\beta}}, u]\|^2 = \sup_{u\in[0,1]} \|\hat{\boldsymbol{J}}_{\boldsymbol{G}[\hat{\boldsymbol{\beta}}]}^{-\frac{1}{2}} \times n^{-\frac{1}{2}} \times \boldsymbol{G}[\hat{\boldsymbol{\beta}}, u|\mathcal{E}]\|^2 \tag{4}$$

where $\hat{\boldsymbol{J}}_{\boldsymbol{G}[\hat{\boldsymbol{\beta}}]}^{-\frac{1}{2}} = n^{-1} \times \sum_{k=1}^{n} \boldsymbol{G}_k[\hat{\beta}, t_k]\boldsymbol{G}_k[\hat{\beta}, t_k]^T$, $\boldsymbol{G}_k[\hat{\beta}, t_k]$ being the individual contribution to process $\boldsymbol{G}[\hat{\boldsymbol{\beta}}, \cdot]$. Under the assumption of adequacy of the model formulation, $\hat{\boldsymbol{W}}[\hat{\boldsymbol{\beta}}, u]$ converges to a vector of q independent Brownian bridges and p-value of KS test can be simulated empirically (Hjort & Koning, 2002).

3 Modelling plant and insect invasions

We fitted and computed the corresponding AIC for 728 model formulations, encompassing all possible combinations of available covariates, with both fixed and time-varying effects. The best model , according to corrected AIC, includes *distance*, *trade*, *colonial ties*, *climatic dissimilarity*, and *urban land-coverage*, modeled with time-varying effects. Figure 1 *Left* shows the estimated smooth time-varying effects of distance. Its estimated negative effect confirms the rarity of long-distance natural invasion occurrences (Juozaitienė et al., 2023). Figure 1 *Right* shows that the effect of distance seems accurately estimated by this model.

References

Borgan, O., Goldstein, L. and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics*, **23**, 1749 – 1778.

Brandenberger, L. (2019). Predicting network events to assess goodness of fit of relational event models. *Political Analysis*, **27**, 556 – 571.

Hjort, N.L. and Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, **14**, 113 – 132.

Juozaitienė, R., Seebens, H., Latombe, G., Essl, F. and Wit, E.C. (2023). Analysing ecological dynamics with relational event models: The case of biological invasions. *Diversity and Distributions*, **29**, 1208 – 1225.

Marzec, L. and Marzec, P. (1998). Testing based on sampled data for proportional hazards model. *Statistics & Probability Letters*, **37**, 303 – 313.

Perry, P.O. and Wolfe, P.J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 821 – 849.

Richardson, D.M., Allsopp, N., D'Antonio, C.M., Milton, S.J. and Rejmánek M. (2000). Plant invasions–the role of mutualisms. *Biological Reviews*, **75**, 65 – 93.

Seebens, H., et al. (2017). No saturation in the accumulation of alien species worldwide. *Nature Communications*, **8**, 1 – 9.

Vu, D., Pattison, P. and Robins, G. (2015). Relational event models for social learning in MOOCs. *Social Networks*, **43**, 121 – 135.

Boosting distributional copula regression for bivariate time-to-event data

Guillermo Briseño Sanchez[1], Nadja Klein[1], Andreas Groll[1], Andreas Mayr[2]

[1] TU Dortmund University, Germany
[2] University of Bonn, Germany

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

Abstract: We propose a flexible distributional copula regression model for right-censored bivariate time-to-event data. The joint survival function is constructed using parametric copulas, allowing for separate specification of the dependence structure between the margins and their respective distributions, which are modelled via well-known parametric distributions such as the log-normal, log-logistic or Weibull. Following the generalised additive models for location, scale and shape (GAMLSS) approach, possibly all parameters may be modelled in an additive fashion through semi-parametric predictors. Estimation is carried out via component-wise gradient boosting, leading to data-driven variable selection, an extremely helpful feature in such a complex model class, especially in high-dimensional ($p \gg n$) settings. Our method is implemented as an add-on function of the `R` package `gamboostLSS`, ensuring transparent and reproducible research.

Keywords: Dependence modelling; GAMLSS; Shrinkage; Survival analysis; Variable selection.

1 Progression time of AMD in left and right eyes

According to the National Eye Institute, age-related macular degeneration (AMD) is the leading cause of blindness in England and the United States (The AREDS Group, 1999). Traditional statistical techniques often rely on specifying a model for a univariate response with a dummy-variable indicating a patient's left or right eye. Modelling the survival function of patient's time to progression in the left and right eyes jointly using a distributional regression approach could provide new insights of AMD progression time, the role clinical characteristics play in said times as well as their dependence. Moreover, we are interested in uncovering which covariates affect certain aspects of the joint survival function as well as estimating their functional form in a data-driven manner, while also keeping individual models for the progression time in each organ. Consequently, we propose

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

a distributional copula regression modelling approach where estimation is carried out using component-wise gradient boosting in order to obtain an interpretable statistical model for the entire bivariate survival function.

2 Methodology

Let T_{vi}, $v = 1, 2$, $i = 1, \dots, n$, denote the true event-time and $\tilde{T}_{vi}$ an independent (of T_{vi}), random, non-informative censoring time. In practice one observes $Y_{vi} = \min\{T_{vi},\ \tilde{T}_{vi}\}$ accompanied by the censoring indicator $\delta_{vi} = \mathbf{1}\{T_{vi} \leq \tilde{T}_{vi}\}$. Within the framework of GAMLSS (Rigby and Stasinopoulos, 2005), the i-th observation of a response is assumed to follow a parametric distribution with density f and survival function $S = 1 - F$, where F is the cumulative distribution function (CDF). For bivariate time-to-event responses, the joint survival function of $(Y_{1i}, Y_{2i})^{\mathrm{T}}$ is written as

$$S_{1,2}(y_{1i}, y_{2i}; \boldsymbol{\vartheta}_i) = C\left(S_1\left(y_{1i}\ ;\ \boldsymbol{\vartheta}_i^{(1)}\right),\ S_2\left(y_{2i}\ ;\ \boldsymbol{\vartheta}_i^{(2)}\right); \vartheta_i^{(c)}\right), \tag{1}$$

where $C(\cdot, \cdot) : [0,1]^2 \rightarrow [0,1]$ is the CDF of a bivariate parametric copula function with parameter $\vartheta_i^{(c)}$ that determines association's strength between the marginal responses (Nelsen, 2006) and $S_1(\cdot)$, $S_2(\cdot)$ denote the univariate marginal survival functions. The $K = K_1 + K_2 + 1$ dimensional vector $\boldsymbol{\vartheta}_i = \left(\boldsymbol{\vartheta}_i^{(1)}, \boldsymbol{\vartheta}_i^{(2)}, \vartheta_i^{(c)}\right)^{\mathrm{T}}$ contains the sub-vectors of margin-specific parameters and the copula dependence parameter. Each component of $\boldsymbol{\vartheta}_i$ is allowed to depend on covariates by means of structured additive predictors and suitable link functions $g(\cdot)$ with corresponding inverse or response functions $h(\cdot) \equiv g^{-1}(\cdot)$ that enforce parameter space restrictions:

$$g_k^{(\bullet)}\left(\vartheta_{ik}^{(\bullet)}\right) = \eta_{ik}^{(\bullet)} = \beta_{0k}^{(\bullet)} + \sum_{r=1}^{P_k^{(\bullet)}} s_{rk}^{(\bullet)}(x_{ir}),\ k = 1, \dots, K_\bullet,\ \bullet \in \{1, 2, c\}, \tag{2}$$

where $\beta_{0k}^{(\bullet)}$ are parameter-specific intercepts and $s_{rk}^{(\bullet)}(\cdot)$ are smooth functions that can accommodate a wide range of functional forms of the covariates. The summation limit $P_k^{(\bullet)}$ from Equation (2) emphasises that the individual parameters $\vartheta_{ik}^{(\bullet)}$ do not necessarily have to be modelled using the same subset of covariates. A major issue of distributional regression models is the determination of a suitable subset of covariates for each parameter. Hence, we resort to component-wise gradient-boosting to estimate the model coefficients. Instead of using the approach proposed by Hans et al. (2023), we conduct estimation in a two-step fashion using the `R` package `gamboostLSS`: First we boost each margin separately using a loss function for univariate right-censored responses and compute $\hat{S}_\bullet$, $\hat{f}_\bullet$, $\bullet = 1, 2$. In a second step we boost the loss corresponding to the following

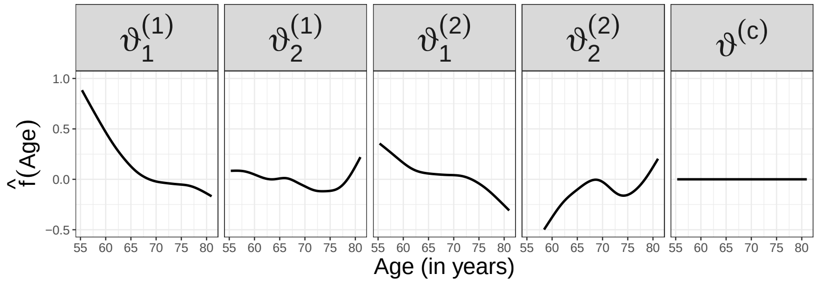


FIGURE 1. Estimated non-linear effect of age across distribution parameters.

log-likelihood function:

$$\begin{aligned}\ell_i = {} & \delta_{1i}\delta_{2i}\Big[\log(c(S_1(y_{1i}), S_2(y_{2i}))) + \log(f_1(y_{1i})) + \log(f_2(y_{2i}))\Big] + \\ & \delta_{1i}(1-\delta_{2i})\left[\log\left(\frac{\partial C(S_1(y_{1i}), S_2(y_{2i}))}{\partial S_1(y_{1i})}\right) + \log(f_1(y_{1i}))\right] + \\ & (1-\delta_{1i})\delta_{2i}\left[\log\left(\frac{\partial C(S_1(y_{1i}), S_2(y_{2i}))}{\partial S_2(y_{2i})}\right) + \log(f_2(y_{2i}))\right] + \\ & (1-\delta_{1i})(1-\delta_{2i})\Big[\log\Big(C(S_1(y_{1i}), S_2(y_{2i}))\Big)\Big], \end{aligned} \tag{3}$$

with $\hat{S}_\bullet$, $\hat{f}_\bullet$ plugged-in, which depends only on the association parameter $\vartheta^{(c)}$. Note that $c(\cdot,\cdot)$ denotes the copula density. The number of fitting iterations is optimised by means of the out-of-bag risk (Hans et al., 2023).

3 Results

The data we analysed consisted of $n = 629$ observations and three clinical covariates: An eye-specific severity score (`Severity`), the patient's age at study enrolment as well as a genetic covariate (`RS2284665`) encoded as a factor. The censoring rates were 46.7% and 44.5% for the AMD progression times for the left and right eye, respectively. We first determined the best-fitting margins by means of the log-score, which led to the log-normal distribution being selected for both margins. This preliminary result already suggests some symmetry regarding the statistical behaviour of the organs' AMD progression time. Using the optimal marginal fits, we compared different copula specifications (Gaussian, Frank, Clayton, Gumbel, Joe, as well as 90°, 180° and 270° rotations of the latter three) and based on the log-score we found the Clayton copula to provide the best predictive performance. See Table 1 for the estimated linear coefficients of the final model as well as the optimal number of fitting iterations of each of the five parameters of the joint survival function. The boosting algorithm did not include the highest `Severity` score level of each respective eye in the sub-model of the parameter $\vartheta_2^{(\bullet)}$. Additionally, the algorithm excluded the `Severity` score `5` from $\vartheta_2^{(1)}$'s sub-model. Figure 1 depicts the estimated non-linear effect of age on the parameters of the joint survival function. It can be observed that age was not

TABLE 1. Estimated coefficients of the Clayton copula model.

	Left eye log-normal		Right eye log-normal		Dependence Clayton
	$\vartheta_1^{(1)}$	$\vartheta_2^{(1)}$	$\vartheta_1^{(2)}$	$\vartheta_2^{(2)}$	$\vartheta^{(c)}$
Fitting iterations	312	156	106	58	7
`Intercept`	3.184	0.311	3.330	0.352	0.231
`Severity_1: 5`	−0.262	−0.050	–	–	0
`6`	−0.510	−0.077	–	–	0
`7`	−1.171	−0.093	–	–	0
`8`	−1.824	0	–	–	−0.037
`Severity_2: 5`	–	–	0	−0.069	0
`6`	–	–	−0.366	−0.367	0
`7`	–	–	−0.972	−0.171	0
`8`	–	–	−1.431	0	−0.023
`RS2284665: GT`	−0.150	−0.154	−0.209	−0.098	0
`TT`	−0.306	0.026	−0.353	−0.152	0
Censoring rates: 46.7% (left), 44.5% (right), $n = 629$.					

selected for the dependence model, whereas a similar shaped, downward-sloping effect of age can be seen on the parameter $\vartheta_1^{(\bullet)}, \bullet = 1, 2$. The effect of age on the parameter $\vartheta_2^{(\bullet)}, \bullet = 1, 2$ showed different forms depending on the margin. The estimated dependence in terms of Kendall's τ lies within $\hat{\tau} \in [0.392;\ 0.401]$, indicating a considerably strong dependence between the progression times of the eyes. Moreover, another important dependence measure obtained from our model is the estimated lower-tail dependence coefficient $\hat{\lambda}_U \in [0.584;\ 0.596]$, indicating that the margins have a strong dependence in the lower tail. In other words, progression times of AMD in the left and right eyes exhibit stronger dependence over time (i.e. at very low survival probabilities, since the survival function is monotonic decreasing in time). Note that the genetic covariate `RS2284665` and `Severity` scores below level `8` of each eye were excluded from the sub-model of the dependence parameter $\vartheta^{(c)}$. Figure 2 displays nine estimated joint survival functions according to Equation (1), based on the fitted Clayton copula model with log-normal margins for an hypothetical individual of median age (69.8 years), baseline expression of `RS2284665` and a combination of three levels of eye-specific `Severity` scores.

Panel (a) corresponds to both eyes having the lowest level of the `Severity` score (Left: 4 Right: 4), showing an optimistic progression prognosis (joint survival function is close to 1). Drops in the joint survival function are not as pronounced if only one eye deteriorates (`Severity` increases), see e.g. panels (b), (c), (d) or (g). The aforementioned panels show that a relatively high survival probability can still be expected from the healthy eye (larger extension of brighter color along one axis). In contrast, Figure 2 panel (i) shows a bleak prognosis for a patient whose left and right eyes have a `Severity` score of `8`, compare for example the values of the estimated joint survival function at 5 years for both eyes in panels (d) or (e) against (i).

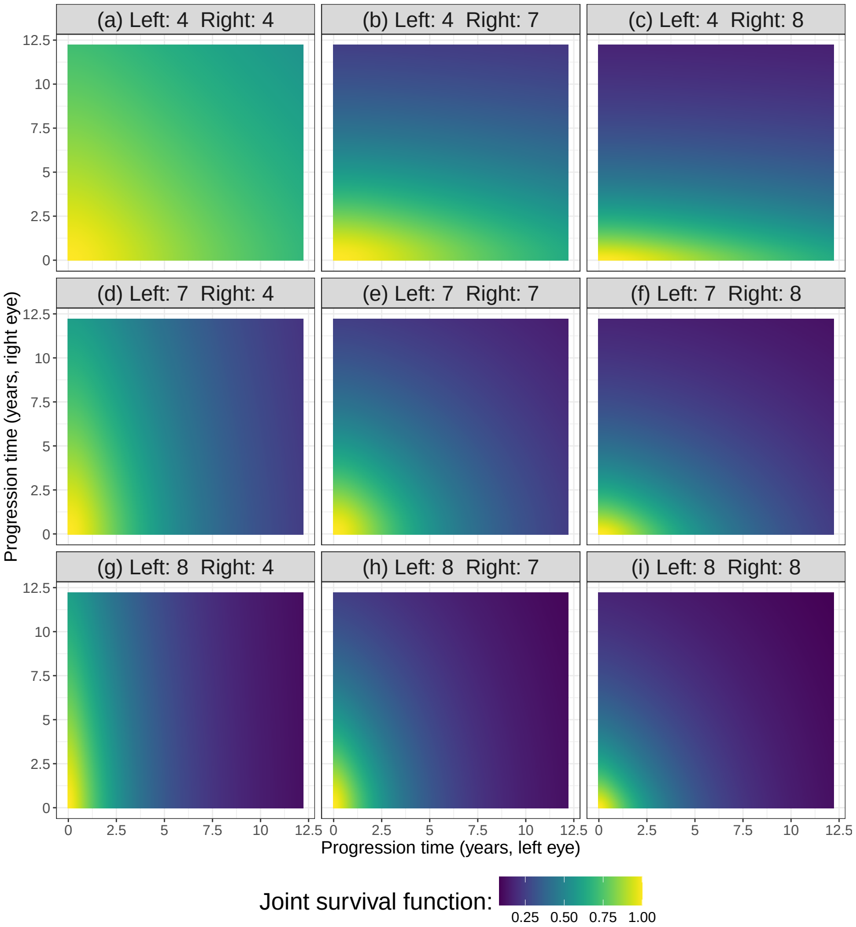


FIGURE 2. Estimated joint survival functions of AMD progression for an individual of median age (69.8 years), baseline expression of `RS2284665` and different combinations of `Severity` scores in the left and right eyes, respectively.

4 Discussion

We proposed a boosting algorithm to conduct data-driven variable selection for dependent bivariate right-censored time-to-event responses modelled using distributional copula regression techniques. Although not shown here, simulation studies indicate that our boosting estimation approach based on two-steps leads to better performance compared to that proposed by Hans et al. (2023), which resulted in independent margins in most scenarios. Our method exhibits a strong shrinkage effect on the dependence parameter $\vartheta^{(c)}$ relative to the parameters of the marginal survival functions, particularly given a large number of covariates. The shrinkage effect on all model coefficients was also affected by the censoring rate in the margins, thus higher censoring rates led to a stronger shrinkage effect.

Despite the fact that the data considered here was low-dimensional (in the number of covariates), we demonstrated how our proposed approach streamlined the model-building process. Potential avenues of future research involve modelling of cure fractions as well as the development of an implementation for general censoring schemes (left, right and interval) akin to Petti et al. (2022) in order to extend the applicability of our approach.

Acknowledgments: The work on this article was supported by the German research foundation (DFG) through the grants KL3037/2-1, MA7304/1-1 (428239776).

References

Hans, N., Klein, N., Faschingbauer, F., Schneider, M. and Mayr, A. (2023). Boosting distributional copula regression. *Biometrics*, **79**, 2298 – 2310.

Hofner, B., Mayr, A. and Schmid, M. (2016). `gamboostLSS`: An `R` package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software*, **74**, 1 – 31.

Nelsen, R. B. (2006). *An Introduction to Copulas.* Springer New York.

Petti, D., Eletti, A., Marra, G. and Radice, R. (2022). Copula link-based additive models for bivariate time-to-event outcomes with general censoring scheme. *Computational Statistics and Data Analysis*, **175**, 107550.

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

The AREDS Research Group. (1999). The age-related eye disease study (AREDS): Design implications AREDS report no. 1. *Controlled Clinical Trials*, **20**, 573 – 600.

Modelling social interaction data

Kevin Burke[1], James P. Gleeson[1], Mike Quayle[2,3]

[1] Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland
[2] Department of Psychology, University of Limerick, Limerick, Ireland
[3] Department of Psychology, University of KwaZulu-Natal, Pietermaritzburg, KwaZulu-Natal, South Africa

E-mail for correspondence: `kevin.burke@ul.ie`

Abstract: In this paper, we consider modelling approaches for social interaction data generated by participants (players) in the VIAPPL (Virtual Interaction APPLication) game environment. We make use of agent-based models to generate null distributions of interest, and also consider multinomial regression modelling.

Keywords: Agent-based model; Linear model; Network data; Multinomial model; Social interaction.

1 VIAPPL environment

In the field of psychology, it is of interest to study the emergence of social norms and group structure shaped by social interactions. The VIAPPL environment records the social interactions of participants (players) over a series of rounds in a controlled (virtual) game. Within a VIAPPL game, players interact with each other by exchanging "tokens". Each player starts with a set number of tokens, and, within each round, they select a player to whom they give a token (and they may select themselves). See Fennell et al (2023), upon which this conference paper is based.

Figure 1 displays the start (left panel) and end (right panel) of one round of a VIAPPL game. Each player is represented as a node, where the node with the thick border denotes the player on their own screen, i.e., Figure 1 displays the screen of a player who is located at the bottom of the on-screen diagram. At the start of a round, a player selects another player to whom they give a token. At this point, they can only see their own selection. Once all players have made their selections, all of these selections are revealed at the end of the round in a network diagram. Having gained this knowledge (of all exchanges), players then enter the next round. Note that players are allocated (randomly) to groups identified by the purple and green node colours.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

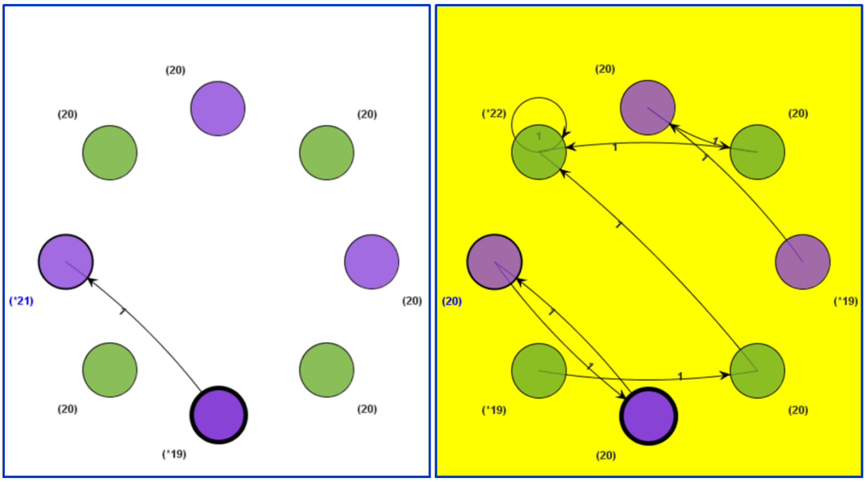


FIGURE 1. VIAPPL game: (left panel) a player selects another player to whom they give a token; (right panel) at the end of a round, all selections are revealed in a network diagram. (Figure adapted from Fennell et al (2023).)

2 Linear token exchange model

Let Y_{ij} denote the number of tokens player i has *received* from player j over the whole course of the game. Then, assume the linear model

$$Y_{ij} = \alpha + \rho Y_{ji} + \gamma G_{ij} \qquad i \neq j \tag{1}$$

such that the number of tokens that player i *receives* from player j (i.e., Y_{ij}) is related to the number of tokens *given* to player j (i.e., Y_{ji}), and also whether or not they are in the same group via the binary indicator G_{ij} (where $G_{ij} = 1$ means that they are *not* in the same group). This is a model for the weighted directed edges of the network formed by all token exchanges, where the edge directed from node j to node i is described by the edge directed from node i to j in combination with a group effect.

From the above model, ρ represents the level of reciprocity in the game, i.e., players giving and receiving tokens in pairs. In particular, if ρ is positive, this indicates the presence of reciprocal behaviour. Behaviour towards players in the opposite group is represented by γ, where a negative value indicates fewer token exchanges with this group, i.e., a preference for the player's own group.

3 Hypothesis testing via synthetic agent-based games

Equation (1) represents a somewhat non-standard statistical model for several reasons: the Y_{ij} counts play a role as both a response and as a covariate; there are constraints on these counts since each player has a fixed number of tokens, exchanging 1 per round, over a fixed number of rounds; we expect there to be high connectivity (dependence) between a small number of players within the same game.

Taken together, these features make it difficult to specify a reasonable likelihood function, and hence conduct inference. Therefore, we propose numerically generating the distribution of estimated parameters under a specified null hypothesis based on the results of an underlying agent-based model. Specifically, we simulate data with the same constraints as the original game (number of players, token exchanges per round, and rounds), where we set up some null behaviour within the simulated agents. A key behaviour of interest in this application area is "agents giving at random" (i.e., $\rho = \gamma = 0$), where player decisions are not influenced by tokens given to them or by the group assignment in the game.

Our setup is as follows.

1. Run an agent-based game, thereby generating synthetic counts Y^*_{ij} under some condition such as agents giving at random.
2. Fit model (1) to the synthetic data using least squares, producing estimates ρ^* and γ^* from the null distribution.
3. Repeat steps 1 and 2 a large number of times (e.g., 10,000) to construct the null distribution numerically.
4. Fit model (1) to the real data, producing estimates $\hat{\rho}$ and $\hat{\gamma}$ to be compared with the null distribution.

4 Results

We consider modelling data from 4 VIAPPL games spanning 40 rounds, where there are 14 different players in each of these games, and each player starts with 20 tokens. We apply the approach described in the previous section, leading to the table of estimates in Table 1.
We can see that the coefficients are numerically very close across games 1, 2, and 4, in each of which both the reciprocity and group effects are statistically significant (compared to a giving-at-random game). In these games, the reciprocity effects are positive and group effects are negative, meaning that players tend to develop reciprocal bonds and give more often to players in their own group.

TABLE 1. Estimated model parameters.

		Game 1		**Game 2**		**Game 3**		**Game 4**	
		Est.	p-val.	Est.	p-val.	Est.	p-val.	Est.	p-val.
Intercept	α	2.96	(0.78)	3.01	(0.66)	0.60	(<0.01)	2.77	(0.76)
Reciprocity	ρ	0.31	(<0.01)	0.29	(<0.01)	0.87	(<0.01)	0.37	(<0.01)
Group	γ	-1.95	(<0.01)	-1.96	(<0.01)	-0.42	(0.11)	-1.99	(<0.01)

Est. = estimate from the real data using least squares, p-val = p-value computed by comparing the estimates to the reference null distribution generated from the agent-based games.

The effects in game 3 differ from the other games, with a stronger reciprocity effect and a weaker group effect. This can largely be explained by two particular players from different groups who formed an unusually strong reciprocal bond. More generally, we have developed an approach for identifying influential players based on comparing the difference in estimated model coefficients upon replacing a player with a random-giving agent (similar to the classical "dfbeta" approach used throughout statistical modelling). Further details on this, and network visualisations of the results, can be found in Fennell et al (2023).

5 Another VIAPPL game and the multinomial model

We have also considered data from another VIAPPL experiment where players first agree/disagree with 4 different topics (rather than being placed in groups), and then, in one go (rather than over rounds) distribute their tokens. For these experiments, we have developed a multinomial regression model where the probability that player i gives a token to player j is proportional to $\exp(\alpha + \beta A_{ij})$ where $A_{ij} \in \{0, 1, 2, 3, 4\}$ is the number of topics on which the players agree. This alternative VIAPPL setup and model is omitted for brevity, but will be discussed further in the presentation.

Acknowledgments: The authors acknowledge funding from Science Foundation Ireland (JPG, grant numbers 16/IA/4470, 12/RC/2289 P2, 16/RC/3918; KB, grant number 16/RC/3918), and the European Research Council (ERC) (MQ, Grant agreement number 802421). For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Fennell, S.C., Gleeson, J.P., Quayle, M., Durrheim, K. and Burke, K. (2023). Agent-based null models for examining experimental social interaction networks. *Scientific Reports*, **13**, 5249.

Estimating dose and time of exposure from a protein-based radiation biomarker

Yilun Cai[1], Jochen Einbeck[1], Stephen Barnard[2], Elizabeth Ainsbury[2]

[1] Durham University, UK,
[2] UKHSA Radiation, Chemicals and Environmental Hazards Division, Chilton, Didcot, UK

E-mail for correspondence: `yilun.cai@durham.ac.uk`

Abstract: In order to analyze the potential damage to the human body caused by exposure to ionizing radiation, one needs to have an estimation of the dose of radiation received by the individual. In the context of a protein-based biomarker for radiation exposure, we present here a new method that, unlike the approaches that produce the estimation with data collected at a predetermined time after exposure, allows us to estimate the dose at any time within a reasonable time interval after exposure, as well as determine the time of exposure if needed. Namely, we take existing calibration curves and generalize them using the decay mechanism of γ-H2AX foci to build a model that describes the functional relationship between the count of γ-H2AX foci in exposed blood cells and the time and dose of exposure. This model is illustrated using both real and simulated data.

Keywords: Biological dosimetry, Poisson regression, Overdispersion.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

A boosted multistate model of partnership trajectories in Germany

Angela Carollo[1], Nicole Hiekel[1], Valeria Ferraretto[2]

[1] Max Planck Institute for Demographic Research, Rostock, Germany ,
[2] University of Trento, Trento, Italy

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

Abstract: Multistate models are an appropriate choice when analysing partnership trajectories over time. Building a correct model for these trajectories involves a series of non-trivial modelling decisions which can be automated in a data-driven way by using statistical boosting. Here, we use boosted multistate models to study partnership trajectories in Germany.

Keywords: Multistate models; Boosting; Partnership trajectories.

1 Introduction

Partnership trajectories involve a series of transitions between states like dating, cohabiting, marrying or separating. Understanding the complexity of these trajectories, and what differentiates individuals who break up their relationships from those who successfully advance the partnership are relevant questions in contemporary family demography research. An appropriate model to analyse partnership trajectories is a multistate model where individuals in a romantic relationship move across different states of the partnership over time. To investigate individual heterogeneity in transitions between these states, a suitable set of covariates is usually identified from the relevant literature and the available data. Usually, studies tend to focus on specific classes of predictors to avoid overfitting and collinearity issues, such as demographic variables, socio-economic factors or variables measuring relationships' quality.

Building a multistate model for such a complex process is not an easy task. On the one hand, the analyst aims to select a model that is simple enough to provide interpretable and general results, on the other hand, by doing so, some important predictors of the transitions of interest may be overlooked. Additional choices need to be made such as the choice of the time scale(s) to be used (clock-forward or clock-reset approach), the linear or non-linear effect of covariates, if any interaction between covariates should be considered, and whether some of

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

the covariates share the same effect on more than one transition, referred to as *cross-transition-type* effects.
Statistical boosting, combines the predictive power of machine learning approaches with interpretable statistical modelling techniques. It also provides a robust solution to multicollinearity issues (Mayr and Hofner, 2018). Hence, boosting might offer a clever solution to some of the problems that arise from building a complex multistate model. The idea of combining statistical boosting and multistate models was introduced by Reulen and Kneib in 2016. In a boosted multistate model cross-transitions-type effects, as well as non-linear effects, are selected automatically in a data-driven way. Here, we apply this boosting approach to multistate model to identify the best predictors of transitions between states of a partnership of young individuals in Germany.

1.1 Model specification

The transition rates in a multistate model can be formulated as:

$$\lambda_q(t) = \lambda_{0,q}(t)\exp(\eta)\,, \tag{1}$$

where t is the time scale, $q = 1, \ldots, Q$ indicates the specific transition, λ_0 is the baseline rate and η is a linear predictor built from covariates and their effects. The baseline hazard is estimated non-parametrically and the linear predictor is estimated through minimization of the negative log-stratified-partial likelihood w.r.t. η. Reulen and Kneib (2016) show that the negative log-statified partial likelihood is a valuable choice as loss function for the gradient boosting algorithm. We indicate with $x_{p.q,i} = x_{p,i} \cdot I_{trans_i=q}$ the transition-specific value of covariate x_p for individual i and transition q. In case only transition-specific covariates are included in the linear predictor η in an additive way, for individual i:

$$\eta_i = \sum_{p=1}^{P}\left(\sum_{q=1}^{Q} f_{x_{p.q}}(x_{p.q,i})\right)\,, \tag{2}$$

where $f_{x_{p.q}}$ is a function of the transition-specific covariate x_p for transition q. The boosting algorithm fits η at the same time as it selects the best $f_{x_{p.q}}$. Continuous covariates can be fitted with non-linear effect by decomposing $f_{x_{p.q}}(x_{p.q,i})$ as the sum of a linear part and a smooth deviation from linearity:

$$f_{x_{p.q}}(x_{p.q,i}) = f_{x_{p.q}}^{\text{linear}}(x_{p.q,i}) + f_{x_{p.q}}^{\text{smooth}}(x_{p.q,i}) \tag{3}$$

by ensuring that the covariates are centered before fitting and that the degrees of freedom of the base learners are set equal to 1.

2 A boosted model for partnerships trajectories

We model partnership trajectories of young women and men living in Germany. We specify the multistate model in Figure 1.
There are three transient states, one absorbing state and 5 possible transitions in this model, identified by the edges in Figure 1. The model is non-reversible, as each relationship is only allowed to move forward, and it can end either by *union dissolution* or as censored observation. Multiple relationships per individual are

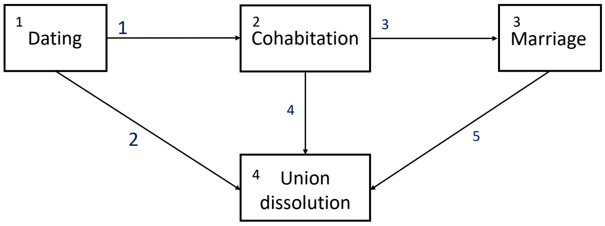


FIGURE 1. Multistate model of partnership trajectories.

allowed, and individuals can enter the model in any of the states *dating, cohabitation* and *marriage*. We excluded direct transitions from the state *dating* to the state *marriage* (without passing through the intermediate state of *cohabitation*) because they are extremely rare in our data.

We use data from the German family panel (pairfam) (Huinink et al, 2011), which is a longitudinal survey providing detailed information on romantic relationships of German individuals sampled from four different cohorts. At each wave, respondents and their partner are asked a set of questions concerning their relationship, their values, their satisfaction and fertility plans, as well as updated information on their relationship status and socio-economic status.

Such rich data provides the perfect opportunity to study predictors of transitions between partnership states and investigate heterogeneity in these transitions. In order to do so, we fit a boosted multistate model using the R-package `mboost` (Hofner et al. (2014)) and specifying `family = multistate()` from the add-on-package `gamboostMSM` (Reulen (2022)).

We consider 307 base-learners of transitions-specific covariates and let the algorithm select the set of best predictors for each transition.

Our preliminary results, presented in Figure 2 and in Table 1, show that the set of predictors selected by the boosting algorithm is rather small. In total, the algorithm selected 31 transition-specific covariates, of which 16 continuous or numerical covariates represented in Figure 2 (two with both linear and non-linear effect, two with only non-linear effects and the other 12 with linear effect only), and 15 dummy variables shown in Table 1. The estimated effects of the numerical variables are all in the expected direction. For example, increasing levels of satisfaction with the relationship are associated with decreasing risk of union dissolution. Similarly, the presence of children is associated with lower risk of dissolving a cohabitation or a marriage, while higher importance of being in a partnership is associated with increasing risk of moving to a cohabitation. Among the stronger estimated effects there is having definite plans for moving-in together and having marriage plans, which increase the risk of experiencing both events respectively, while not having any plans is associated with lower risk of both events. Being infertile (or missing information on infertility) are associated with negative partnership's transitions, while higher education level of both the respondent and the partner is associated with positive transitions.

In the next steps of our analysis, we will use consider different cross-transition-type effects, non-linear effect of continuous variables, possibly time-varying effects and interactions effects. We will also refine the initial choices of the base-learners and the tuning of the model. Finally, we plan to estimate the standard errors

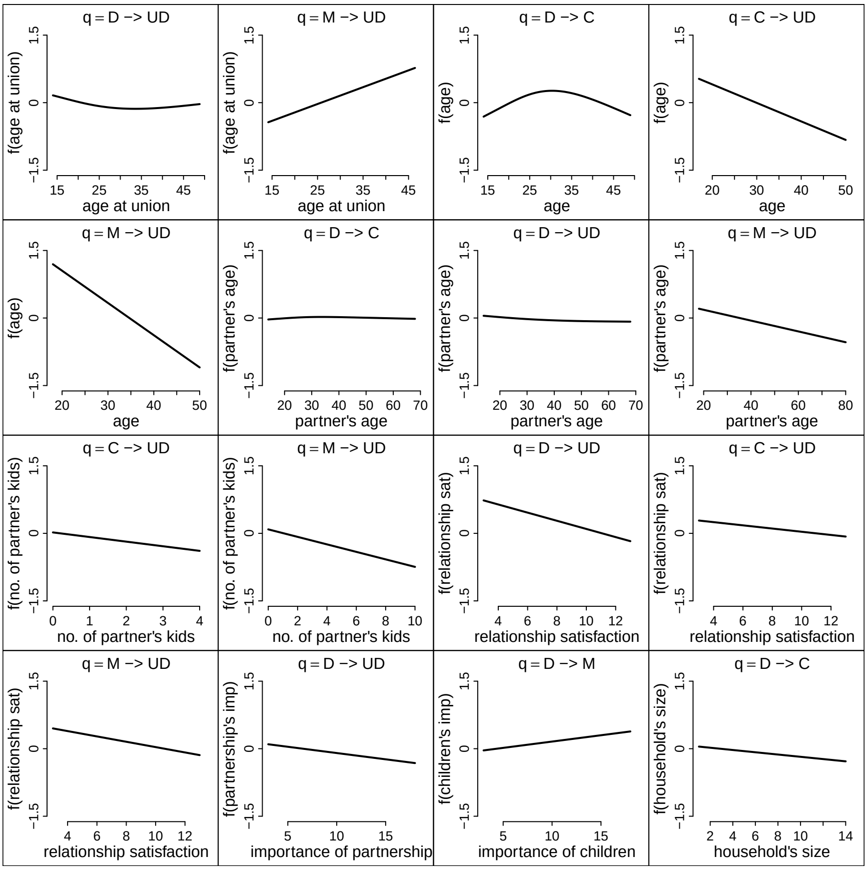


FIGURE 2. Selected continuous predictors of partnership transitions, with linear and non-linear effects. The title of each plot indicates the transition for which the predictor has been selected.

TABLE 1. Transition-specific covariate's effects.

Covariate	D→C	D→UD	C→M	C→UD
infertile/incomplete	-0.108	0.017		0.118
education: high	0.036			-0.032
partner's educ.: high	0.005			-0.021
partner's educ.: missing		-0.359		
plans for coh.: yes	1.25	-0.047		
plans for coh.: no	-0.05			
plans for mar.: yes			0.875	-0.117
plans for mar.: no			-0.653	
household inc.: 2000+				-0.201

via bootstrapping, which would also correct for extra heterogeneity introduced by observing multiple relationships per individual.

Acknowledgments: We are incredibly grateful to *Guillermo Briseno-Sanchez* for the help with navigating the treacherous waters of statistical boosting and the helpful comments to this abstract.

References

Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Computational Statistics*, **29**, 3-–35.

Huinink, J., Brüderl, J., Nauck, B., Walper, S., Castiglioni, L. and Feldhaus, M. (2011). Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. *Zeitschrift für Familienforschung - Journal of Family Research*, **23**, 77 – 101.

Mayr, A. and Hofner, B. (2018). Boosting for statistical modelling - A nontechnical introduction. *Statistical Modelling*, **18**, 365 – 384.

Reulen, H. and Kneib, T. (2016). Boosting multi-state models. *Lifetime Data Analysis*, **22**, 241 – 262.

Reulen, H. (2022). R-package `gamboostMSM`: Boosting Multistate Models.

Nonparametric Bayesian modeling of nonstationary joint extremes

Miguel de Carvalho[1,2], Vianey Palacios Ramirez[3]

[1] School of Mathematics, University of Edinburgh, UK
[2] Department of Mathematics, University of Aveiro, Portugal
[3] School of Mathematics, Statistics and Physics, Newcastle University, UK

E-mail for correspondence: Miguel.deCarvalho@ed.ac.uk

Abstract: We propose a novel Bayesian model for inferring about the intensity of observations in the joint tail over time, and for assessing if two stochastic processes are asymptotically dependent. To model the intensity of observations exceeding a high threshold, we develop a Bayesian nonparametric approach that defines a prior on the space of what we define as EDI (Extremal Dependence Intensity) functions. An application of the proposed methodology to a set of big tech stocks—known as FAANG—sheds light on some interesting features on the dynamics of their combined losses over time.

Keywords: FAANG stocks; Mixture of finite Polya trees; Nonparametric prior; Nonstationary extremal dependence; Statistics of extremes.

1 Introduction

Record-breaking extreme events—such as stock market crashes, widespread flooding, wildfires and heatwaves—call for a better understanding and quantification of their risk. Extreme value theory offers a sound probabilistic and statistical setup for dealing with those challenges, given its ability to extrapolate into the tails of a distribution (e.g., Coles, 2001). In a multivariate context, the degree of association between the extreme observations of a random vector with common margins, (X, Y), is often evaluated by,

$$\chi = \lim_{z \to \infty} P(X > z \mid Y > z). \tag{1}$$

The measure χ quantifies the probability of X being extreme, given that Y is extreme. If $0 < \chi \leq 1$ the variables are asymptotically dependent (AD), whereas if $\chi = 0$ they are said to be asymptotically independent (AI).

In this paper we develop a Bayesian model for learning about the *intensity of extreme observations* of a random vector over time, as well as for assessing if two stochastic processes are asymptotically dependent. As it will be shown below, our methods have links with a time-varying version of (1) (i.e., $\chi(t) = \lim_{z\to\infty} P(X_t > z \mid Y_t > z)$).

2 Modeling time-changing joint extremes

Framework. Let $\{(X_t, Y_t)\}_{t\in[0,T]}$ be a sequence of independent random vectors, and following standard practice in extreme value theory suppose that $\{X_t\}$ and $\{Y_t\}$ are unit Fréchet distributed, i.e., $P(X_t < z) = P(Y_t < z) = \exp(-1/z)$, with $z > 0$ for all t.

For AD processes, their degree of dependence can be characterized by what we will refer to as the EDI (Extremal Dependence Intensity) function,

$$f(t) = \frac{\chi(t)}{\int_0^T \chi(\tau)\,\mathrm{d}\tau} = \frac{\lim_{z\to\infty} P(Z_t > z)}{\int_0^T \lim_{z\to\infty} P(Z_\tau > z)\,\mathrm{d}\tau}, \tag{2}$$

where $Z_t = \min(X_t, Y_t)$. The EDI carries information on the intensity of observations in the joint tail over time, $A = [u,\infty)^2 \times [0,t]$ for u large. This follows from the fact that for a sufficiently large u,

$$f(t) \propto \lim_{z\to\infty} P(Z_t > z) \approx P(Z_t > u) = \frac{\mathrm{d}\Lambda(A)}{\mathrm{d}t}, \tag{3}$$

since the intensity measure

$$\Lambda(A) = E\left(\int_0^t J_\tau\,\mathrm{d}\tau\right) = \int_0^t P(Z_\tau > u)\,\mathrm{d}\tau,$$

as $J_\tau = \mathbb{1}_{\{Z_\tau > u\}} \sim \mathrm{Bern}\{P(Z_\tau > u)\}$, where $\mathbb{1}$ is the indicator function.

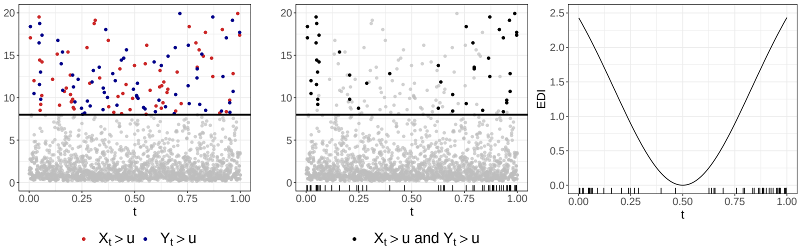


FIGURE 1. Left: Simulated data above threshold from a time-varying Gumbel copula. Middle: Rug of times of exceedances of $Z_t = \min(X_t, Y_t)$ above threshold and corresponding exceedances. Right: EDI.

A flat EDI, $f(t) \propto 1$, indicates a constant intensity of joint extremes over time, whereas a peaking EDI signals higher intensity in that period. See Fig. 1 an illustration.

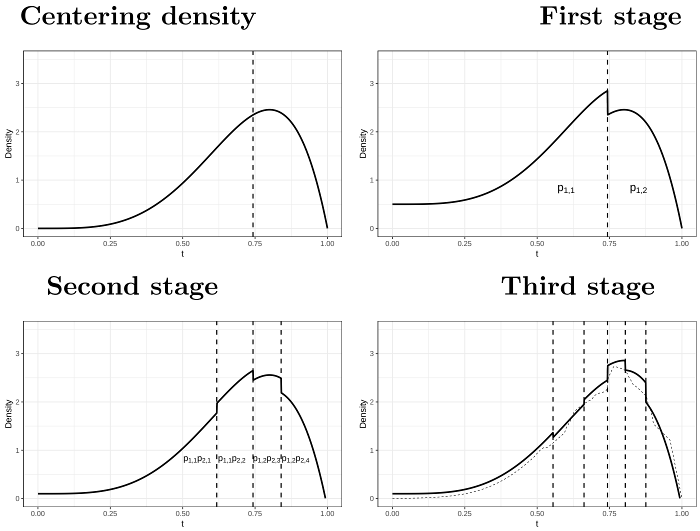
Learning from data. Bayesian inference for the EDI function involves defining a prior over the space of EDI functions. Our prior consists of a mixture of finite

Polya trees (Hanson, 2006). A Polya tree of level J can be regarded as an extension of a parametric model; see Fig. 2 for an illustration.

Let $I = \{t/T : Z_t > u\} = \{\tau_1, \ldots, \tau_k\}$ be the standardized times of $k = o(T)$ joint observations exceeding a high threshold. The hierarchical representation of our model for the EDI is as follows

$$I \mid F \sim F, \quad F \sim \mathrm{PT}_J(\alpha, F_{0,\theta}), \quad \theta \sim \mathsf{p}(\theta). \tag{4}$$

Here, $\mathrm{PT}_J(\alpha, F_{0,\theta})$ is a Polya tree with two parameters: A centering cumulative EDI ($F_{0,\theta}(t)$); a precision parameter ($\alpha > 0$). The parameter α controls how much deviations from the centering are allowed, in the sense that the smaller the α the more one allows for deviations from the centering.



Centering density

First stage

Second stage

Third stage

FIGURE 2. Example of Polya tree densities centred at a Beta$(5, 2)$ density over stages 1–3; the third stage also shows a mixture of Polya trees mixing over $a \sim \mathrm{LN}(\log 2, .05)$ and $b \sim \mathrm{LN}(\log 5, .05)$. The dashed line represents the quantiles defining the bins.

3 Data illustration

We now apply the proposed methods to track the dynamics governing extreme joint losses of FAANG (Meta's **F**acebook, **A**pple, **A**mazon, **N**etflix and Alphabet's **G**oogle) stocks. These stocks have attracted retail investors, money managers, and other professional stakeholders. Since our focus is on extreme losses, we use weekly negative returns as a unit of analysis.

As can be seen from Fig. 3, most EDIs tend to peak around 2016–19, thus indicating that extreme joint losses have occurred mostly around that time. From a financial outlook the dynamics portrayed by the EDI in Fig. 3 may look surprising at first, keeping in mind that the 2020 pandemic crisis has led to some sharp sell-offs worldwide. And in fact economists have painted a doomsday scenario for the real economy in the short-run since early 2020. Yet, Fig. 3 simply claims that the *relative frequency* of extreme joint losses has been higher over 2016–2019, than over the 2020 pandemic outbreak. Many geopolitical issues (e.g., US–China trade war) and US policy issues (e.g., former President Trump impeachment) may have been the drivers for some of these joint sell-offs over 2016–19.

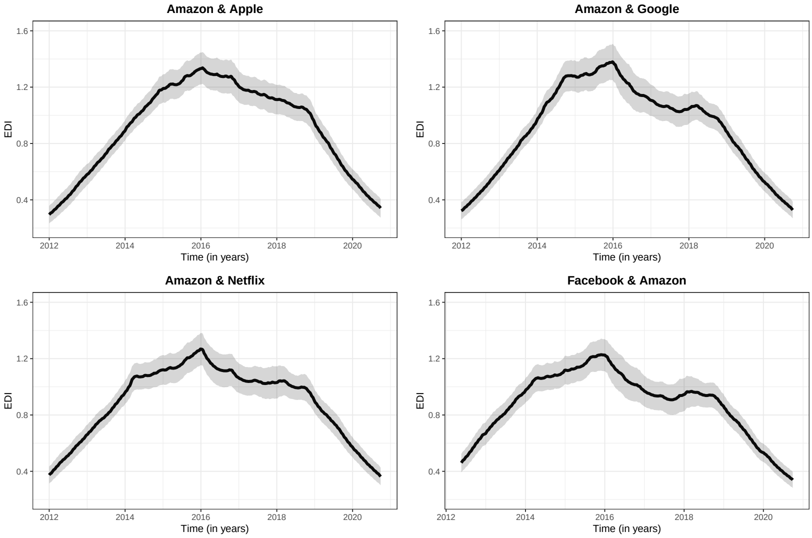


FIGURE 3. Pairwise EDI for FAANG stocks: Posterior mean of EDI based on a mixture of finite Polya trees along with pointwise credible bands.

Acknowledgments: This work was funded by the Royal Society of Edinburgh and by CIDMA

`https://doi.org/10.54499/UIDB/04106/2020`
`https://doi.org/10.54499/UIDP/04106/2020`

References

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Hanson, T.E. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, **101**, 1548 – 1565.

Approximated Gaussian random field under different parameterizations for MCMC

Joaquin Cavieres[1], Cole C. Monnahan[2], David Bolin[3], Elisabeth Bergherr[1]

[1] Chair of Spatial Data Science and Statistical Learning, , University of Göttingen, Germany

[2] National Marine Fisheries Service (NOAA), United States

[3] King Abdullah University of Science and Technology (KAUST), Saudi Arabia

E-mail for correspondence: `joaquin.cavieres@uni-goettingen.de`

Abstract: Fitting spatial models with a Gaussian random field as spatial random effect poses computational challenges for Markov Chain Monte Carlo (MCMC) methods, primarily due to two factors: computational speed and convergence of chains for the hyperparameters. To deal with this, a Gaussian random field can be approximated by a Gaussian Markov random field using stochastic partial differential equations. This methodology is commonly used in "latent Gaussian models", where the inference is done by the Integrated Nested Laplace Approximations, but rarely used in an MCMC method. In this contribution, we evaluated different parameterizations of the approximated Gaussian random field, specifically using the Hamiltonian Monte Carlo algorithm of the Stan software. A simulation study demonstrated that models using the hyperparameters ρ and σ_u were better able to estimate the values used to simulate the spatial random field. Their speed computation were faster compared to models parameterized with κ and τ. In real data application, the index of relative abundance estimated for Pollock indicates similar trends for the six models proposed. However, models incorporating ρ and σ_u demonstrated faster computation compared to those utilizing κ and τ, corroborating the results found in the simulation. Even more important, none of these models encountered convergence issues, as indicated by the Rhat statistic.

Keywords: Approximated Gaussian random field; Index of relative abundance; Bayesian spatial modelling.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

Integrating single index effects in generalized additive models

Claudia Collarin[1], Matteo Fasiolo[1]

[1] School of Mathematics, University of Bristol, Bristol, UK

E-mail for correspondence: claudia.collarin@phd.unipd.it

Abstract: Linearly combining the elements of a vector of covariates to get a scalar-valued feature is common practice in regression modelling. In this work, we propose a novel approach to integrate single index effects in Generalised Additive Models (GAMs). In particular, model fitting and inference are performed by exploiting the efficient methods proposed in Wood et al (2016) [*JASA* **111**, 1548-1563]. We consider an application to daily electricity load consumption data, demonstrating improved predictive performance relative to traditional GAMs. This integrated approach provides a valuable tool to capture complex relationships in real-world applications, while preserving interpretability.

Keywords: Generalised additive models; Single index models; Projection pursuit regression.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Combining biomarkers through PCA

Tahani Coolen-Maturi[1]

[1] Department of Mathematical Sciences, Durham University, UK

E-mail for correspondence: `tahani.maturi@durham.ac.uk`

Abstract: The area under the receiver operating characteristic curve (AUC) is a useful tool for evaluating the ability of a diagnostic test to distinguish between two groups or classes. In practice, multiple diagnostic tests or biomarkers are often combined to improve diagnostic accuracy. This paper utilises principal component analysis (PCA) to identify the best linear combination of multiple biomarkers to enhance diagnostic accuracy, specifically by maximising the AUC.

Keywords: Diagnostic accuracy; Combining biomarkers; AUC; PCA.

1 Introduction

Measuring the accuracy of diagnostic tests is crucial in many application areas, including medicine, machine learning and credit scoring. The receiver operating characteristic (ROC) curve is a useful tool for assessing the ability of a diagnostic test to discriminate between two classes. However, one diagnostic test may not be enough to draw a useful decision; thus, in practice, multiple biomarkers may be combined to improve diagnostic accuracy (Pepe and Thompson, 2000). This paper uses PCA to improve diagnostic accuracy by identifying the best linear combination of biomarkers to maximise AUC. PCA is a statistical technique used for data visualisation and dimensionality reduction in various fields, including machine learning and signal processing. Suppose that X is a continuous random quantity of a diagnostic test result and that larger values of X are considered more indicative of disease. X^1 and X^0 are used to refer to test results for the disease and non-disease groups, respectively. The ROC curve is defined as $\{(\text{FPF}(c), \text{TPF}(c)), c \in (-\infty, \infty)\}$, where $\text{FPF}(c) = P(X^0 > c)$ and $\text{TPF}(c) = P(X^1 > c)$. The area under the ROC curve, AUC, is a useful summary that measures the overall performance of a diagnostic test. Higher values indicate more accurate tests, with $\text{AUC} = 1$ for perfect tests and $\text{AUC} = 0.5$ for uninformative tests. Consider test data from the disease group $\{x_1^1, \ldots, x_{n_1}^1\}$ and from the non-disease group $\{x_1^0, \ldots, x_{n_0}^0\}$, with the two groups fully independent.Thus,

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

the empirical AUC is given by

$$\widehat{\mathrm{AUC}} = \frac{1}{n_1 n_0} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \left[\mathbf{1}\left\{x_i^1 > x_j^0\right\} + \frac{1}{2}\mathbf{1}\left\{x_i^1 = x_j^0\right\} \right].$$

2 Methods

A practical question that often arises is how to effectively combine information from multiple biomarkers to accurately differentiate between diseased and non-diseased groups. The concept of linearly combining biomarkers to improve diagnostic accuracy by maximising the AUC has been explored by Su and Liu (1993) under the assumption of normality, while others have proposed distribution-free approaches, e.g., Pepe and Thompson (2000) and Coolen-Maturi (2017). Consider a set of p biomarkers, $X_1, X_2, \ldots, X_p$, with their linear combination $Y = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_p X_p$. The goal is to find optimal values for $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ to maximise the AUC and, consequently, enhance the diagnostic accuracy. To this end, we search for the best $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ values by evaluating the AUC for the combined test $\sum_j \alpha_j X_j$ across 101 equally spaced values for each $\alpha_j \in [0, 1]$, $j = 1, \ldots, p$ such that $\sum_j \alpha_j = 1$. Various methods have been proposed in the literature to determine the optimal values with different restrictions, e.g., Pepe and Thompson (2000). However, Coolen-Maturi (2017) has shown that when dealing with uncorrelated or weakly correlated tests, implementing the above restriction when combining biomarkers leads to greater improvement compared to other restrictions. This makes it suitable for combining biomarkers using PCA, as all principal components are uncorrelated from one another. More discussion about the use of this restriction and its advantages is given in Coolen-Maturi (2017). The principal components can be written as linear combinations of $X_1, X_2, \ldots, X_p$ $(j = 1, \ldots, p)$ as

$$PC_j = a_{j1} X_1 + a_{j2} X_2 + \ldots + a_{jp} X_p$$

The first principal component is the linear combination of $X_1, X_2, \ldots, X_p$ that has maximum variance (among all linear combinations); that is, it accounts for as much variation in the data as possible. These coefficients for the first component are obtained to maximise its variance, subject to $\sum_{j=1}^{p} a_{1j}^2 = 1$. The second principal component has the second maximum variance, and so on. So the aim of this paper is to find the optimal $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)^T$, that maximises the empirical AUC, where $\gamma_j \in [0, 1]$, and $\sum_{j=1}^{q} \gamma_j = 1$, $q \leq p$, that is

$$T = \gamma_1 PC_1 + \gamma_2 PC_2 + \ldots + \gamma_p PC_q$$

To combine multiple biomarkers, the biomarker measurements should be comparable, or some form of normalisation should be conducted.

3 Application

In a study, blood samples were collected from 120 patients, with 82 normal and 38 carriers, to screen carriers of a genetic disorder. Four measurements, M_1, M_2, M_3, M_4, were taken, transformed to a logarithmic scale, and standardised

(Cox et al., 1982). The empirical AUC for these biomarkers are $\widehat{AUC}_1 = 0.9034$, $\widehat{AUC}_2 = 0.7526$, $\widehat{AUC}_3 = 0.8232$, and $\widehat{AUC}_4 = 0.8789$. So, M_1 outperformed other biomarkers, followed by M_4, then M_3, and M_2. M_1 has a strong correlation with M_3 and M_4 ($r = 0.64$), and M_3 has a strong correlation with M_4 ($r = 0.56$), whereas M_2 has weak correlation ($r < 0.3$) with other measurements. After PCA, the extracted loadings are:

$$\begin{aligned}
PC_1 &= 0.56M_1 + 0.26M_2 + 0.55M_3 + 0.56M_4\\
PC_2 &= -0.30M_1 + 0.94M_2 - 0.13M_3 - 0.01M_4\\
PC_3 &= 0.07M_1 - 0.07M_2 - 0.73M_3 + 0.68M_4\\
PC_4 &= 0.77M_1 + 0.19M_2 - 0.39M_3 - 0.47M_4
\end{aligned}$$

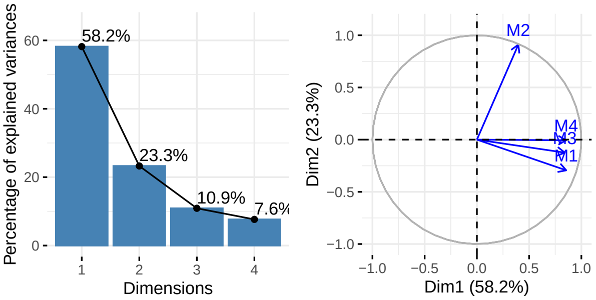


FIGURE 1. Scree plot of the components (left) and Biplot of the biomarkers with respect to the principal components (right).

It seems that M_1, M_3, and M_4 tend to move together, contributing positively and similarly to PC_1. PC_2 is primarily influenced positively by M_2 and negatively by M_1 to a lesser extent. M_3 and M_4 have weaker contributions to PC_2. M_3 and M_4 have similar projections onto PC_3 but with opposite signs. Finally, PC_4 is primarily influenced positively by M_1 and, to some extent, negatively by M_3 and M_4. Figure 1 indicates that PC_1 alone explains 58.2% of the variance, and together with PC_2, they explain 81.5%. In Table 1, we used different combinations of these PCs to find the optimal values that maximise the empirical AUC. Using PC_1 alone yields $\widehat{AUC} = 0.9445$, higher than any other biomarker alone. When considering two PCs together, adding PC_1 improves performance over using it alone; however, the best improvement is achieved by combining PC_1 and PC_4 ($\widehat{AUC} = 0.9564$), followed by PC_1 and PC_2 ($\widehat{AUC} = 0.9506$). Similarly, when considering three PCs, the best performance is achieved by combining PC_1, PC_2 and PC_4 ($\widehat{AUC} = 0.9615$). Combining all PCs results in a minor improvement ($\widehat{AUC} = 0.9618$), with more weights on PC_1 and PC_4.

TABLE 1. Blood samples data set, combining biomarkers through PCA

Principal Components	$\hat{\gamma}_{opt}$	$\widehat{AUC}$
PC_1	(1, 0, 0, 0)	0.9445
PC_2	(0, 1, 0, 0)	0.5257
PC_3	(0, 0, 1, 0)	0.5273
PC_4	(0, 0, 0, 1)	0.5870
$PC_1 + PC_2$	(0.78, 0.22, 0, 0)	0.9506
$PC_1 + PC_3$	(0.95, 0, 0.05, 0)	0.9464
$PC_1 + PC_4$	(0.68, 0, 0, 0.32)	0.9564
$PC_2 + PC_3$	(0, 0.52, 0.48, 0)	0.5597
$PC_2 + PC_4$	(0, 0.11, 0, 0.89)	0.5931
$PC_3 + PC_4$	(0, 0, 0.26, 0.74)	0.6088
$PC_1 + PC_2 + PC_3$	(0.76, 0.20, 0.04, 0)	0.9519
$PC_1 + PC_2 + PC_4$	(0.47, 0.15, 0, 0.38)	0.9615
$PC_1 + PC_3 + PC_4$	(0.59, 0, 0.08, 0.33)	0.9586
$PC_2 + PC_3 + PC_4$	(0, 0.21, 0.29, 0.59)	0.6178
$PC_1 + PC_2 + PC_3 + PC_4$	(0.51, 0.10, 0.10, 0.29)	0.9618

References

Coolen-Maturi, T. (2017). Predictive inference for best linear combination of biomarkers subject to limits of detection. *Statistics in Medicine*, **36**, 2844 – 2874.

Cox, L.H. and Johnson, M.M. and Kafadar, K. (1982) Exposition of statistical graphics technology. *ASA Statistical Computing Section*, 55 – 56.

Liu, A. and Schisterman, E.F. and Zhu, Y. (2005). On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine*, **24**, 37 – 47.

Su, J.Q. and Liu, J.S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, **88**, 1350 – 1355.

Pepe, M.S. and Thompson, M.L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, **1**, 123 – 140.

Non-parametric frailty model for the natural history of prostate cancer; using data from a screening trial

Ilse Cuevas Andrade[1], Ardo van den Hout[1], Nora Pashayan[2]

[1] University College London, UK
[2] Department of Public Health and Primary Care, University of Cambridge, UK

E-mail for correspondence: ilse.andrade.21@ucl.ac.uk

Abstract: Mixed-effects models for survival, known as frailty models, can be used to capture individual or cluster-specific unobserved heterogeneity. A common choice is assuming the random effects follow a parametric distribution, e.g. a normal distribution. However, computing the marginal likelihood can be computationally expensive and infeasible in a high-dimension setting. Alternatively, a non-parametric approach avoids the normality assumption for random effects and can be less computationally demanding. The data used are from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. The aim is to model prostate cancer progression to evaluate screening strategies, considering the PSA longitudinal biomarker, accounting for unobserved heterogeneity, interval-censored, left-truncation, and right-censored data.

Keywords: Mixed models; Precision Matrices; Sparse Matrix Algebra.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Species sampling with misidentification: a Bayesian parametric approach

Davide Di Cecco[1], Andrea Tancredi[2]

[1] Università Unitelma Sapienza, Italy
[2] Università di Roma La Sapienza, Italy

E-mail for correspondence: andrea.tancredi@uniroma1.it

Abstract: We introduce a new species sampling model to account for species misidentification. The assumption is that each sampled individual or specimen has an unknown probability of being misclassified. Misclassified units constitute fictitious cases that artificially inflate the number of species observed only once. We consider standard parametric models for the true number of occurrences of each species, and present a Gibbs sampler algorithm to perform Bayesian inference. The proposed model is applied to a real-world microbial diversity dataset.

Keywords: Capture-recapture models; Species richness problem; Bayesian inference; Thinned processes.

1 Introduction

Consider a sample of individuals from a population, and a procedure classifying each individual into distinct "species". The problem of the estimation of the total number N of distinct species in the population based on such a sample is often termed the "species sampling problem" and it is commonly tackled by assuming that the classification procedure is error-free. In this framework, let X_i^*, for $i = 1, \ldots, N$, be the number of sampled individuals belonging to the $i-th$ species. Unobserved species correspond to $X_i^* = 0$. Let $n_j^* = \sum_{i=1}^N I(X_i^* = j)$ be the number of species with j sampled occurrences. Then, n_0^* represents the number of unobserved species that must be estimated to obtain $N = n_0^* + \sum_{j\geq 1} n_j^*$. Inference can be conducted by specifying a model for the random variable X_i^* for $i = 1, \ldots, N$, and standard assumptions are that X_i^*, for $i = 1, \ldots, N$, are independent and identically distributed observations from a Poisson, or more realistically, a mixed-Poisson distribution (see, for instance, Bunge et al. (2014)). Misidentification of units has not been commonly addressed in species sampling problems although the problem has received some attention in the similar framework of capture-recapture. In animal abundance estimation, errors in identification have been studied in different areas. For example, abundance estimates based

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

on photo–identification heavily depend on the photographic quality and distinctiveness of markings on the body of the animals. In genotype-based identification, animal samples are collected in an area and analyzed to extract DNA. It is commonly assumed that genotyping errors lead to fictitious genotypes that cannot be associated to other existing cases, thus erroneously increasing the number of single captures, which lead to an overestimation of the population size.
The latter assumptions are adopted in the so-called $M_{t,\alpha}$ capture-recapture model, (see Link et al (2010)). This model generalizes the classical M_t model by assuming that unit misidentification occurs independently and with the same probability α at the different sampling occasions. In addition, misidentified units cannot be confused with other population units with the effect of creating a *ghost* (i.e., fictitious) capture history with precisely one capture.

2 The missing link model

In this Section, we outline our model for the species sampling problem with misidentification. We assume that the quantities X_i^* for $i = 1, \dots, N$ represent the latent number of captures in our sample for a certain species i we would have observed without identification errors. Let $f^*(\cdot;\theta)$ be the baseline distribution of the error-free captures, i.e. $f^*(j;\theta) = P(X_i^* = j)$. The generating mechanism of our model is the following: for each species i with X_i^* captured individuals/specimens, we have a latent number M_i, $(M_i \le X_i^*)$ of unidentified individuals (missing links) such that we erroneously count $X_i = (X_i^* - M_i)$ occurrences for species i. The observed data are represented by the vector $(n_1, n_2, \dots)$ where

$$n_1 = \sum_{i=1}^{N} I(X_i = 1) + \sum_{i=1}^{N} M_i \quad n_j = \sum_{i=1}^{N} I(X_i = j) \quad j \ge 2.$$

Notice that we implicitly assume that each time an individual is not correctly identified, this creates a new non-existing species observed exactly once.
To complete the model we assume that $M_i | X_i^* = x_i^*$ is Binomial(x_i^*, μ) so that the marginal distribution of X_i can be generally written as

$$\tilde{f}(j;\theta,\mu) = P(X_i = j \mid \theta, \mu) = \sum_{m_i=0}^{\infty} \binom{j+m_i}{m_i} \mu^{m_i} (1-\mu)^j f^*(j+m_i;\theta) \quad j \ge 0.$$

Straightforward calculations show that assuming for f^* a Poisson(λ) distribution, the resulting distribution $\tilde{f}$ is Poisson$(\lambda(1-\mu))$ and M_i is Poisson$(\mu\lambda)$ and it is independent on X_i. Similarly, taking for f^* a Negative Binomial(r, p) distribution, the resulting distribution $\tilde{f}$ is Negative Binomial$(r, p/(1-\mu(1-p)))$ and $M_i | X_i = x_i$ is Negative Binomial$(x_i + r, 1 - \mu(1-p))$.

2.1 Bayesian inference

In the following, we show how to perform Bayesian inference under the missing link model when the baseline distribution f^* is Poisson(λ). Let $\mathbf{n}$ be the observed data vector $(n_1, ..., n_j, ...)$ and let $\tilde{\mathbf{n}}$ be the vector $(\tilde{n}_0, \tilde{n}_1, ..., \tilde{n}_j, ...)$, where $\tilde{n}_j = \sum_{i=1}^{N} I(X_i = j)$ is the number of species with j identified occurrences. Note that

$N = \sum_{j\geq 0} \tilde{n}_j$ and $n_j = \tilde{n}_j$ for $j \geq 2$. We can exploit a Gibbs sampler algorithm to simulate from the posterior distribution

$$p(\tilde{n}_0, \tilde{n}_1, \lambda, \mu|\mathbf{n}) \; \propto \; p(n_1|\tilde{\mathbf{n}}, \lambda, \mu) p(\tilde{\mathbf{n}}|\lambda, \mu) p(\lambda) p(\mu). \qquad (1)$$

Note that, given $\tilde{\mathbf{n}}$, we know that there are $N = \sum_{j=0}^{\infty} \tilde{n}_j$ species. Moreover, since $n_1 = \tilde{n}_1 + M$, where $M = \sum_{i=1}^{N} M_i$, and each M_i is Poisson($\mu\lambda$) independently on $\tilde{\mathbf{n}}$, we have that

$$n_1|\tilde{\mathbf{n}}, \lambda, \mu \sim \tilde{n}_1 + Poisson(N\lambda\mu).$$

As for the second factor of the right hand side of (1), we have:

$$p(\tilde{\mathbf{n}}|\lambda, \mu) = \binom{N}{\tilde{n}_0, \tilde{n}_1, \ldots, \tilde{n}_j, \ldots} \prod_{j=0}^{\infty} \left(\frac{e^{-\lambda(1-\mu)} (\lambda(1-\mu))^j}{j!} \right)^{\tilde{n}_j} p(N).$$

We adopt a Gamma(a, b) prior for λ, a Beta(f, g) prior for μ, and $p(N) \propto 1/N$ as a prior for N. Then, the full conditional for λ, is a Gamma$(s + a, N + b)$, where s is the total number of captures $\sum_{j>0} j n_j$. The full conditional distribution of μ is Beta$(n_1 - \tilde{n}_1 + f, s - (n_1 - \tilde{n}_1) + g)$. Setting $n_{2+} = \sum_{j=2}^{\infty} n_j$, the probability mass function of the full conditional distribution for $\tilde{n}_0$ is given by

$$p(\tilde{n}_0|\cdots) \propto e^{-\mu\lambda(\tilde{n}_0+\tilde{n}_1)} (\tilde{n}_0 + \tilde{n}_1 + n_2)^{n_1 - \tilde{n}_1} \frac{(\tilde{n}_0 + \tilde{n}_1 + n_{2+})!}{\tilde{n}_0!(\tilde{n}_0 + \tilde{n}_1 + n_{2+})} (e^{-\lambda(1-\mu)})^{\tilde{n}_0}$$

where $\tilde{n}_0 \geq 0$. The full conditional distribution for $\tilde{n}_1$ is given by

$$\begin{aligned} p(\tilde{n}_1|\cdots) \quad &\propto \quad \frac{e^{-\mu\lambda(\tilde{n}_0+\tilde{n}_1)} (\tilde{n}_0 + \tilde{n}_1 + n_2)^{n_1-\tilde{n}_1}}{(n_1 - \tilde{n}_1)!} \frac{(\tilde{n}_0 + \tilde{n}_1 + n_{2+})!}{\tilde{n}_1!(\tilde{n}_0 + \tilde{n}_1 + n_{2+})} \\ &\times \quad (e^{-\lambda(1-\mu)} \lambda(1-\mu))^{\tilde{n}_1} \end{aligned}$$

where $0 \leq \tilde{n}_1 \leq n_1$.

3 Microbial diversity application

In this Section, we apply the missing link model to a real-world dataset of marine microbial diversity study. The data summarized in Figure 1 (top-left panel), were analyzed in Hong et al. (2006). The detected number of singletons species is 381, which constitutes 74% of the observed species and 39% of the observed specimens. Since both the sequencing process and the clustering process utilized to construct the data might have been subject to errors, the proposed missing link model represents a valuable option for analyzing these data. To highlight the potentiality of our approach, in Figure 1 we report the posterior distribution of N (top-right panel) and the posterior distribution of μ (bottom-left panel) obtained with the Gibbs sampler scheme described in the previous Section. Finally, we also show the posterior distribution of Shannon's diversity index $E = \exp\left(-\sum_{j\geq 1} n_j^* \frac{j}{s} \ln \frac{j}{s}\right)$ (bottom right panel).

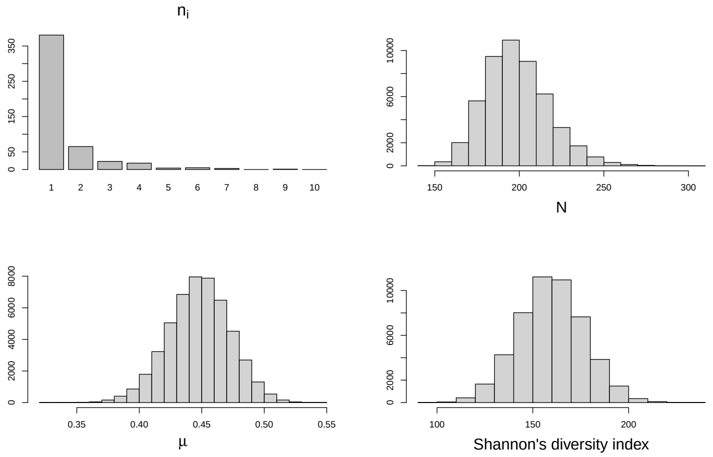


FIGURE 1. Microbial diversity application: data set and posterior distributions for N, μ and the Shannon's diversity index.

4 Conclusions

We have considered several other parametric families as baseline for our model, e.g. the lognormal Poisson distribution, Bulmer (1974), the inverse Gaussian Poisson distribution, Ord and Whitmore (1986), the Poisson-Lindley distribution, Pathak et al. (2024), the Consul's generalized Poisson and Conway-Maxwell-Poisson distributions, Anan et al (2017). However, it is not just as easy to adapt our algorithm to other parametric choices. In fact, in our MCMC scheme we take advantage of the simple form of the distribution of the thinned counts X_i, and of the conditional distribution of $M_i|X_i$ given their sum, and it appears that those variables do not have a simple distribution under the aforementioned cases. Of course, it is possible to adopt different algorithms, such as those based on the approximate Bayesian computation (ABC) techniques, as illustrated in Di Cecco and Tancredi (2024).

References

Anan, O., Böhning, D. and Maruotti, A. (2017). Uncertainty estimation in heterogeneous capture–recapture count data. *Journal of Statistical Computation and Simulation*, **87**, 2094 – 2114.

Bulmer, M. G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, **30**, 101 – 110.

Bunge, J., Willis, A. and Walsh, F. (2014). Estimating the number of species in microbial diversity Studies *Annual Review of Statistics and Its Application*, **1**, 427 – 445.

Di Cecco, D. and Tancredi, A. (2024) Estimating the number of sequencing errors in microbial diversity studies. *Environmental and Ecological Statistics.*

Hong, S. H., Bunge, J., Jeon, S. O. and Epstein, S. S. (2006). Predicting microbial species richness. *Proceedings of the National Academy of Sciences*, **103** , 117–122.

Link, W. A., Yoshizaki, J., Bailey, L. L. and Pollock, K. H. (2010). Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. *Biometrics,* **66** , 178 – 185.

Ord, J. K. and Whitmore, G. A. (1986). The Poisson-inverse gaussian disiribuiion as a model for species abundance. *Communications in Statistics - Theory and Methods*, **15**, 853 – 871.

Pathak, A., Kumar, M., Singh, S. K., Singh, U. and Kumar, S. (2024). Bayesian estimation of the number of species from Poisson-Lindley stochastic abundance model using non-informative priors. *Computational Statistics*, https://doi.org/10.1007/s00180-024-01464-7.

gateTree: A user-informed tree algorithm for population identification in flow cytometry

Ultán P. Doherty[1], Rachel McLoughlin[2], Tracey Claxton[2], Arthur White[1]

[1] School of Computer Science and Statistics, Trinity College Dublin, University of Dublin, Ireland

[2] Host-Pathogen Interactions Group, Trinity Biomedical Sciences Institute, School of Biochemistry and Immunology, Trinity College Dublin, University of Dublin, Ireland

E-mail for correspondence: dohertyu@tcd.ie

Abstract: Flow cytometry is a high-throughput single-cell analysis technique for immunological research. Manual gating has traditionally been used to identify cell populations in cytometry. Data clustering algorithms have also been developed for this purpose. We present a semi-supervised decision tree algorithm which utilises user-provided information to implement population-specific variable selection, outlier removal, and pruning. We apply our algorithm to a flow cytometry data set and show that it can outperform state-of-the-art cytometry clustering algorithms.

Keywords: Flow cytometry; Decision tree; Gating; Semi-supervised clustering.

1 Introduction

Flow cytometry allows an immunological researcher to indirectly measure the expression level of a range of protein markers for every cell in a tissue sample. The identification of cell populations in cytometry data is crucial as it facilitates downstream analysis of how specific cell types differ between subjects from different experimental groups. Population identification has traditionally been carried out using a manual approach called gating.

To perform manual gating, a researcher selects a pair of variables on which to view a scatterplot of the data, and then uses software to draw a polygonal boundary around the population of interest. Often, boundaries differentiate between cells with high or low values for one or both of these variables. The subset within the boundary can be further refined by repeating the process with another pair of variables. This process is carried out for every cell population in the data set. Manual gating is time-consuming and is becoming increasingly infeasible as the

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

number of variables which can be analysed by modern cytometers continues to grow. Manually drawing boundaries can also lead to reproducibility issues.

2 Method

Algorithm 1: Tree Construction

```
$\mathcal{P}$ := Cell populations, $\mathcal{X}$ := Individual cells, $\mathcal{V}$ := Variables.
$T(p,\ v) \in \{-1,\ 0,\ +1\}$ := Table value
$x_v$ := Value of cell $x$ for variable $v$.
$(l_v,\ u_v)$ := Lower and upper cutoff values for variable $v$.
for $p$ in 1 to $|\mathcal{P}|$
    $s \leftarrow 0;\ p' \leftarrow p;\ \mathcal{P}_p^{(0)} \leftarrow \mathcal{P}$
    $\mathcal{V}_p^{(0)} \leftarrow \{v \in \mathcal{V} : T(q,v) \neq 0\ \forall\ q \in \mathcal{P}\}$
    $\mathcal{X}_p^{(0)} \leftarrow \{x \in \mathcal{X} : x_v \in (l_v,\ u_v)\ \forall\ v \in \mathcal{V}_p^{(0)}\}$
    $\mathcal{W}_p^{(0)} \leftarrow \{\}$
    while $p' = p$
        Select a split variable, $v^* \in V_p^{(s)}$, and a split location, $y^*$.
        $\mathcal{W}_p^{(s+1)} \leftarrow \mathcal{W}_p^{(s)} \cup \{v^*\}$
        if the split selection algorithm is successful
            Remove populations with conflicting descriptions:
            $\mathcal{P}_p^{(s+1)} \leftarrow \{q \in \mathcal{P}_p^{(s)} : T(q,v^*) = T(p,v^*)\}$
            Identify new variables for which these populations are defined:
            $\mathcal{V}_p^{(s+1)} \leftarrow \{v \in \mathcal{V} : T(q,v) \neq 0\ \forall\ q \in \mathcal{P}_p^{(s+1)}\} \setminus \mathcal{W}_p^{(s+1)}$
            Remove cells which are on the wrong side of $y^*$:
            $\mathcal{X}_p^{(s+1)} \leftarrow \{x \in \mathcal{X}_p^{(s)} : \mathrm{sign}(x_{v^*} - y^*) = T(p,v^*)\}$
            Remove cells which are outside the cutoffs:
            $\mathcal{X}_p^{(s+1)} \leftarrow \{x \in \mathcal{X}_p^{(s+1)} : x_v \in (l_v,\ u_v)\ \forall\ v \in \mathcal{V}_p^{(s+1)}\}$
            $s \leftarrow s + 1$
        else
            $p' \leftarrow p + 1$
        end if else
    end while
    $\mathcal{P}_p \leftarrow \mathcal{P}_p^{(s)};\ \mathcal{X}_p \leftarrow \mathcal{X}_p^{(s)}$
end for
The unique elements of $\{\mathcal{X}_p :\ p \in \mathcal{P}\}$ are the final clusters.
```

Algorithm 2: Split Selection

```
for $v$ in $\mathcal{V}_p^{(s)}$
    $\hat{f}_v$ := Kernel density estimate of $\mathcal{X}_p^{(s)}$ for variable $v$.
    peak($v$) := Global maximum of $\hat{f}_v$.
    peak($v,1$), ..., peak($v,M$) := Other local maxima of $\hat{f}_v$.
    valley($v,m$) := Minimum of $\hat{f}_v$ between peak($v$) and peak($v,m$).
    depth($v,m$) := $(\text{peak}(v,m) - \text{valley}(v,m)) \times \frac{100}{\text{peak}(v)}$
    $m^* \leftarrow^{\arg\max}_{m} \{\text{depth}(v,m)\}$
    valley($v$) ← valley($v,m^*$); depth($v$) ← depth($v,m^*$)
end for
if $^{\max}_{v} \{\text{depth}(v)\}$ > minimum depth
    $v^* \leftarrow^{\arg\max}_{v} \{\text{depth}(v)\}$ ; $y^* \leftarrow \text{valley}(v^*)$
else
    for $v$ in $\mathcal{V}_p^{(s)}$
        boundary($v,1$), ..., boundary($v,R$) ← regularly spaced points.
        BIC($v,r$) := BIC of a two-component univariate GMM with
          components defined by boundary($v,r$).
        BIC($v$) := BIC of a one-component univariate GMM.
        $\text{diff}(v,r) \leftarrow (\text{BIC}(v,r) - \text{BIC}(v)) \Big/ \left(2\log\left|\mathcal{X}_p^{(s)}\right|\right)$
        $r^* \leftarrow^{\arg\max}_{r} \{\text{diff}(v,r)\}$
        boundary($v$) ← boundary($v,r^*$); diff($v$) ← diff($v,r^*$)
    end for
    if $^{\max}_{v} \{\text{diff}(v)\}$ > minimum diff
        $v^* \leftarrow^{\arg\max}_{v} \{\text{diff}(v)\}$ ; $y^* \leftarrow \text{boundary}(v^*)$
    else
        Split variable, $v^*$, and split location, $y^*$, not identified.
    end if else
end if else
```

To apply the gateTree algorithm, the user must describe a set of cell populations of interest as positive, negative, or neutral / undefined for a selection of variables. We refer to a population as positive (negative) for a given variable if it has a high (low) expression level. The algorithm recursively partitions the cells based on whether their values for one of the selected variables are higher or lower than the threshold constructed by the algorithm to optimally split that variable. In particular, the sequence in which the variables are split is designed to identify the described populations.

gateTree partitions a selected variable either at the deepest density valley of a univariate kernel density estimate or at the optimal boundary between two univariate Normal distributions, according to BIC. If several of the variables which gateTree is trying to split have viable density valleys, then the variable whose valley has the greatest depth is chosen. If none of these variables have viable density valleys, then the algorithm attempts to use mixture model boundaries instead. If several of these variables have viable mixture models, then it chooses the one with the best BIC score.

A detailed description of the procedure followed by gateTree is provided in Algorithm 1 and Algorithm 2. It is implemented in an open-source R package available at https://github.com/UltanPDoherty/gateTree.

We refer to gateTree as a semi-supervised clustering algorithm because it uses minimal descriptions of the populations to be identified. However, the term semi-supervised is often used to refer to algorithms which use a subset of true class labels as training data. gateTree does not require any of the observations to be labelled and does not require any training data.

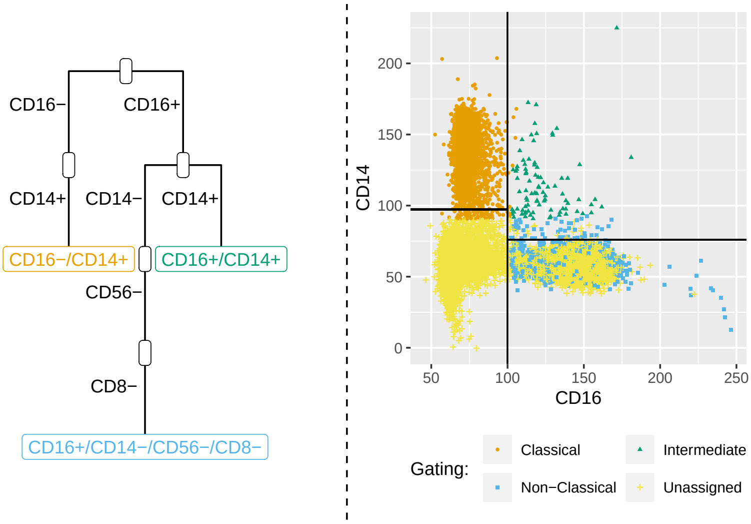


FIGURE 1. Left: Tree diagram showing how gateTree partitioned the data for the Monocyte pathway of the haemodialysis data set. Right: A CD14 vs CD16 scatterplot of the haemodialysis data set coloured according to the manual gating for the Monocyte pathway. The CD16 and CD14 splits constructed by gateTree for this pathway are also displayed.

We will use the Monocyte gating pathway from the haemodialysis data set discussed in Section 3 as an illustrative example. The user information for this pathway described three populations of monocytes: Classical (CD16-CD14+), Intermediate (CD16+CD14+), and Non-Classical (CD16+CD14-CD56-CD8-). CD16, CD14, CD56, and CD8 are variables in the data set. The Classical Monocytes are described as CD16-CD14+ because they have negative and positive expression levels for CD16 and CD14, respectively.

The tree diagram in Figure 1 illustrates how gateTree partitions the haemodialysis data for the Monocyte pathway. Based on the user-provided information, the first split has to be on either CD16 or CD14, as these are the variables which the descriptions of the three populations have in common. Any observations with extreme values for the CD16 or CD14 variables, according to user-provided cut-off thresholds, were automatically allocated to the "Unassigned" subset. After this first split, the Classical Monocytes (CD16-CD14+) lie on the CD16- branch, while the Intermediate (CD16+CD14+) and Non-Classical (CD16+CD14-CD56-CD8-) Monocytes lie on the CD16+ branch. Since the only population on the CD16- branch is the Classical Monocytes (CD16-CD14+) and it is only described with respect to CD16 and CD14, the next split for this branch must be on CD14. The resulting CD16-CD14- branch is immediately pruned because there are no user-described populations lying on it. All observations on the pruned branch are moved to the "Unassigned" subset. Meanwhile, the CD16-CD14+ branch cannot be split further without subdividing the Classical Monocytes. We can now associate the observations on this branch with the Classical Monocyte population. The process used to identify this subset of observations featured user-informed and population-specific variable selection, outlier removal, and pruning.

3 Application and results

To demonstrate gateTree's performance, we applied it to a flow cytometry data set from a haemodialysis study. We only included the pre-gated single-cell subset of 32,624 observations and the 9 fluorescence channels. We also applied two publicly available population identification algorithms: cytometree (Commenges et al., 2018) and FlowSOM (van Gassen et al., 2015). For cytometree's AIC parameter, we ran t = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, and 1. For FlowSOM's number of metaclusters parameter, we ran nClus = 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, and 50.

The manual gating consisted of four gating pathways with gates from different pathways allowed to overlap. The user information utilised for the Monocyte pathway is described in Section 2. Each of the other three pathways consisted of two populations with similar user descriptions. For every pathway, we ran cytometree and FlowSOM both on the full data set and on the subset of variables which featured in that pathway's manual gating. For every pathway, the performance of each algorithm was evaluated using the unweighted mean F_1 measure with respect to the manual gating.

TABLE 1. Maximum unweighted mean F1 values per pathway for each algorithm.

	B & CD8+	γ-δ	Monocyte	NK
gateTree	0.984	0.991	0.864	0.938
cytometree (All)	0.976	0.281	0.775	0.511
cytometree (Selected)	0.980	0.025	0.876	0.866
FlowSOM (All)	0.955	0.924	0.584	0.685
FlowSOM (Selected)	0.975	0.988	0.627	0.852

Table 1 shows that, for each of the four pathways, a single clustering solution constructed by gateTree was able to compete with or outperform the best-performing clustering solution from each other algorithm. This is true even when some of the information that gateTree utilises was provided to the other algorithms via variable selection.

4 Conclusion

Our user-informed tree algorithm outperformed two state-of-the-art population identification algorithms. gateTree achieved this strong performance by using a tree structure to mimic manual gating's sequential subsetting approach and by utilising information about the populations of interest.

References

Commenges, D., Alkhassim, C., Gottardo, R., Hejblum, B. and Thiébaut, R. (2018). cytometree: A binary tree algorithm for automatic gating in cytometry analysis. *Cytometry Part A*, **93**, 1132–1140.

Lee, H. C., Kosoy, R., Becker, C. E., Dudley, J. T. and Kidd, B. A. (2017). Automated cell type discovery and classification through knowledge transfer. *Bioinformatics*, **33**, 1689—1695.

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N. et al. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, **87**, 636-645.

Derivatives of the log of a determinant

Paul H.C. Eilers[1], Martin P. Boer[1]

[1] Wageningen University and Research, The Netherlands

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

Abstract: We present an efficient way to calculate effective model dimensions, using automated differentiation of the Cholesky algorithm. The method is illustrated with two examples using P-splines: adaptive smoothing and smoothing of over-dispersed counts.

Keywords: Choleksy decomposition; Sparse linear algebra; Automated differentiation, Effective model dimension.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

A biclustering approach via mixture of latent trait analyzers for the analysis of digital divide in Italy

Dalila Failli[1], Bruno Arpino[2], Maria Francesca Marino[1], Francesca Martella[3]

[1] Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Firenze, Italy
[2] Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Padua, Italy
[3] Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Rome, Italy

E-mail for correspondence: `dalila.failli@unifi.it`

Abstract: We employ an extension of the Mixture of Latent Trait Analyzers (MLTA) model to analyse the digital divide in Italy in a biclustering perspective. In detail, units (individuals) are partitioned into clusters (components) via a finite mixture of latent trait models; in each component, variables (digital skills) are partitioned into clusters (segments) by modifying the linear predictor's specification of the original MLTA model. This allows us to identify homogeneous groups of individuals with respect to subsets of digital skills, also accounting for the influence of demographic features on the probability of being digitally skilled.

Keywords: Model-based clustering; Co-clustering; Finite mixtures; Latent variables; EM algorithm.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Variance partitioning-based priors for species distribution models

Luisa Ferrari[1], Massimo Ventrucci[1]

[1] University of Bologna, Italy

E-mail for correspondence: luisa.ferrari5@unibo.it

Abstract: Species distribution models for community ecology data are usually quite complex because of the need to account for many abiotic factors, with potentially non-linear effects, as well as residual spatio-temporal correlation, which capture abiotic phenomena. The use of variance partioning-based priors recently emerged in the literature could be an effective and intuitive strategy to deal with the high flexibility often required in this field. In this work, we discuss how to extend this new class of priors to species distribution models containing spatial and temporal smooth effects.

Keywords: Bayesian species distribution models; Intuitive priors; IGMRF.

1 Introduction

Surveys in the field of community ecology collect large datasets on the abundance of different species at certain locations and time points. Multiple factors are believed to influence abundance patterns. Species distribution models (SDM) are often expressed as generalized linear mixed models (GLMM) with fixed effects for the *abiotic* factors and random effects capturing the residual spatio-temporal correlation, reflecting the so-called *biotic* phenomena (e.g. predator-prey abundance cycles, species' spatial segregation, symbiotic or competitive relationships). Further complexity arises in a joint SDM framework, where several approaches to model between-species correlation structures have been proposed including latent variable models (Tikhonov et al. (2020)) and spatio-temporal basis functions (Hui et al. (2023)). All these approaches involve richly-parametrized GLMMs that require regularization to avoid overfitting.
Regardless of the chosen approach, it is often the case that ecologists have prior insight into the relative importance of each factor in explaining the response. As a consequence, a Bayesian approach would be particularly beneficial in these applications to impose a regularization based on prior information. We argue that thinking in terms of quantities like proportions of variance due to the individual model components is more intuitive than considering the original variance

parameters. This can be achieved using *variance partitioning* (VP)-based priors (Franco-Villoria et al. (2022), Fuglstad et al. (2020)) which make use of a reparametrization of the variance parameters of a mixed model into a total variance and a simplex vector containing the proportional contributions to the total variance from each model component.

The common advantage of VP priors consists in the fact that it is much easier to introduce prior information in the model using these new parameters. One can easily implement very different types of prior knowledge on the variance contributions of the different model components, based on what is known about the case study at hand. As an example, assuming a Uniform distribution on the simplex would reflect ignorance a priori about the relative importance of each term, while a Dirichlet inducing sparsity on the proportions of variance would provide a suitable solution to perform variable selection in sparse linear regression. Furthermore, a hierarchical decomposition of the total variance through subsequent splits can be chosen to favour shrinkage towards simpler model structures (Franco-Villoria et al. (2022), Fuglstad et al. (2020)).

The VP-based priors proposed so far only deal with specific effects, e.g. stationary or linear effects. Challenges arise in their extension to complex models, such as SDMs which often contain smooth effects of continuous covariates as well as Intrinsic Gaussian Markov random fields (IGMRFs) for spatial and time effects. The goal of this paper is to develop a unified VP framework applicable to more complex settings, such as SDMs.

2 Proposal

Consider the following SDM in which the linear predictor of a generic abundance response can be written as an additive model of P linear effects for the $X_1, \ldots, X_P$ covariates, a smooth effect over spatial coordinates (S_1, S_2), and another smooth effect for time T. The smooth effects are both expressed using a finite-dimensional basis, $\mathbf{B}_S(\cdot)$ and $\mathbf{B}_T(\cdot)$, and a corresponding set of coefficients, $\mathbf{u}$ and $\mathbf{v}$ respectively:

$$\eta = \mu + \sum_{p=1}^{P} X_p \beta_p + \mathbf{B}_S(S_1, S_2)^{\mathrm{T}}\mathbf{u} + \mathbf{B}_T(T)^{\mathrm{T}}\mathbf{v}. \tag{1}$$

A latent Gaussian model is assumed on all coefficient sets, i.e. they are specified as Normally distributed with 0 mean and fixed precision matrix conditional upon a single scale parameter: $\beta_p|\sigma_p^2 \sim N(0, \sigma_p^2) \quad p = 1, \ldots, P$, $\mathbf{u}|\sigma_S^2 \sim N(\mathbf{0}, \sigma_S^2 \mathbf{Q}_S^{-1})$, $\mathbf{v}|\sigma_T^2 \sim N(\mathbf{0}, \sigma_T^2 \mathbf{Q}_T^{-1})$. The VP parameters can then be defined as:

$$V = \sum_{p=1}^{P} \sigma_p^2 + \sigma_S^2 + \sigma_T^2 \qquad \boldsymbol{\omega} = \left[\frac{\sigma_1^2}{V}, \ldots, \frac{\sigma_P^2}{V}, \frac{\sigma_S^2}{V}, \frac{\sigma_T^2}{V}\right] \tag{2}$$

The great advantage of VP-based priors comes from the possibility of assigning priors directly on the total variance in the linear predictor (i.e. V) and the set of proportions of variance due to each effect (i.e. $\boldsymbol{\omega}$). However, it is not guaranteed that these intuitive interpretations actually match the VP parameters in (2). This only occurs if all model components in (1) are processes on a comparable, *standardized* scale so that the elements of $\boldsymbol{\omega}$ actually represent the corresponding variance contributions.

For linear effects, it is sufficient to use the standardized version of X_p for $p = 1, ..., P$. However, it is not as simple for effects defined using a generic basis matrix, e.g. the spatial and temporal effects in this model. We propose a scaling procedure inspired by the work of Sørbye and Rue (2014) on IGMRFs that guarantees that the parameter of V and $\boldsymbol{\omega}$ match their intuitive interpretation. This is achieved by scaling each of the bases in the model by the square root of a term-specific constant C defined as the variance of the corresponding process conditional on $\sigma^2 = 1$ and marginalizing over the covariates' distribution. For example, the constant for the temporal effect is defined as:

$$C_T = \int_{t \in \mathcal{T}} \mathbf{B}_T(t)^{\mathrm{T}} \mathbf{Q}_T^{-1} \mathbf{B}_T(t) \cdot \pi(t) \; \mathrm{d}t \tag{3}$$

where $\mathcal{T}$ is the support of interest for variable T and $\pi(t)$ is its probability distribution. C_S is analogously defined using a given $\mathcal{S}$ support and $\pi(s_1, s_2)$ density. This scaling procedure can be viewed as a generalization of the standardization procedure used for linear effect, as C simplifies to the variance of the corresponding covariate in this case. We argue that VP-based priors can be safely employed only after scaling each term in the model according to this procedure. An advantage of the scaling procedure lies in the possibility of immediately evaluating the variance partition structure of the model considering the posterior distribution of the $\boldsymbol{\omega}$ vector. This is possible because after scaling each entry will represent the proportional contribution of a model component to the response variability. A challenging aspect in the scaling constant definition in Equation 3 is that it requires the choice of a distribution $\pi(\cdot)$ for the corresponding covariate. While it is reasonable to assume a Uniform distribution over the spatio-temporal support, this becomes a non-trivial choice in the case in which the procedure must be applied to other types of effects, such as smooth effects of continuous covariates.

3 Application

3.1 Data

The model defined in Equation (1) is applied to the NOAA-NEFSC fall bottom trawl survey dataset, studied in Hui et al. (2023) and publicly available at: https://github.com/fhui28/CBFM. The survey contains presence/absence data for 39 fish species from $N = 5892$ different space-time locations in the North-West Atlantic region, spanning a 20-year period. Figure 1 shows the study region with the number of species found in each location. Information about 5 environmental covariates is also available: surface temperature and salinity, bottom temperature and salinity, depth. A binary variable indicating the type of vessel collecting the data at each location can be used as an additional covariate.

3.2 Model and results

The model of Equation 1 is applied to each of the 39 species from the survey to illustrate how the proposed method provides a simple and intuitive way to study the contributions of different factors on the variability of an occurrence response. A logistic model is chosen to link the linear predictor to the binary presence-absence response for each species. The five environmental covariates

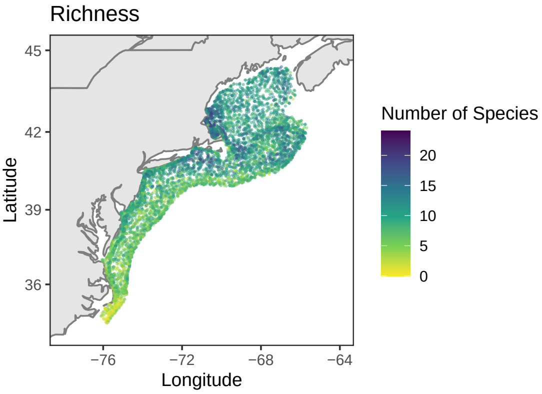


FIGURE 1. Number of different species, i.e. richness, detected in each of the locations from the survey.

and the vessel dummy are entered into the model with linear effects, following standardization. A 2-dimensional B-Spline basis with an Intrinsic CAR model (Besag et al. (1991)) precision matrix is used for the spatial effect, whose knots are equally spaced on a grid of 50x50km cells. A B-Spline with 20 basis functions is chosen for the temporal effect, with a 1^{st} order random walk prior on the coefficients. A Uniform distribution is assumed over the observed spatio-temporal support for the computation of the scaling constants C_S and C_T.
In this case study, the VP-based prior approach is used to reflect the assumption that not all effects are likely to affect the abundance of each species, but rather a few (species-specific) factors are assumed to be responsible for most of the variability. This assumption can be introduced through the choice of a symmetric Dirichlet prior on the vector of proportions: $\boldsymbol{\omega} \sim \text{Dir}(0.5)$. The marginal prior induced on each of the $\boldsymbol{\omega}$ elements is represented as a solid black line in the left panel of Figure 2: as we can see, this prior assigns most probability mass near 0 indicating that $\omega_j = 0$ (no effect) is favoured a priori. The prior specification is completed by a vague prior on the intercept μ and a Jeffreys on V.
The models are fitted using the R-INLA software. Thanks to scaling, the posterior distribution of $\boldsymbol{\omega}$ can directly answers questions about variance partitioning without further transformations. The left panel of Figure 2 shows the marginal posterior distributions of the proportions of variance $\boldsymbol{\omega}$ entries for a single species (*Weakfish*). The plot shows how the prior choice helps in the identification of the most important factors affecting occurrence as most factors are shrunk towards 0. The right panel shows the posterior median of $\boldsymbol{\omega}$ for six different species. Along with conclusions about individual species, this plot can help assess the variance partitioning for the community as a whole: for example, the spatial component appears to be a relevant term for all the species in this subset.

4 Discussion

This work proposes a new way to analyze SDMs that can incorporate prior knowledge about the relative importance of different factors affecting species abundance

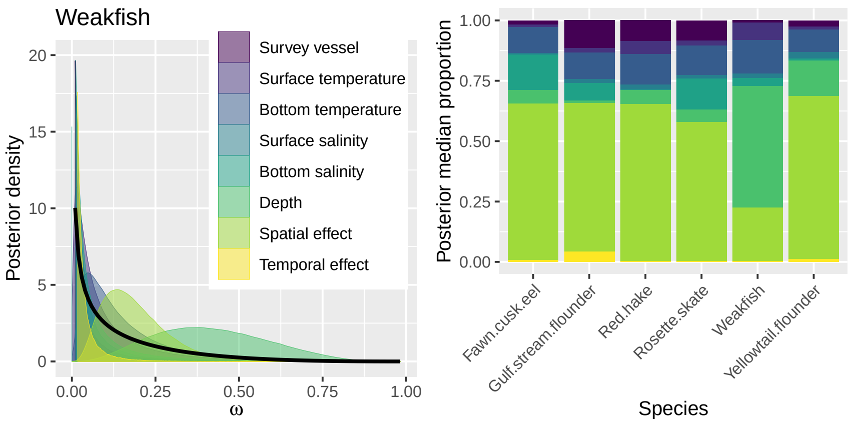


FIGURE 2. Left panel: comparison between the prior distribution on each ω_j (solid black line) and their posterior density for the *Weakfish* species. Right panel: posterior median of each ω_j for six different species.

and give immediate and intuitive posterior outputs about variance partitioning. The class of models of Equation 1 represents just an illustration of a larger theoretical framework developed to correctly apply VP-based priors to a broader class of SDMs, which can include for example smooth effects of abiotic factors, among others. Future challenges include exploring the application of VP-based priors in the context of joint species distribution models.

Acknowledgments: This work was funded by the European Union under the NextGeneration EU Programme within the Plan “PNRR - Missione 4 “Istruzione e Ricerca” - Componente C2 Investimento 1.1 “Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)” by the Italian Ministry of University and Research (MUR), Project title: “METAbarcoding for METAcommunities: towards a genetic approach to community ecology (META2)”, Project code: 2022PA3BS2 (CUP E53D23007580006), MUR D.D. financing decree n. 1015 of 07/07/2023.

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 – 20.

Franco-Villoria, M., Ventrucci, M. and Rue, H. (2021). Variance partitioning in spatio-temporal disease mapping models. *Statistical Methods in Medical Research*, **31**, 1566 — 1578.

Fuglstad, G.A., Hem, I.G., Knight, A., Rue, H. and Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*, **15**, 1109 – 1137.

Hui, F. K., Warton, D. I., Foster, S. D. and Haak, C. R . (2023). Spatiotemporal joint species distribution modelling: A basis function approach. *Methods* *in Ecology and Evolution*, **58**, 2150 – 2164.

Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, **8**, 39 – 51.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J. and Ovaskainen, O. (2020). Joint species distribution modelling with the R-package Hmsc. *Methods in ecology and evolution*, **11**, 442 – 447.

Functional copula graphical regression model for analysing brain–body rhythm

Rita Fici[1], Luigi Augugliaro[1], Ernst C. Wit[2]

[1] University of Palermo, Italy
[2] Università della Svizzera Italiana, Switzerland

E-mail for correspondence: rita.fici@unipa.it

Abstract: In physiology, organ functions can be modelled as networks with individual regulatory mechanisms, forming a broader system through continuous interactions. The system not only interacts with itself, but can also respond to outside impulses. The paper proposes a functional graphical regression model to describe interconnected brain activities partly in response to other organs. The analysis focuses on the conditional independence structure of brain waves given the RR interval of the electrocardiographic waveform, the respiration amplitude and the blood volume pulse.

Keywords: EEG waves; Network physiology; Functional data analysis; Functional graphical regression models, Gaussian copula.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Inference for quasi-reaction models with covariate-dependent rates

Matteo Framba[1], Veronica Vinciotti[1], Ernst C. Wit[2]

[1] Università degli Studi di Trento, Italy
[2] Università della Svizzera Italiana, Switzerland

E-mail for correspondence: matteo.framba@unitn.it

Abstract: Statistical models of quasi-reaction systems are typically described by constant reaction rates. This assumption is too restrictive in many applications, as rates may vary dynamically, spatially or within groups of the population. In this paper, we capture this heterogeneity with the inclusion of covariates in the dynamic model. In particular, we propose an extension of a recently developed latent event history model, by allowing log-reaction rates to be linearly dependent on a vector of covariates. We describe an inferential approach for parameter estimation of the resulting model and evaluate its performance via a simulation study. Finally, we show an illustration on COVID-19 data, where the approach is able to measure the effect of environmental factors and governmental interventions on the disease spreading and severity.

Keywords: Quasi-reaction models; EM algorithm; Kalman filtering; Latent event history models; Epidemic modelling.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

A joint modelling approach for longitudinal patient-reported outcomes and survival analysis

Cristina Galán-Arcicollar[1,2], Danilo Alvares[3], Josu Najera-Zuloaga[2], Dae-Jin Lee[4]

[1] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain, Country
[2] Departamento de Matemáticas, Universidad del País Vasco UPV/EHU, Spain
[3] MRC Biostatistics Unit, University of Cambridge, United Kingdom
[4] School of Science and Technology, IE University, Madrid, Spain

E-mail for correspondence: cgalan@bcamath.org

Abstract: Recently, there has been an increasing interest in using longitudinal biomarkers to characterize the occurrence of an event, such as death. In this context, two outcomes from the same subject are simultaneously observed: repeated measures and time-to-event. The inherent association between them has brought the joint modelling framework. Furthermore, there is a growing priority on placing patients at the centre of healthcare research. In this context, patient-reported outcomes (PROs) are helpful tools for informing clinicians about patients' health status and quality of life. We propose a joint modelling Bayesian approach for longitudinal PRO measurements and survival data that includes adequate distributional fits of PRO by considering its nature and characteristics.

Keywords: Beta-binomial distribution; Chronic obstructive pulmonary disease; Joint models; Patient-reported outcome.

1 Motivation

Patient-reported outcomes (PROs) are helpful tools that provide reports about patient's health status considering their health, quality of life, or functional status associated with the health care or treatment they received. They are now widely utilized for routine monitoring and assessment of care outcomes in adult patients. Its use is strongly recommended, especially in chronic illnesses. PROs come directly from the patient, without any intervention from a clinician, and are frequently collected by supplying validated questionnaires. Thus, PROs are often built as a sum of responses to several questionnaire items, so they can be considered discrete and bounded random variables that are usually overdispersed due to subject-specific characteristics. The Beta-Binomial distribution (BB) is proposed in the literature as an adequate distribution to fit overdispersed discrete and bounded outcomes, particularly in PRO analysis (Arostegui et al. (2007)). The question of whether an association exists between the survival data and longitudinal PROs is widely extended. However, the BB does not belong to the exponential family of distributions, which makes its inclusion into the joint modelling framework not straightforward. Therefore, the inclusion of BB into the joint modelling methodology has yet to be thoroughly investigated. In this work, we propose a Bayesian joint modelling approach for longitudinal PRO measurements and survival data that includes adequate distributional fit of PRO by considering its nature and characteristics.

2 Model definition

Our joint model formulation uses two submodels, one for the longitudinal and another one for the survival that share the same random effects. The complete set of parameters is jointly estimated by means of a full likelihood approach. For each subject i the following data is considered:

- For survival outcome (T_i, δ_i), indicates the time-to-event and status.
- For longitudinal outcome $Y_i = (y_{i1}, \dots, y_{in_i})$, indicates the vector of measurements taken at $(t_{i1}, \dots, t_{in_i})$ times.

The key assumption of this methodology is full conditional independence, i.e., it is assumed that conditional to random effects u_i, time-to-event and longitudinal outcomes are independent, as well as the different measurements for the same subject (Rizopolous (2012)). Then, the joint posterior distribution can be written as:

$$f(\theta, u_i|T_i, \delta_i, Y_i) \propto \left[\prod_j f(y_{ij}|u_i, \theta_y)\right] f(T_i, \delta_i|u_i, \theta_y, \theta_t) f(u_i|\theta_u)\pi(\theta)$$

where $\theta = (\theta_y, \theta_t, \theta_u)$. We propose the use of BB in the longitudinal model when dealing with PROs. To that aim, following a Beta-Binomial mixed effect model (Najera-Zuloaga et al. (2019)), we have $y_{ij}|u_i \sim BB(m, p_{ij}, \phi)$ with bound m, probability parameter p_{ij}, and dispersion parameter ϕ. The BB density is described as:

$$f(y_{ij}|u_i) = \binom{m}{y_{ij}} \frac{\Gamma\left(\dfrac{1}{\phi}\right)}{\Gamma\left(\dfrac{1}{\phi} + p_{ij}\right)} \frac{\Gamma\left(\dfrac{p_{ij}}{\phi} + y_{ij}\right)}{\Gamma\left(\dfrac{p_{ij}}{\phi}\right)} \frac{\Gamma\left(\dfrac{1 - p_{ij}}{\phi} + m - y_{ij}\right)}{\Gamma\left(\dfrac{1 - p_{ij}}{\phi}\right)},$$

where the probability parameter and the subject linear tendency are connected through the logit function. Particularly, for subject i at time t it is denoted as:

$$\text{logit}(p_{it}) = (\beta_0 + u_{i0}) + (\beta_1 + u_{i1})t.$$

According to this longitudinal model setting, $\theta_y = (\beta_0, \beta_1, \phi)$.

The density function for the time-to-event outcome is defined in terms of the hazard function:

$$f(T_i, \delta_i \mid u_i; \theta_t) = \lambda_i(T_i)^{\delta_i} \exp\left\{-\int_0^{T_i} \lambda_i(s)ds\right\},$$

which, assuming the proportional hazards model, is defined as:

$$\lambda_i(t) = \lambda_0(t)\exp(\alpha w_{it}) \quad \text{with} \quad w_{it} = mp_{it},$$

where w_{it} is the true and unobserved value of the longitudinal outcome at time t, according to the previously specified longitudinal model. Thus, the set of parameters for the survival model is $\theta_t = (\alpha, \theta_{\lambda_0})$, where the parameters for the baseline function will vary according to its definition.

Finally, the random parameters distribution is assumed as a multivariate normal density with zero mean and variance-covariance matrix D. Thus, the parameter vector for random effects distribution is $\theta_u = (\sigma_{u_0}, \sigma_{u_1})$.

To overcome the computational complexities, we proceed by using a Bayesian approach. The parameter estimation is performed using the Hamilton Monte Carlo algorithm through `rstan` R-package (Stan Development Team (2024)).

3 Application

We assessed data from a 5-year follow-up study of 543 patients with chronic obstructive pulmonary disease (COPD) from Galdakao-Usansolo Hospital. COPD is one of the major causes of mortality world-wide and its overall impact on the subject is multifaceted, and more than clinical biomarkers are needed to assess its evolution. In this sense, the COPD study considered survival data and one to four Health-Related Quality of Live (HRQoL) measurements per individual collected during the follow-up period. Two questionnaires were used to evaluate the HRQoL: Short-Form 36, and St. George's Respiratory Questionnaire.

We applied our methodology to measure the impact on the questionnaires' scores into the patients' risk of death. Noticing that HRQoL is considered an important outcome itself and a predictor of mortality in COPD patients.

4 Simulation study

A simulation study was carried out to assess the performance of the methodology. The overall scenario settings are mainly based on COPD study, considering the same maximum number of measurements per patient, different entry times, measurement times, follow-up period, and censoring.

We set two main scenarios for the model parameters based on previously developed work by Galán-Arcicollar et al. (2024). The first one considers a positive association parameter, while the second is negative. Different bounds are also considered for the longitudinal model. Table 1 shows these two main parameter setting.

	β_0	β_1	σ_{u_0}	σ_{u_1}	m	α
Scenario 1	-0.19	0.03	1.2	0.05	24	> 0
Scenario 2	0.40	-0.15	1.5	0.3	8	< 0

TABLE 1. Setting the true parameter values for two main scenarios.

To offer a variety of simulation scenes, we varied the strength of the association parameter and also the dispersion parameter of the longitudinal measurements.

These scenario variations allow us to evaluate the performance when there is almost no, moderate and strong association between the outcomes and to contemplate different shapes of the longitudinal distribution. Furthermore, a Weibull baseline hazard was considered for both scenarios:

$$\lambda_0(t) = \nu t^{\nu-1} \exp(\gamma),$$

whose parameters where fixed as $(\nu, \gamma) = (1.6, -2.3)$.
Next figures 1 and 2 show the bias results for the association parameter according to the scenarios set in Table 1 respectively, as well as and well as the sub-scenarios for association and dispersion parameter variations.

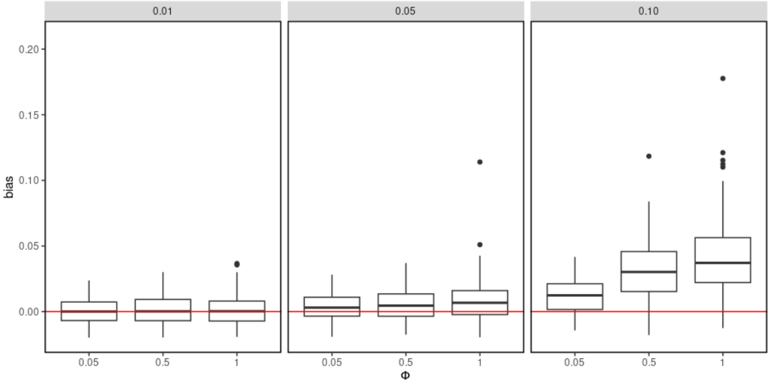


FIGURE 1. Bias boxplot for the α estimation when applying the proposed model according to Scenario 1. Different sub-scenarios are shown according to $\phi \in \{0.05, 0.5, 1\}$ and $\alpha \in \{0.01, 0.05, 0.10\}$.

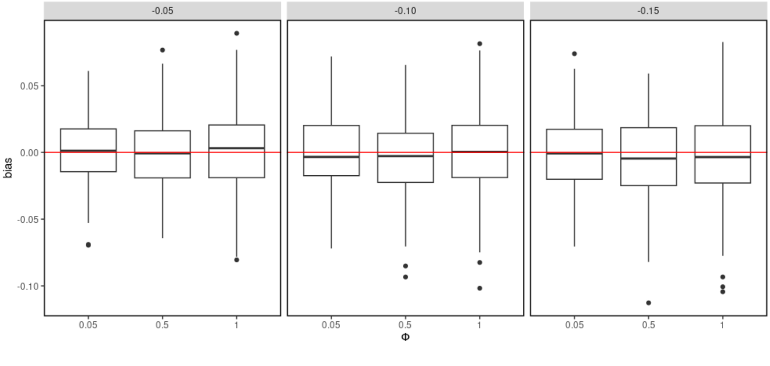


FIGURE 2. Bias boxplot for the α estimation when applying the proposed model according to Scenario 2. Different sub-scenarios are shown according to $\phi \in \{0.05, 0.5, 1\}$ and $\alpha \in \{-0.05, -0.10, -0.15\}$.

It is noticed in Scenario 1 that according to the set bound for the longitudinal outcome, the increment in the association parameter strength leads to biased results. This fact is mainly based on the early occurrence of events and, according to the low number of longitudinal measurements generated. Nevertheless, the other sub-scenarios showed low bias, even leading to unbiased results as shown in figure 2.

5 Concluding remarks

The modelling framework presented provides a suitable way to consider the nature and characteristics of PRO longitudinal data into the joint modelling frame. Furthermore, it presents easy results interpretation in terms of odds and hazard ratio. The validity of the approach was supported by the simulation study and applied to COPD data. In the case study, we got that COPD patients's perception on their health and functional status could lead to an impact in their risk of death, specially for their physical status.

Acknowledgments: This research is supported by AEI and EJ-GV under Grants PID2020-115882RB-I00 / AEI / 10.13039 / 501100011033 with acronym "S3M1P4R", Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco [IT1456-22], by the Basque Government through the BERC 2022-2025 program, by the Ministry of Science and Innovation: BCAM Severo Ochoa accreditation CEX2021-001142-S / MICIN / AEI / 10.13039 / 501100011033, and by the U.K. Medical Research Council grant MC_UU_00002/5.

References

Arostegui, I., Núñez-Antón, V. and Quintana, J. M. (2007). Analysis of the short form-36 (SF-36): The beta-binomial distribution approach. *Statistics in Medicine*, **26**, 1318 – 1342.

Galán-Arcicollar, C., Najera-Zuloaga, J. and Lee, D.-J. (2024). Patient-reported outcomes and survival analysis of chronic obstructive pulmonary disease patients: a two-stage joint modelling approach. *SORT*, **48**, 1 – 28.

Najera-Zuloaga, J., Lee, D.-J. and Arostegui, I. (2019). A beta-binomial mixed-effects model approach for analysing longitudinal discrete and bounded outcomes. *Biometrical Journal*, **61**, 600 – 615.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data With Applications in R.* CRC Press.

Stan Development Team (2024). *RStan: the R interface to Stan.* R package version 2.32.6, https://mc-stan.org/.

A Bayesian Markov-switching for smooth modelling of extreme value distributions

Vincenzo Gioia[1], Gioia Di Credico[1], Francesco Pauli[1]

[1] University of Trieste, Department of Economic, Business, Mathematical and Statistical Sciences "Bruno de Finetti", Trieste, Italy

E-mail for correspondence: vincenzo.gioia@units.it

Abstract: Markov-switching models are attractive for analysing time series that exhibit different stochastic processes along different periods, and where the regime-switching is controlled by an unobservable Markovian process. Model flexibility can be enhanced considering regime-specific distributions, whose distributional parameters may be modelled using smooth functions of covariates. Here, we propose a two-state Markov-switching model using full Bayesian inference and accounting for extreme value modelling. The proposal is illustrated by analysing energy prices.

Keywords: Distributional regression, Energy price modelling, Regime-switching.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Additive mixed models for location, scale and shape via gradient boosting techniques

Colin Griesbach[1], Elisabeth Bergherr[1]

[1] Georg-August-Universität Göttingen, Germany

E-mail for correspondence: colin.griesbach@uni-goettingen.de

Abstract: In this work we adapt recent findings from statistical boosting in order to construct an estimation approach for distributional regression including random effects. The algorithm is applied to registry data provided by the German Cystic Fibrosis Registry where the subject-specific evolution of each patients lung function and its corresponding distributional parameters are modelled.

Keywords: Statistical boosting; Distributional regression; Random effects.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

Statistical modeling of UEFA EURO soccer matches with focus on player market value and other hybrid variables

Jacob Grytzka[1], Andreas Groll[1]

[1] TU Dortmund University, Germany

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

Abstract: We analyze the UEFA EUROs 2004 to 2020 to find the best feature set for predicting future UEFA EUROs. We use 21 features covering the economic strength of the country and the sporting performance of each national team. In addition, three so-called hybrid features are included, which are based on separate statistical models. These are based on historical match data, bookmakers' odds and a special player ranking. We create 31 different feature sets and compare them on seven different performance measures using two cross validation approaches. We found the feature set consisting only of the bookmaker odds and the special player ranking to perform best. For this feature set, we finally took a closer look at the effect of the features.

Keywords: Random forest; Cross validation; Feature set; UEFA EURO; Soccer.

1 Introduction

When a big soccer tournament like the UEFA EURO takes place, many experts try to predict the winning team. Statistical learning models are a good way to do this. At UEFA EURO 2020, Groll et al. (2021) compared the performance of different statistical models and found the Conditional Random Forest (`cforest`; Hothorn et al., 2006) to be a good choice. To better predict the winner of the next UEFA EURO, the cforest is used in this study to find the best feature set. We consider economic and sporting factors as well as three hybrid characteristics based on separate statistical models.

2 Dataset and statistical methods

We have a total of 21 features at our disposal. We look at economic factors such as gross domestic product (GDP) and the population of a country. We also

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

regard features relating to sporting success, such as the FIFA world ranking, points in the UEFA five-year ranking and the resulting number of starting places in international competitions. We also examine the team structure based on the average age of the squad, the number of players playing abroad and the number of Champions League and Europa League players in the squad. We also look at the groups of players within a squad who also play together at the club. Two further features relate to the coach. We use the age of the coach and a feature that indicates whether the coach is from the respective country. In addition, the home advantage and the tournament phase of each match are taken into account. Special focus is on the *market value* of the teams and the so-called hybrid variables, which were extracted from separate statistical models and are presented below. A detailed description of these features can be found in Groll et al. (2021).

Abilities

According to a comparison by Ley et al. (2019), a bivariate Poisson model is well suited for estimating team strength parameters based on historical match data. Let $n \in \mathbb{N}$ be the number of teams and $M \in \mathbb{N}$ the number of matches. Let Y_{ijm} be the random variable representing the number of goals scored by team i against team j $(i, j \in \{1, \dots, n\}, i \neq j)$ in match $m \in 1, \dots, M$. We assume that Y_{ijm} and Y_{jim} follow a bivariate Poisson-distribution (see Karlis and Ntzoufras, 2003). The expected goals λ_{ijm} of a team are modeled using

$$\log(\lambda_{ijm}) = \begin{cases} \beta_0 + (r_i - r_j) + h, & \text{if team } i \text{ plays at home in match } m, \\ \beta_0 + (r_i - r_j), & \text{otherwise.} \end{cases}$$

Here, β_0 is the intercept, h the effect of the home advantage and r_i the team-specific strength parameters, which finally form the feature *Abilities*. For this study, matches from the eight years preceding each UEFA EURO are utilized to estimate the strength parameters. The $n + 3$ parameters of this model are estimated using the maximum likelihood method. To ensure that the estimation of the parameters is identified, the likelihood function is optimized under the constraint $\sum_{i=1}^{n} r_i = 0$. Individual matches are weighted differently depending on how long ago they took place. For this purpose, we define a half-period $H \in \mathbb{N}$, which specifies after how many days a match is weighted only half as much. The weight w_m of match m, which took place t_m days ago, is then given by

$$w_m = \left(\frac{1}{2}\right)^{\frac{t_m}{H}}.$$

As in Ley et al. (2019), the half-period is set to three years, i.e. $H = 1095$ days. There, a weighting is also made based on the importance of the match, which is not used here, as in Groll et al. (2021).

Logability

One possibility to predict a UEFA EURO is based on the betting odds of various bookmakers in relation to winning the respective tournament. Leitner et al. (2010) used these to carry out an inverse tournament simulation. Let n be the number of teams and B the number of bookmakers. The $quoted\_odds_{ib}$ indicate how large the profit per staked Euro is for bookmaker $b \in \{1, \dots, B\}$ in the case of a successful bet on team $i \in \{1, \dots, n\}$. Similarly, the "true" odds are denoted by $odds_{ib}$. The parameter $\delta_b \in [0,1]$ indicates the proportion of bets that are actually paid out by bookmaker b. Thus, $1 - \delta_b$ corresponds to the profit margin of bookmaker b. The following relationship is assumed:

$$quoted\_odds_{ib} = odds_{ib} \cdot \delta_b + 1.$$

The +1 is necessary because the invested stake is also paid out again if the bet is successful. In Leitner et al. (2010) it is assumed that δ_b is the same for all teams in a tournament. According to the equation, all odds are thus determined and then logarithmized to obtain the *log-odds*. These are averaged over all bookmakers and then converted into winning probabilities using the logit transformation. Finally, these probabilities finally form the bookmakers' consensus model.

The tournament to which the odds apply is then simulated several times. To determine the winner in a single match, the probability of victory is calculated as follows:

$$p_{ij} = \mathrm{P}(\text{Team } i \text{ defeats Team } j) = \frac{s_i}{s_i + s_j} \quad i \neq j,$$

where s_i and s_j $(i, j \in \{1, \ldots, n\}; i \neq j)$ are strength parameters for team i and team j. Since there is no probability of a draw in this model and also no information about goals scored and thus goal difference, it may be necessary to simulate additional matches in the group phase of a tournament in order to obtain a clear ranking of the teams in a group.

The simulation of the tournament is carried out 100,000 times in Leitner et al. (2010). In this way, a team's probability of winning can be estimated on the basis of its share of tournament victories. The strength parameters $s_1, \ldots, s_n$ are chosen in such a way that the winning probabilities approximate those from the bookmakers' consensus model. These strength parameters finally form the feature *Logability*.

Plus-Minus player rating

The third hybrid variable is based on the Plus-Minus player rating (PM) by Hvattum (2019). In this approach, the strength of a team is not determined directly, but with the help of the strength of its players in the squad. To estimate the strength of a particular player $j \in \{1, \ldots, S\}$, all the matches for both the national team and his national club, in which the player has played actively, are considered. The matches are divided into time periods, in each of which the same players are on the pitch. The first period starts with the kick-off of the match. When a player is substituted or sent off (e.g. by receiving a red or double-yellow card), the current period ends and a new one begins. The last period ends with the end of the match. For each period, the number of goals scored by the home and away team is also recorded.

The basis of the plus-minus player score is a simple linear model for each time period $i \in \{1, \ldots, n\}$ of the following form:

$$y_i = \sum_{j=1}^{S} \beta_j x_{ij} + \varepsilon_i.$$

The response variable of the linear models y_i represents the goal difference from the home team's point of view with respect to all goals scored in time period i.

ε_i is the error term and

$$x_{ij} = \begin{cases} 1, & \text{if player } j \text{ plays for the home team in time period } i, \\ -1, & \text{if player } j \text{ plays for the away team in time period } i, \\ 0, & \text{otherwise,} \end{cases}$$

is an indicator for the presence of a player. The fitted $\hat{\beta}_j$ represent the rating of the players. Two features are extracted from the PM of each squad. First, *MeanPM* denotes the average rating of all players in a squad. The second determined feature *XMissing* indicates how many players were not in the squad of their country for the respective tournament, although they played at least once for the national team and, according to the PM, belong to the best 11 players of the national team.

3 Study

We have divided the features into five groups. The *market value* feature and each of the three models for the hybrid features form a group of features. The fifth group contains the remaining basic features. We then built feature sets from the 31 combinations of these groups. We use the conditional random forest to model the expected goals of a team. The expected goals of both teams can be used to predict the outcome of the match. The 31 feature sets are compared via both a classical 10-fold and a tournament-specific cross validation (CV) using seven different performance measures. These were the mean squared error (MSE) and mean absolute error (MAE), each in terms of goals scored by a team and goal difference in a match. In addition, three measures were calculated, which refer to the specific match outcome (i.e. win, draw and defeat). The corresponding predictive Multinomial Likelihood (PL) indicates which probability the prediction model assigned to the actual outcome. The Classification Rate (CR) is the proportion of correctly predicted results and the Rank Probability Score (RPS) is an error measure that takes the ordinal structure of the results into account. In selecting the best feature set, the focus was primarily on the latter three performance measures. Table 1 exemplarily shows the best results of both CVs. The feature set containing only *market value* achieved the best CR in both CVs. Taking the *market value* and the PM results in the best RPS in the leave one tournament out CV. The best feature set with respect to PL and RPS consists of the two groups *Logability* and *PM.*

We want to investigate the effect of the winning feature set in more detail. For a fitted random forest this is not trivial and we need to employ methods from the field of interpretable machine learning (see Molnar, 2023). Figure 1 shows a Partial Dependence plot (PDP) for the feature *Logability* on the left side. The graph exhibits a trend where higher values of the feature basically lead to more goals, which seems plausible. However, the curve is only roughly monotonic. At some points the effect is briefly and slightly decreasing, before the curve makes a jump and continues the trend. The influence of *MeanPM* is not as clear as the one for *Logability.* We can see an increasing trend, but local deviations are quite strong. E.g., after an initial rise, the curve actually falls below the initial level. The effect of *Logability* seems to be slightly more meaningful overall than the one of *MeanPM.* While the left graph ranges from 0.75 to 1.75, and hence covers

TABLE 1. Best settings for all PLs, CRs and RPSs with 10-fold CV (above) and leave one tournament out CV (below)

Basic	MW	abilities	Logability	PM	PL	CR	RPS
X	X		X	X	0.3997	0.5171	0.1990
X			X	X	0.3959	0.5171	0.2007
	X		X	X	0.4067	0.4971	0.1983
	X				0.3808	**0.5224**	0.2052
		X	X	X	0.4049	0.5021	0.1999
			X	X	**0.4075**	0.5018	**0.1973**
	X		X	X	0.4070	0.5004	0.2004
	X		X		0.3919	0.5108	0.2039
	X			X	0.4059	0.5044	**0.1996**
	X				0.3826	**0.5276**	0.2042
			X	X	**0.4084**	0.5018	0.2005

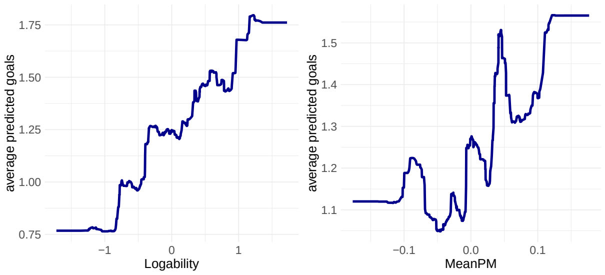


FIGURE 1. Partial dependence plots showing the influence of two features on the goals scored

a larger range of the response, the effect for *MeanPM* on the right is roughly between 1.05 and 1.55.

References

Groll, A., Hvattum, L. M., Ley, C., Popp, F., Schauberger, G., Van Eetvelde, H. and Zeileis, A. (2021). Hybrid Machine learning forecasts for the UEFA EURO 2020. *arXiv preprint arXiv:2106.05799*

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, **15**, 651 – 674

Hvattum, L. M. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport*, **18**, 1 – 23.

Karlis, D. and Ntzoufras I. (2003). Analysis of sports data by using bivariate Poisson models. *International Journal of Computer Science in Sport, Series D*, **52**, 381 – 393.

Leitner, C., Zeileis A. and Hornik K. (2010). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, **26**, 471 – 481.

Ley, C., Van de Wiele T. and Van Eetvelde H. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, **19**, 55 – 73.

Molnar, C. (2023). Interpretable Machine Learning.
url: `https://christophm.github.io/interpretable-ml-book/`.

Learning Bayesian networks from ordinal data - The Bayesian way

Marco Grzegorczyk[1]

[1] Bernoulli Institute, FSE, Groningen University, NL

E-mail for correspondence: m.a.grzegorczyk@rug.nl

Abstract: We propose a new Bayesian method for Bayesian network structure learning from ordinal data. Our Bayesian method is similar to a recently proposed non-Bayesian method, referred to as the ordinal structural expectation maximization (OSEM) method. Both methods assume that the ordinal variables originate from Gaussian variables, which can only be observed in discretized form, and that the dependencies in the unobserved latent Gaussian space can be described in terms of Gaussian Bayesian networks. In our simulation studies the new Bayesian method yields significantly higher network reconstruction accuracies than the OSEM method.

Keywords: Bayesian networks; Ordinal data; Latent Gaussian space; Markov chain Monte Carlo (MCMC).

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Statistical models for patient-centered outcomes in clinical studies

Gillian Heller[1], Andrew Forbes[2], Stephane Heritier[2]

[1] NHMRC Clinical Trials Centre, University of Sydney, Australia,,
[2] School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

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

Abstract: Days alive and out of hospital is recommended as a patient-centered outcome in perioperative clinical studies. It is defined as the number of days, out of the first M postoperative days, that the patient has been discharged from hospital, or zero if the patient dies within M days of surgery. This composite measure presents statistical challenges in its unusual distributional shape, and its inability to distinguish between the qualitatively different outcomes of death, and a hospital stay longer than M days. We propose a mixed binary-continuous model that overcomes these difficulties, and illustrate its use on a clinical trial of a drug administered in cardiac surgery.

Keywords: Distributional regression; Censoring; Hospital length of stay; GAMLSS; Patient-centered outcomes.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Modelling of overdispersed count rates

John Hinde[1], Alberto Alvarez-Iglesias[2], John Ferguson[1], Clarice G.B. Demétrio[3] , John Crown[4,5], Bryan T. Hennessy[6], Vicky Donachie[7]

[1] School of Mathematical and Statistical Sciences, University of Galway, Galway, Ireland,,
[2] HRB Clinical Research Facility Galway, University of Galway, Galway, Ireland
[3] ESALQ/USP, Piracicaba, Brazil
[4] Department of Medical Oncology, St Vincent's University Hospital, Dublin, Ireland
[5] Cancer Clinical Research Trust, National Institute for Cellular Biotechnology, Dublin City University, Dublin, Ireland
[6] Department of Medical Oncology/ICORG/Cancer Trials Ireland, Beaumont Hospital, Royal College of Surgeons in Ireland, Dublin, Ireland
[7] Cancer Trials Ireland/ICORG, Dublin, Ireland

E-mail for correspondence: `john.hinde@universityofgalway.ie`

Abstract: This paper revisits the common problem of analysing counts recorded over time through the modelling of the underlying rate, motivated by the analysis of a cancer treatment related study. The baseline Poisson model is simply implemented though the inclusion of an offset for the different exposure/recording times and the underlying Poisson process gives other nice well-known properties. We consider how this approach can be extended to models for overdispersed data. The use of a simple offset with a negative binomial model is common practice and we consider the appropriateness of this and the resulting implications. We discuss how these ideas extend to more general mixed Poisson models, including ZIP, to handle zero-inflation, and ZINB for zero-inflation and overdispersion. The simple offset approach does not extend to other extended count models such as the COM-Poisson and general weighted Poisson distributions.

Keywords: Counts; Rates; Overdispersion models.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

Induced robust inference for the concordance correlation coefficient

Vanda Inácio[1], Richard A. Parker[2]

[1] School of Mathematics, University of Edinburgh, UK
[2] Edinburgh Clinical Trials Unit, Usher Institute, University of Edinburgh,, UK

E-mail for correspondence: `vanda.inacio@ec.ac.uk`

Abstract: The concordance correlation coefficient is a widely used standardized index in agreement studies. We develop robust inference for the concordance correlation coefficient for the case of repeated or clustered measurements data, thus minimising the impact of outlying observations that otherwise may lead to erroneous conclusions about the agreement between devices. Our methods are illustrated in a study involving chronic obstructive pulmonary disease patients and matched repeated respiratory rate observations.

Keywords: Agreement studies; Concordance correlation coefficient; Linear mixed effects model; Repeated measurements; Robust statistics.

1 Introduction

Measurements are the basis for evaluation in almost all scientific disciplines, among which the medical sciences are possibly the most prominent. Agreement studies quantify the closeness of the measurements of the same variable made by two different devices and are typically motivated when newer, less invasive, and/or cheap devices become available and their agreement with the gold standard device needs to be evaluated. If the measurements generated by each device are close together most of the time, we conclude that the devices agree and they can be used interchangeably.

Several indices for assessing the agreement of continuous data have been proposed in the literature, of which the concordance correlation coefficient (CCC) is one of the most popular and widely used. When repeated measurements are available, the CCC can be calculated via a linear mixed effects model. As we shall see in the next section, the expression for the CCC is a function of the variance components of the underlying mixed effects model. The usual assumption behind a linear mixed effects model is that all variance components follow a normal distribution. When there are outlying observations, either an outlying measurement among the repeated measurements within an individual or/and an

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

outlying invidual in the sample of subjects, the use of the normal linear mixed effects model may lead to misleading conclusions about the agreement between devices. Indeed, recently, Keu et al. (2021) have provided both theoretical and empirical evidence that estimates of the variance components can be strongly biased when the distribution of the random effects is misspecified.

To close this gap in the agreement literature, we propose to estimate the linear mixed effects model that underlies the computation of the CCC based on robust methods that mitigate the impact of outlier measurements. We are motivated by a study involving chronic obstructive pulmonary disease (COPD) patients, where respiratory rate measurements (in breaths per minute) from 21 subjects with COPD were measured simultaneously by a new and a gold standard device, both worn at the same time. Multiple time-matched respiratory rate measurements were taken on each patient. Specifically, eleven different activities, ranging from slow walking to climbing stairs, and which were chosen to be representative of the activities encountered in daily life, were performed by participants. Not everyone performed exactly the same number of activities because some tasks were too difficult for some participants, with most activities having just one respiratory rate reading per subject.

2 Induced robust estimation for the concordance correlation coefficient

In the context of our COPD example, we assume the following linear mixed effects model

$$y_{ijlt} = \mu + \alpha_i + \beta_j + \gamma_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{il} + (\beta\gamma)_{jl} + \epsilon_{ijlt}, \qquad (1)$$

where y_{ijlt} represents the respiratory rate measurement made on subject i by device j when performing activity l at time t, μ is the overall mean, $\alpha_i \sim \mathrm{N}(0, \sigma^2_\alpha)$ is the random subject effect, β_j is the fixed effect of device j, and $\gamma \sim \mathrm{N}(0, \sigma^2_\gamma)$ denotes the random activity effect. Further, $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{il}$, and $(\beta\gamma)_{jl}$ denote, respectively, the random interaction between subject and device, between subject and activity, and between device and activity and they all follow a normal distribution with mean zero and with variance $\sigma^2_{\alpha\beta}$, $\sigma^2_{\alpha\gamma}$, $\sigma^2_{\beta\gamma}$, respectively. Finally, $\epsilon_{ijlt} \sim \mathrm{N}(0, \sigma^2_\epsilon)$ is the error. All random effects are assumed to be independent.

We justify these modelling choices as follows. We regard subjects as random effects, therefore implicitly assuming they are a sample from a wider population of COPD patients. We regard activity as a random effect as well, mainly so that we can generalize the results to any activity from a wider ‘population’ of activities performed by participants in daily life, but also so that activities with small numbers of respiratory rate readings are not weighted too highly in the model . All possible two-way interactions were included in the model and they take into account the variability in subjects across devices, in subjects across activities, and in devices across activities.

Based on (1), the CCC can be written as

$$\rho_{CCC} = \frac{\mathrm{cov}(y_{i1lt}, y_{i2lt})}{\mathrm{var}(y_{ijlt})} = \frac{\sigma^2_\alpha + \sigma^2_\gamma + \sigma^2_{\alpha\gamma}}{\sigma^2_\alpha + \phi^2_\beta + \sigma^2_\gamma + \sigma^2_{\alpha\gamma} + \sigma^2_{\alpha\beta} + \sigma^2_{\beta\gamma} + \sigma^2_\epsilon},$$

where $\phi^2_\beta = \sum_{j=1}^{2} \beta^2_j$ (to ensure identifiability we assume $\beta_1 + \beta_2 = 0$) which accounts for the systematic differences between the two devices. The CCC can

take values between -1 and 1, where 1 indicates perfect agreement and -1 indicates perfect disagreement. The CCC, in this particular case, thus reflects the proportion of the total overall variability explained by the subject and activity effects (and their interaction) and a CCC of 1 implies that there is no variability in the device across subjects and activities.
In order to mitigate the adverse effect of outlying measurements, either within each participant or across participants, we follow the approach of Koller (2016), implemented in the `robustlmm` `R` package, which robustifies the scoring equations arising from Equation (1) by replacing the residuals and predicted random effects with bounded functions, namely with Huber type of functions. Replacing terms by bounded functions thereof downweights terms with a large absolute value. In the robustness literature, these weights are called robustness weights. Observations or random effects with low robustness weights are classified as outliers by the robust method. Inference for the CCC is performed through a bootstrap scheme by resampling at the patient level.

3 Application to COPD data: results

We begin by employing the non-robust linear mixed effects model and the CCC was estimated to be 0.69, with a 95% bootstrap CI of $(0.60, 0.74)$, based on 500 resamples. The normal quantile quantile plots for all variance components and error term are presented in Figure 1. We can observe that for the random interaction between subject and activity, some observations fall outside the 95% simulation envelope. Additionally, the random error exhibits pronounced heavy-tail behavior. This evidence calls into question the validity of the non-robust linear mixed effects model and consequently, the computed value for the CCC. Indeed, according to the robust linear mixed effects model, two observations for the subject actvity interaction have a robust weight of less than 0.2, with the number being 20 for the error term, thus indicating that these measurements are potential outliers. The CCC estimate based on the robust approach is 0.77 $(0.71, 0.82)$. Note that the non-robust CCC point estimate is not even included in the robust 95% CI. Further, there is little overlap between the two intervals. Now onto the conclusions based on the robust approach. A CCC of $0.77(0.71, 0.82)$ shows moderate to good agreement between the new and the gold standard device. The variance components estimates for the subject and activity effects are the highest according to the robust model and therefore are the main source of diseagreement. In turn, the random interaction between subject and device is estimated to be zero, indicating no evidence of a difference in the device effect across subjects.

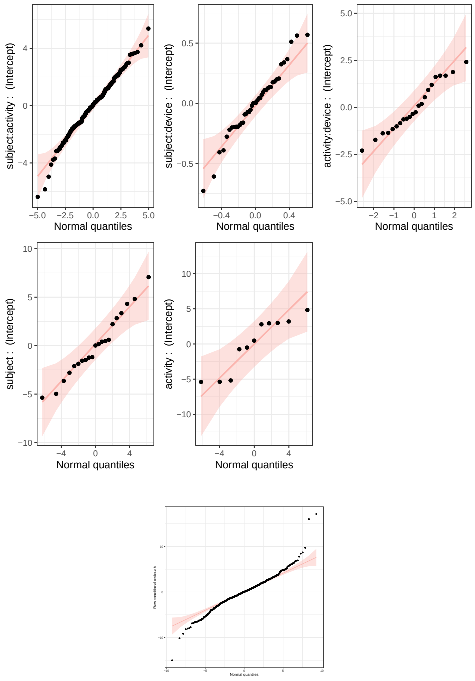


FIGURE 1. Top (rows 1 and 2): normal QQ plots for the random effects. Bottom (row 3): normal QQ plot of the conditional raw residuals.

References

Hui, F. K. C., Muller, S. and Welsh, A. H (2021). Random effects misspecification can Hhave severe consequences for random effects inference in linear mixed models *International Statistical Review*, **89**, 186 – 206.

Koller, M. (2016). `robustlmm`: An `R` package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, **75**, 1 – 24.

Optimism correction of the AUC with complex survey data

Amaia Iparragirre[1], Irantzu Barrio[1,2]

[1] Univesity of the Basque Country UPV/EHU, Spain,
[2] BCAM - Basque Center for Applied Mathematics, Spain

E-mail for correspondence: `irantzu.barrio@ehu.eus`

Abstract: Special statistical techniques are required to develop valid prediction models for complex survey data. Recently, a weighted estimator has been proposed to estimate the area under the receiver operating characteristic curve in this context. However, the proposed estimator has shown an optimistic behaviour. Thus, the goal of this work is to analyze the performance of replicate weights methods to correct for the optimism of the AUC in the context of complex survey data.

Keywords: Area under the receiver operating characteristic curve (AUC); Complex survey data; Optimism correction; Replicate weights.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

A distance-based statistic for goodness-of-fit assessment

Darshana Jayakumari[1], Jochen Einbeck[2], John Hinde[3], Rafael A. Moral[1]

[1] Maynooth University, Ireland
[2] Durham University, United Kingdom
[3] University of Galway, Ireland

E-mail for correspondence: `darshana.jayakumari.2021@mumail.ie`

Abstract: The modelling of count data in real world scenarios often requires models that address overdispersion. Within the generalized linear modeling framework, various overdispersion models are available as extensions of the basic Poisson model. Graphical assessment of goodness-of-fit commonly involves half-normal plots, for which a simulated envelope can be added to aid interpretation. The simulated envelope is such that, under a well-fitted model, the majority of points should fall within its bounds. Nonetheless, closely related models tend to produce very similar graphs. Here, we propose an objective statistic based on half-normal plots with a simulated envelope to aid goodness-of-fit assessment.

Keywords: Count data; Half-normal plot; Model selection.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Misleading effects in relational event models

Rūta Juozaitienė[1], Ernst C. Wit[2]

[1] Vytautas Magnus University, Lithuania
[2] Università della Svizzera italiana, Switzerland

E-mail for correspondence: `ruta.juozaitiene@vdu.lt`

Abstract: Relational event models are event history models for dynamic network interactions. Computational considerations have resulted in frequently used simplified network statistics. This study explores the impact of diverse endogenous effect definitions. Simulations and real-world studies focused on reciprocity and transitivity effects emphasize the need for more complex effect definitions to avoid possible contradictory interpretation of the results. We introduce a flexible computational framework to infer such effects efficiently.

Keywords: Relational events; Reciprocity; Triadic closure; Nested case control sampling.

1 Introduction

Endogenous network effects, such as reciprocity and triadic closure, encompass social and temporal information within relational events, offering comprehensive measures to describe social interactions. The scientific literature offers a variety of exogenous effect definitions, ranging from simple binary indicators to intricate formulations considering event timing. This research delves into a subset of qualitatively distinct effects, aiming to comprehend how these diverse definitions may influence modelling outcomes.

2 The relational event model

Relational event models (REMs) offer a flexible framework for studying time-ordered sequences of relational events (Butts, 2008). A relational event is defined as a discrete event initiated at time t by a social actor, specifically a sender s, directed towards a receiver r, represented as (s, r, t). The counting process $N_{sr}(t)$ of event (s, r) can be modelled by the conditional intensity function,

$$\lambda_{sr}(t) = \lambda_0(t) \exp\left[\boldsymbol{\beta}^{\mathrm{T}} x_{sr}(t) + f^R_{sr}(t) + f^T_{sr}(t) + \boldsymbol{b}^{\mathrm{T}} z_{sr}\right],$$

where $\lambda_0(t)$ represents the baseline hazard function, $x_{sr}(t)$ is the set of endogenous and exogenous variables, $\boldsymbol{\beta}$ denotes effect sizes, $f^R_{sr}(t)$ and $f^T_{sr}(t)$ represents reciprocity and triadic closure effects, and $\boldsymbol{b} \sim \mathcal{N}(\mathbf{0}, \Sigma(\boldsymbol{\phi}))$ is a vector of random effects.

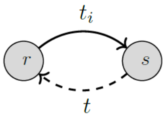
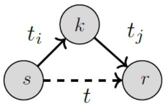
2.1 Definition of exogenous effects

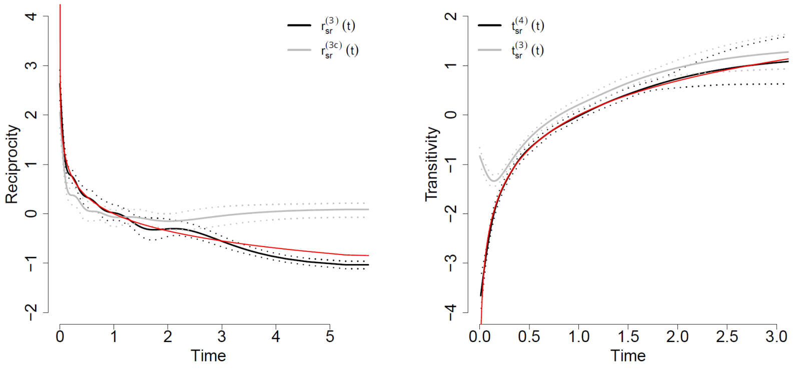
Two traditional definitions for reciprocity (see Table 1) involve a binary variable $r^{(1)}_{sr}(t)$, indicating whether sender s received a relational event from actor r up to time t, or an exponential decay function with a *half-life* parameter T $r^{(2)}_{sr}(t)$. These definitions were proposed mainly for computational convenience. Instead, we suggest modelling reciprocity as a smooth function of time $r^{(3)}_{sr}(t)$, where the time to reciprocity (Δt_{sr}) is defined as the difference between t and the most recent event $r \to s$. In cases where $r \to s$ was not observed at all, $\Delta t_{sr} = \infty$. It could be the reciprocity effect vanishes after the occurrence of the event $s \to r$. For example, in the context of repaying a debt, the debt disappears after being settled. However, reciprocity might also exhibit a continuous nature, where the effect persists once activated, as defined in $r^{(3c)}_{sr}(t)$.
Analogous definitions are introduced for triad closure effect, including an additional definition $t^{(4)}_{sr}(t)$, which enforces a strict order in the creation of a two-path. In contrast, definition $t^{(3)}_{sr}(t)$ posits that a two-path can be formed by placing the link $k \to r$ before $s \to k$. Although this paper primarily focuses on transitivity, similar definitions can be derived for other triadic effects, such as cyclic closure and sending/receiving balance.

2.2 Case-control partial likelihood inference via GAM

The vector of model parameters β in the conditional intensity function can be estimated using the partial likelihood approach (Cox, 1972). Unfortunately, for

TABLE 1. Definitions of the exogenous network effects.

Effect	Structure	Definitions
Reciprocity		$r^{(1)}_{sr}(t) = \begin{cases} 1 & \text{if } t_i < t \\ 0 & \text{otherwise} \end{cases}$ $r^{(2)}_{sr}(t) = \sum_{i:s_i=r,r_i=s,t_i<t} \mathrm{e}^{-(t-t_i)\frac{\ln 2}{T}} \frac{\ln 2}{T}$ $r^{(3)}_{sr}(t) = f^R(\Delta t_{sr}), \Delta t_{sr} = t - t_i$
Transitive closure		$t^{(1)}_{sr}(t) = \begin{cases} 1 & \text{if } t_i < t \text{ and } t_j < t \\ 0 & \text{otherwise} \end{cases}$ $t^{(2)}_{sr}(t) = \sum_{\substack{k: \\ (s,k,t_i),t_i<t \\ (k,r,t_j),t_j<t}} \mathrm{e}^{-\left(t-\max(t_i,t_j)\right)\frac{\ln 2}{T}} \frac{\ln 2}{T}$ $t^{(3)}_{sr}(t) = f^T(\Delta t_{sr}), \Delta t_{sr} = t - \max(t_i, t_j)$ $t^{(4)}_{sr}(t) = f^T(\Delta t_{sr}), \Delta t_{sr} = t - t_j$



(a) Reciprocity (b) Transitivity

FIGURE 1. Average estimates of the time-varying network effects according to different definitions along with their confidence bands. It shows that the incorrect definition (grey) can lead to misleading conclusions.

large networks its computation grows quadraticly with the number of nodes. To address this computational complexity we employ nested case–control sampling (Vu, D. et al., 2015). This approach reduces the risk set to the observed event and one randomly sampled non-event (s_i^*, r_i^*, t_i). The resulting sampled partial likelihood corresponds to the likelihood of a generalized linear mixed model without an intercept for binary outcomes,

$$PL_{NCC}(\boldsymbol{\beta}, \boldsymbol{b}) = \prod_{i=1}^{n} \frac{\lambda_{s_i r_i}(t_i)}{\lambda_{s_i r_i}(t_i) + \lambda_{s_i^* r_i^*}(t_i)} =$$

$$\prod_{i=1}^{n} \frac{\mathrm{e}^{\boldsymbol{\beta}^T \left[x_{s_i r_i}(t_i) - x_{s_i^* r_i^*}(t_i)\right] + \sum_{k \in \{R,T\}} \left[f^k_{s_i r_i}(t) - f^k_{s_i^* r_i^*}(t)\right] + \boldsymbol{b}^T \left[z_{s_i r_i} - z_{s_i^* r_i^*}\right]}}{1 + \mathrm{e}^{\boldsymbol{\beta}^T \left[x_{s_i r_i}(t_i) - x_{s_i^* r_i^*}(t_i)\right] + \sum_{k \in \{R,T\}} \left[f^k_{s_i r_i}(t) - f^k_{s_i^* r_i^*}(t)\right] + \boldsymbol{b}^T \left[z_{s_i r_i} - z_{s_i^* r_i^*}\right]}}.$$

To estimate parameters, we employ a generalized additive model using the *mgcv* package in R (Wood, 2017).

3 Misleading effect estimates in REMs

To show the importance of proper effect definition, we perform a simulation study, simulating 10000 events among 20 nodes, and using REMs with various reciprocal effects reveals conflicting patterns (Table 2). In parallel experiments we explore the transitive effect, assuming the order of events is crucial. Results from 20 simulations show conflicting conclusions based on different definitions (Table 2). Moreover, Figure 1b highlights the crucial role of subtle details, such as the ordering of two-path formation, in the estimation process.

Simulations reveal how different definitions of the same network effect can lead to different conclusions, emphasising the need for a thorough consideration of the appropriate definition in the efficient estimation of network effects.

TABLE 2. Average values of the fixed effects estimates indicate that different models suggest contradictory results.

	Estimate(SE)	AIC	Interpretation
$r_{sr}^{(1)}(t)$	0.08(0.03)	13856	Reciprocal events have a higher rate of occurrence
$r_{sr}^{(3)}(t)$	-278.49(14.83)	13430	r sending events to s reduces the likelihood of s responding with a subsequent event
$t_{sr}^{(1)}(t)$	-0.32(0.05)	13812	Transitive events have a lower rate of occurrence
$t_{sr}^{(2)}(t)$	57.94(2.17)	13080	As past events from s to other nodes contacting r increase, the rate of events from s to r rises.

References

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, **38**, 155 – 200.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187 – 202.

Vu, D., Pattison, P. and Robins, G. (2015). Relational event models for social learning in MOOCs. *Social Networks* , **43**, 121 – 135.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R.* CRC press.

Likelihood ratio tests in penalized logistic regression with categorical covariates

Lea Kaufmann[1], Maria Kateri[1]

[1] Institute of Statistics, RWTH Aachen University, Germany

E-mail for correspondence: kaufmann@isw.rwth-aachen.de

Abstract: In a broad field of applications, high-dimensional problems occur, i.e. problems where the number of parameters in a regression-type model is too high compared to the sample size or can even exceed it. Especially in presence of categorical explanatory variables (i.e. factors), such problems can occur easily even if the number of candidate factors is moderate. While penalized regression approaches enable a simultaneous variable selection and regression coefficients' estimation, the implementation of further statistical inference procedures, e.g., likelihood ratio tests (LRT), is not straightforward, due to the high-dimensionality of the problem. For this, we propose a two-stage penalized logistic regression approach for a penalty function enforcing both factor selection and levels fusion simultaneously. In particular, we extend the (multiple) sample splitting approach, which is introduced for penalization methods performing only *variable selection*, to a method performing *factor selection* as well as *levels fusion*. We specify and adjust the regularity conditions for penalization methods of this type, considering two different approaches for multiplicity adjustments, i.e. the Benjamini-Hochberg procedure and Bonferroni correction. We further investigate asymptotic properties, such as type-I-error control, concluding that the proposed two-stage approach is adequate for applications.

Keywords: High dimensional statistics; Likelihood ratio test; Sample splitting; Logistic regression; Penalized regression.

1 Penalized logistic regression

Consider a logistic regression problem with a binary response variable Y and $J \in \mathbb{N}$ categorical explanatory variables (i.e. factors) $\mathcal{X}_1, \dots, \mathcal{X}_J$, each having $p_j + 1$ levels, where $p_j \in \mathbb{N}$, $j \in \{1, \dots, J\}$. We code these levels as $0, \dots, p_j$ and choose zero as the reference category. For coding of the factors, we follow the commonly used dummy-coding scheme introducing p_j dummy variables $\mathcal{X}_{j,1}, \dots, \mathcal{X}_{j,p_j}$ being zero or one for each factor $\mathcal{X}_j$, where $\mathcal{X}_{j,k} = 1 \Leftrightarrow \mathcal{X}_j = k$ for $k \in \{1, \dots, p_j\}$. We define $\mathcal{X} := (1, \mathcal{X}_{1,1}, \dots, \mathcal{X}_{1,p_1}, \dots, \mathcal{X}_{J,1}, \dots, \mathcal{X}_{J,p_J})$ as the vector of the dummy random variables and $\boldsymbol{x} \in \mathbb{R}^{p+1}$ as a realization of $\mathcal{X}$. The underlying logistic regression model is

$$\mathbb{E}(Y|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x\beta})}{1 + \exp(\boldsymbol{x\beta})},$$

where $\boldsymbol{\beta} := (\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)^T \in \mathbb{R}^{p+1}$ is the vector of regression coefficients, with β_0 denoting the intercept and $\boldsymbol{\beta}_j := (\beta_{j,1}, \ldots, \beta_{j,p_j})$ being the parameter subvector corresponding to the j-th factor.
Given an observed sample of size $n \in \mathbb{N}$, the general approach of penalized regression consists of the selection of a penalty function $P_\lambda(\boldsymbol{\beta})$ and the minimization of the objective function

$$M_{pen}(\boldsymbol{\beta}) := -L_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}),$$

where $-L_n(\boldsymbol{\beta})$ denotes the negative log-likelihood function and λ is a tuning parameter. There are approaches performing *variable selection* (e.g. Lasso) and extensions performing *factor selection* (e.g. Group Lasso), where the latter either includes or excludes a *whole* factor from the model. Furthermore, there exist methods for levels fusion, i.e. merging levels of a factor having the same influence, which apply a penalization method to all (pairwise or adjacent) differences of levels of a factor. Examples of such approaches are given by Gertheiss and Tutz (2010) and Oelker et. al. (2014). For the purpose of performing both factor selection as well as levels fusion *simultaneously*, we introduced the L_0-Fused Group Lasso (L_0-FGL) method in Kaufmann and Kateri (2022), that is, the penalty function

$$P_\lambda(\boldsymbol{\beta}) := \lambda_1 \sum_{j=1}^{J} \sqrt{p_j} ||\boldsymbol{\beta}_j||_2 + \lambda_0 \sum_{j=1}^{J} \sum_{0 \le r < s \le p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0,$$

where $\lambda := (\lambda_1, \lambda_0) \in \mathbb{R}^{\ge 0} \times \mathbb{R}^{\ge 0}$ and $w_0^{(j,rs)}$ are optional weights for $j \in \{1, \ldots, J\}$, $r, s \in \{0, \ldots, p_j\}$, $r \neq s$. For details on tuning, as well as coefficient paths, computational approaches, simulation studies and theoretical properties we refer to Kaufmann and Kateri (2022).

2 Likelihood ratio tests

The next step is to investigate statistical inference analysis for the (eventually merged) influential factors *after* the application of the regularization method. To do so, we extend the sample splitting approach introduced by Wassermann and Roeder (2009), as well as the multiple sample spliting approach introduced by Meinshausen et. al. (2009), for the L_0-FGL penalization method. The crucial issue is to adjust (and ensure) the regularity conditions needed to guarantee convenient properties of the likelihood ratio test statistic, e.g. its convergence to a χ^2 distribution, to ensure applicability in practice (see Section 2.2).

2.1 Sample splitting

We *randomly* split the dataset $\mathcal{D}$ of sample size $n \in \mathbb{N}$ into two different parts, denoted by $\mathcal{D}_1$ and $\mathcal{D}_2$, of (approximately) equal size. On the first dataset $\mathcal{D}_1$ we fit our regularization method L_0-FGL considering the *full* parameter space $\Omega_1 := \mathbb{R}^{p+1}$ of dimension $p+1$. The resulting estimate $\hat{\boldsymbol{\beta}}^{L_0-\text{FGL}}$ and the selected model $\tilde{S}$ induce a *reduced* parameter space Ω_2, i.e. $\Omega_2 := \mathbb{R}^{p^{af}+1}$, where $p^{af} := \sum_{j=1}^{J} p_j^{af}$ and $p_j^{af} := \dim(\hat{\boldsymbol{\beta}}_j^{L_0-\text{FGL}})$, with 'af' standing for 'after fusion'. That is, p_j^{af} is the dimension of the parameter sub-vector of factor j after (possible) fusion occurred by L_0-FGL and p^{af} the dimension of the whole parameter vector, including all by L_0-FGL as *influential* evaluated factors. Then, we perform maximum likelihood estimation (MLE) on $\mathcal{D}_2$ considering the *reduced* parameter space Ω_2. In this step, ensuring necessary regularity conditions, we can perform LRT, assign p-values and so on. The procedure described above is called *two-stage L_0-FGL*. For a visualization we refer to Figure 2.1.

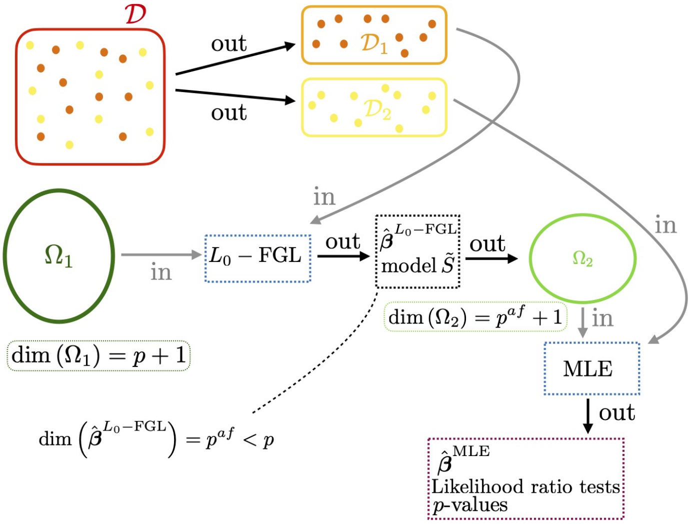

$\mathcal{D}$
out
out
$\mathcal{D}_1$
$\mathcal{D}_2$
in
in
Ω_1
in
$L_0 - \text{FGL}$
out
$\hat{\boldsymbol{\beta}}^{L_0-\text{FGL}}$
model $\tilde{S}$
out
Ω_2
$\dim(\Omega_2) = p^{af} + 1$
in
$\dim(\Omega_1) = p + 1$
MLE
out
$\dim\left(\hat{\boldsymbol{\beta}}^{L_0-\text{FGL}}\right) = p^{af} < p$
$\hat{\boldsymbol{\beta}}^{\text{MLE}}$
Likelihood ratio tests
p-values

FIGURE 1. Visualization of two-stage L_0-FGL with single sample splitting.

2.2 Regularity conditions and details on the tests

To ensure convenient theoretical properties of the MLE/LRT framework, we need, amongst other things, to ensure that the truth $\boldsymbol{\beta}^*$ is an interior point of the parameter space Ω_2 (Casella and Berger, Section 10.6.2). We introduce and analyze *screening properties* for fusion and factor selection (being extensions of known screening properties for variable selection) to guarantee the latter. Further, we consider how we can ensure that the dimension of the parameter space Ω_2 does *not* exceed the sample size of $\mathcal{D}_2$. These conditions need to be discussed in detail to ensure that this method can be applied in practice. Going through all influential factors after L_0-FGL regularization, we test the nested models $\mathcal{M}_0^{(j)}$ (factor j non-influential) against $\mathcal{M}_1^{(j)}$ (factor j influential). Since we execute these tests for all influential factors, we consider two multiplicity adjustments: Bonferrroni correction and Benjamini-Hochberg correction (Benjamini and Hochberg (1995)). For the first approach in the single split, we show that we can asymptotically bound the type-I-error, whereas we show similar results for the Benjamini-Hochberg procedure. Model selection consistency results for two-stage L_0-FGL are also investigated. Further, we consider the multiple split case, where the procedure above is executed a pre-fixed number of times $B \in \mathbb{N}$ aggregating the resulting p-values of each split.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, **57**, 289 – 300.

Casella, G. and Berger, R.L. (2002). *Statistical Inference.* Duxbury, Second Edition.

Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorical explanatory variables. *Annals of Applied Statistics*, **4**, 2150 – 2180.

Kaufmann, L. and Kateri, M. (2022). Simultaneous factors selection and fusion of their levels in penalized logistic regression. *arXiv:2212.10073.*

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-Values for high-dimensional regression. *Journal of the American Statistical Association*, **104**, 1671 – 1681.

Oelker, M.-R., Pößnecker, W. and Tutz, G. (2014). Selection and fusion of categorical predictors with L_0-type penalties. *Statistical Modelling: An International Journal*, **15**, 389 – 410.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics*, **37**, 2178 – 2201.

Bias-reducing adjustments for generalised additive models

Oliver Kemp[1], Ioannis Kosmidis[1]

[1] Department of Statistics, University of Warwick, UK

E-mail for correspondence: Oliver.Kemp@warwick.ac.uk

Abstract: We derive an expression for the asymptotic bias of parameter estimators of generalised additive models, under the assumption that smoothing parameters are $O(1)$. The asymptotic bias decomposes into a term due to penalisation to prevent overfitting of functional components, and a term that mimics the asymptotic small sample bias in generalised linear models. An adjustment to the estimating functions is proposed that eliminates the leading term of the bias expansion. A simulation study on a binomial generalised additive model illustrates the improved properties of the new estimators.

Keywords: Adjusted estimating functions; Penalised likelihood; Smoothing parameter.

1 Introduction

Consider a generalised additive model (GAM), Wood (2017), with

$$g(\mu_i) = \sum_{k=1}^{n_k} \alpha_k z_{ik} + \sum_{j=1}^{p} f_j(x_{ij}) \qquad (i = 1, \dots, n),$$

with $\mu_i = E(Y_i)$, where Y_i is a response variable with an exponential family distribution, $Y_i \sim EF(\mu_i, \phi)$, and $Y_1, \dots, Y_n$ are independent conditionally of the covariance vectors $z_1, \dots, z_n$, $x_1, \dots, x_n$ with $z_i \in \mathbb{R}^{n_k}$, $x_i \in \mathbb{R}^p$. The functions f_j are smooth, each typically represented by a basis expansion. This leads to a generalised linear model (GLM) structure

$$g(\mu_i) = X_i\theta, \quad Y_i \sim EF(\mu_i, \phi),$$

with full parameter vector θ and i-th row of a complete design matrix X. Maximising the likelihood $l(\theta)$ for this model would typically lead to overfitting, so parameter estimation is by maximising a penalised likelihood

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

$$l^P(\theta) = l(\theta) - \frac{1}{2\phi}\theta^{\mathrm{T}}S\theta, \tag{1}$$

with $S = \sum_j \lambda_j S_j$, where S_j is a penalty matrix for function f_j, and λ_j is a smoothing parameter to control the trade off between how well f_j fits the data, and its smoothness. In practice, this maximisation is achieved through a penalised iteratively reweighted least squares (PIRLS) procedure, illustrated for example in Wood (2017).

The aim of this work is to provide a framework for reducing the asymptotic bias of parameters in generalised additive models, in particular the linear parameters which are biased by the penalty term in (1), as well as finite sample bias.

Firth (1993) showed that applying an appropriate adjustment $A(\theta)$ to the score $\nabla_\theta l(\theta)$, and solving the adjusted score equations $\nabla_\theta l(\theta) + A(\theta) = 0_p$ results in an estimator of θ with asymptotically reduced bias compared to the maximum likelihood estimator. In the above, ∇_θ denotes gradient with respect to the parameter θ, and 0_p denotes a p-vector of zeros.

However, the difficulty of GAMs is that the estimating equations are not unbiased due to the presence of the functional penalty term, meaning the adjustments of Firth (1993) can not be directly applied. Therefore, we derive a similar approach in order to obtain bias-reduced estimators of parameters in GAMs.

2 Bias-reducing GAM adjustment

To consider the bias of estimated parameters in a generalised additive model, suppose that $\tilde{\theta}$ is a consistent solution of the equations $\nabla l^P(\theta) = \{l^P_r(\theta)\} = 0$. Assuming that $\lambda = O(1)$, an expansion of $l^P_r(\tilde{\theta})$ about θ using a similar approach to Pace and Salvan (1997) gives that the bias of $\tilde{\theta}$ is

$$\mathbb{E}(\tilde{\theta} - \theta)^r = \underbrace{-\frac{1}{2}j^{rs}j^{tu}(\nu_{s,tu} + \nu_{s,t,u})}_{\text{Bias from the GLM part}} \quad \underbrace{-\lambda j^{rs}S_{st}\theta^t}_{\text{Bias from smoothing penalty}} \quad + \; O(n^{-3/2}). \tag{2}$$

The above expression employs index notation with Einstein's summation convention, and j^{rs} is the inverse of $j_{rs} = \mathbb{E}(l^P_{rs})$, $S_{st}\theta^t$ is the s-th element of the vector $S\theta$, $\nu_{s,tu} = \mathbb{E}(l_s l_{tu})$ and $\nu_{s,t,u} = \mathbb{E}(l_s l_t l_u)$. We have for example $l_{rs} = \partial^2 l(\theta)/\partial\theta^s\partial\theta^r$ for the unpenalised part of the likelihood, with the same notation for the full penalised likelihood l^P. Suppose that estimation is performed by solving $l^P_r + A_r = 0$, where $A_r = O(1)$. By choosing $A_r = \{A(\theta)\}_r$ appropriately, we obtain a vector of bias-reducing adjustments that eliminates the leading term of bias expansion (2) by setting

$$\mathbb{E}(A_r) = i_{rs}\left\{j^{rs}\left(\frac{1}{2}j^{tu}(\nu_{s,tu} + \nu_{s,t,u}) + \lambda S_{st}\theta^t\right)\right\} + O(n^{-1/2}), \tag{3}$$

where i^{rs} is the inverse of $i_{rs} = \mathbb{E}(l_{rs})$.

3 Simulation study

We conduct a simulation study to investigate the properties of the reduced-bias estimators, using the model $Y_i \sim \text{Binomial}(10, \psi_i)$, where

$$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \alpha_0 + \sum_{k=1}^{10} \alpha_k x_{ik} + f(x_{i11}).$$

We use the covariate setup described in Šinkovec et al. (2019) by generating 10 independent standard Gaussian random variables, then applying the listed transformations to obtain a mixture of binary, ordinal and continuous covariates. We examine two cases, first setting the generated Gaussian variables to be uncorrelated, then setting the Gaussian variables to have correlation matrix elements described in Table 1. Continuous covariates are truncated at their third quartile plus five times their interquartile range. The functional covariate x_{i11} is a uniform random variable in $[0, 1]$, and $f(x) = 2\sin(\pi x)$. The true parameters are set to $\log(4)$ for binary covariates, $\log(2)$ for ordinal covariates, $\log(\sqrt{2})$ for α_7, 0 for α_8 and α_9, and finally $-\log(\sqrt{2})$ for α_{10}. The intercept α_0 was estimated numerically in order to ensure that $E(Y_i) = 4$ approximately.

TABLE 1. Structure of linear covariates to be used in the simulation study, as used in Šinkovec et al. (2019). $I(\cdot)$ is the indicator function, [.] is a truncation function, removing the decimal part of a number.

Gaussian variable	Correlation of z_{ik}	Transformation	$\mathbb{E}(x_{ik})$
z_{i1}	$z_{i2}(0.6), z_{i3}(0.5), z_{i7}(0.5)$	$x_{i1} = I(z_{i1} < 0.84)$	0.8
z_{i2}	$z_{i1}(0.6)$	$x_{i2} = I(z_{i2} < -0.35)$	0.36
z_{i3}	$z_{i1}(0.5), z_{i4}(-0.5), z_{i5}(-0.3)$	$x_{i3} = I(z_{i3} < 0)$	0.5
z_{i4}	$z_{i3}(-0.5), z_{i5}(0.5), z_{i7}(0.3), z_{i8}(0.5), z_{i9}(0.3)$	$x_{i4} = I(z_{i4} < 0)$	0.5
z_{i5}	$z_{i3}(-0.3), z_{i4}(0.5), z_{i8}(0.3), z_{i9}(0.3)$	$x_{i5} = I(z_{i5} \geq -1.2) + I(z_{i5} \geq 0.75)$	1.11
z_{i6}	$z_{i7}(-0.3), z_{i8}(0.3)$	$x_{i6} = I(z_{i6} \geq 0.5) + I(z_{i6} \geq 1.5)$	0.37
z_{i7}	$z_{i1}(0.5), z_{i4}(0.3), z_{i6}(-0.3)$	$x_{i7} = [10z_{i7} + 55]$	54.5
z_{i8}	$z_{i4}(0.5), z_{i5}(0.3), z_{i6}(0.3), z_{i9}(0.5)$	$x_{i8} = [\max(0, 100\exp(z_{i8}) - 20)]$	138.58
z_{i9}	$z_{i4}(0.3), z_{i5}(0.3), z_{i8}(0.5)$	$x_{i9} = [\max(0, 80\exp(z_{i9}) - 20)]$	106.97
z_{i10}	-	$x_{i10} = [10z_{i10} + 55]$	54.5

For $n = 100$ and $k = 10$ knots used with a first order P-spline, 1000 datasets were generated using the above covariates for each of the uncorrelated and correlated cases. Parameters were estimated by solving $\nabla_\theta l^P(\theta) = 0$, and by using adjustments (3), using the `nleqslv` package in `R` version 4.3.2. The smoothing parameter λ was set to each of $\{0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4\}$, as well as a run where the smoothing parameter was estimated using `mgcv::gam()`. Estimated biases and MSEs for the scalar parameters are plotted in Figures 1 and 2 below. The average bias and MSE for the parameters when we use `mgcv` estimated smoothing parameters are plotted as horizontal dotted lines. We observe a general reduction in bias for the linear parameters in both the fixed smoothing parameter and `mgcv` selected smoothing parameter cases, particularly for the uncorrelated covariates. The biases for parameters α_2 and α_3 increase in the correlated case for the larger smoothing parameter values, but other parameters still obtain bias reduction. We also observe MSE remaining similar or slightly smaller when using the bias-reducing adjustments.

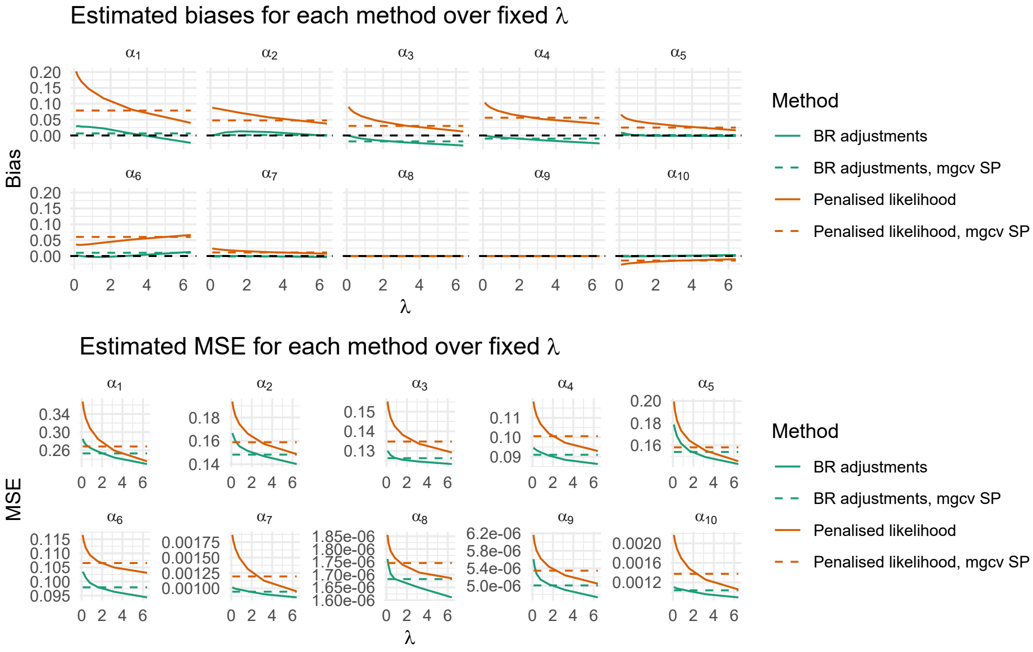


FIGURE 1. Estimated biases and MSEs when solving $\nabla_\theta l^P(\theta) = 0$ (penalised likelihood), and when applying bias-reducing adjustments; uncorrelated case.

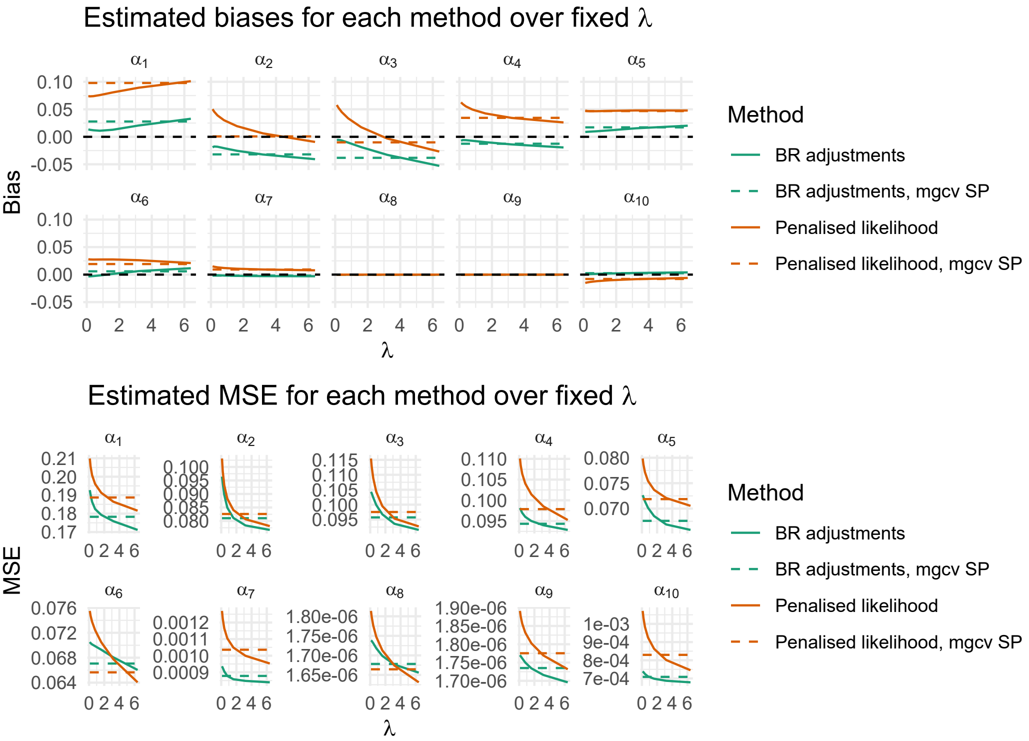


FIGURE 2. Estimated biases and MSEs when solving $\nabla_\theta l^P(\theta) = 0$ (penalised likelihood), and when applying bias-reducing adjustments; correlated case.

Ongoing work involves developing bias reduction methods when $\lambda = O(n^{1/2})$. A seamless smoothing parameter selection when using this method is also a current point of research.

References

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

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*, Volume 4, World scientific.

Šinkovec, H., Geroldinger, A. and Heinze, G. (2019). Bring more data!-A Good Advice? Re-moving separation in logistic regression by increasing sample size. *International Journal of Environmental Research and Public Health*, **16**, 4658.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall / CRC Texts in Statistical Science.

Anchor regression to enhance transferability of genetic prediction models

Hannah Klinkhammer[1,2], Andreas Mayr[1], Carlo Maj[3], Peter M. Krawitz[2], Christian Staerk[4,5]

[1] Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Germany
[2] Institute for Genomic Statistics and Bioinformatics, University of Bonn, Germany
[3] Center for Human Genetics, Philipps University Marburg, Germany
[4] IUF–Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany
[5] Department of Statistics, TU Dortmund University, Germany

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

Abstract: In polygenic risk modelling, finding prediction models that yield competitive performance across different ancestry groups remains one of the greatest challenges. Anchor regression is a recent regression technique that combines ideas from causal inference and predictive modelling to achieve robust predictions also in unseen environments. In this work, we investigate the potential of anchor regression to reduce the gap in transferability in genetic prediction models. Specifically, we incorporate anchor regression in the statistical boosting framework snpboost, which enables the computationally efficient derivation of multivariable, sparse polygenic risk models from individual-level genotype data. As a proxy for ancestry, we use the first ten genetic principal components as anchors to derive prediction models that are robust against perturbations across different populations. We then apply anchor regression to large-scale BMI data from the UK Biobank and analyse the prediction performance.

Keywords: Anchor regression; Polygenic risk scores; Robust prediction.

1 Introduction

Polygenic risk scores (PRS) capture the genetic predisposition to a specific trait (e.g., a disease or phenotype) based on common genetic variants, so-called single nucleotide polymorphisms (SNPs). PRS are most often derived from large cohort studies like the UK Biobank, which are often biased towards European

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

ancestry in their population structure. The predictive accuracy of PRS that were derived from individuals with a shared ancestry (e.g., European) tends to be strongly decreased in individuals from other ancestries (e.g., African or Asian). To overcome this issue of lacking PRS transferability, there is a strong need for more diverse biobanks, as well as for methods that effectively include different ancestries to achieve more robust predictions across out-of-target populations or admixed individuals.

The ancestry of an individual can be estimated from their genetic principal components (PCs). Therefore, genetic PCs are often used as covariates in PRS analyses to account for different population structures. Tanigawa and Kellis (2023) observed that including non-European individuals in the training data of PRS considerably improved the prediction performance on non-European populations without reducing the performance on the European population. However, it is unclear how to best account for the ancestry in the training of the PRS model. Therefore, in a first analysis, we tested whether it is crucial to include the PCs before deriving the PRS or if it is sufficient to correct for them in the subsequent analysis. We did not observe a consistent pattern here; however, in some cases and particularly when only training on European samples, including the PCs as simple covariates in a linear regression even decreased transferability.

In previous works, we have introduced snpboost, a statistical boosting framework to infer PRS directly from individual-level genotype data (Klink-hammer et al., 2023). Within snpboost, a specific regression task is solved by iteratively fitting simple linear base-learners, each base-learner corresponding to one genetic variant. The regression task is defined by the chosen loss function; so far, the snpboost framework incorporates targeted loss functions for linear and logistic regression, as well as for quantile regression. Furthermore, it is possible to adequately model time-to-event and count data by using corresponding loss functions. In Klinkhammer et al. (2023), we found that snpboost, when applied to European data, yields sparser models with competitive or better prediction performance compared to multivariable approaches (e.g., snpnet, BayesR) and approaches based on summary statistics from genome-wide association studies (e.g., PRScs, LDpred). However, none of the incorporated loss functions specifically targets the gap in transferability to different populations.

Anchor regression is a regression technique that aims to provide robust predictions across out-of-target environments (Rothenhäusler et al., 2021). The environments are represented by so-called anchors and account for shift perturbations in the covariates as well as the outcome. In this work, we extend the statistical boosting framework snpboost to include anchor regression by incorporating the corresponding loss function. The aim is to derive sparse polygenic risk scores that are potentially more transferable to out-of-target populations. To increase robustness with respect to an individual's ancestry, we use the genetic PCs as anchors. We apply anchor regression on data from the UK Biobank, a large cohort study with extensive phenotypic and genotypic information. In particular, we examine the influence of the tuning parameter γ, which determines the strength of the anchors' influence.

2 Boosting polygenic risk scores via anchor regression

Anchor regression transfers ideas from causal inference into predictive modelling (Rothenhäusler et al., 2021). It assumes a situation where we are interested in the effects of some covariates X on an outcome Y, but the distribution of X is perturbed in different environments, represented by some anchors A. It then aims to estimate the effects such that the prediction performance is more robust across different environments including even unobserved ones.

In the context of genetic prediction models, for n individuals and a continuous outcome Y, let $\boldsymbol{y} \in \mathbb{R}^n$ be the vector of observations and $\boldsymbol{X} \in [0,2]^{n \times p}$ the observed genotype matrix. Its j-th column $\boldsymbol{x}_j$ corresponds to the j-th variant which is encoded as $x_{i,j} = 0$ if individual i has no mutation in variant j compared to the reference genome and $x_{i,j} = 1$ and $x_{i,j} = 2$ in case of heterozygous and homozygous mutations, respectively. Furthermore, let $\boldsymbol{\mu} = \boldsymbol{X\beta}$ with $\boldsymbol{\beta} \in \mathbb{R}^p$ be some linear predictor of $\boldsymbol{y}$. Moreover, for k anchors, let $\boldsymbol{A} \in \mathbb{R}^{n \times k}$ be the anchor matrix. Then, for a tuning parameter $\gamma \geq 0$, the loss function of anchor regression is given by

$$\rho(\boldsymbol{y}, \boldsymbol{\mu}) = \|(\mathbb{I} - \boldsymbol{\Pi_A})(\boldsymbol{y} - \boldsymbol{\mu})\|_2^2 + \gamma\|\boldsymbol{\Pi_A}(\boldsymbol{y} - \boldsymbol{\mu})\|_2^2,$$

where $\boldsymbol{\Pi_A} = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ is the projection matrix onto the column space of $\mathbf{A}$ and $\mathbb{I}$ is the identity matrix. For $\gamma = 1$, the loss function equals the L_2-loss, i.e., in this case anchor regression coincides with ordinary least squares regression without taking the anchors into account. For $\gamma = 0$, anchor regression yields the same estimator as regressing both $\boldsymbol{y}$ and $\boldsymbol{X}$ on $\boldsymbol{A}$ first and then only considering the residuals in an ordinary least squares regression. On the other extreme, for $\gamma \to \infty$, and under instrumental variable assumptions, anchor regression corresponds to two-stage least squares regression, a technique to estimate causal effects. Anchor regression interpolates between those edge cases and offers a flexible range of models under less assumptions than the classical instrumental variable set-up (Rothenhäusler et al., 2021).

To incorporate anchor regression in the snpboost framework, we implemented $\rho(\cdot,\cdot)$ as a loss function and added γ as an additional input parameter. Furthermore, to include the special case $\gamma = \infty$, we implemented the loss function

$$\rho_{\mathrm{IV}}(\boldsymbol{y}, \boldsymbol{\mu}) = \|\boldsymbol{\Pi_A}(\boldsymbol{y} - \boldsymbol{\mu})\|_2^2.$$

The choice of anchors depends on the kind of perturbations that should be accounted for. As we aim for robust predictions across different ancestry populations, we use the first ten genetic PCs as anchors. The genetic PCs reflect the individual ancestry and are often incorporated in polygenic risk modelling to account for population structure. Unlike a categorical variable encoding an individual's ancestry to a population or ethnic group, PCs can portray ancestry more adequately as a continuum and therefore also capture admixed individuals.

In anchor regression, the choice of the tuning parameter γ is crucial. Rothenhäusler et al. (2021) suggest that this parameter could be determined based on the expected strength of perturbations on future data relative to the training set. Building up on this, in our work we particularly investigate the impact of γ on the prediction performance across different ancestries.

3 Analysis of large-scale UK Biobank data

We applied the new anchor regression via snpboost approach on BMI data from the UK Biobank (under application number 81202). The genotyped data comprised p = 619,773 genetic variants. For varying γ, models were trained on

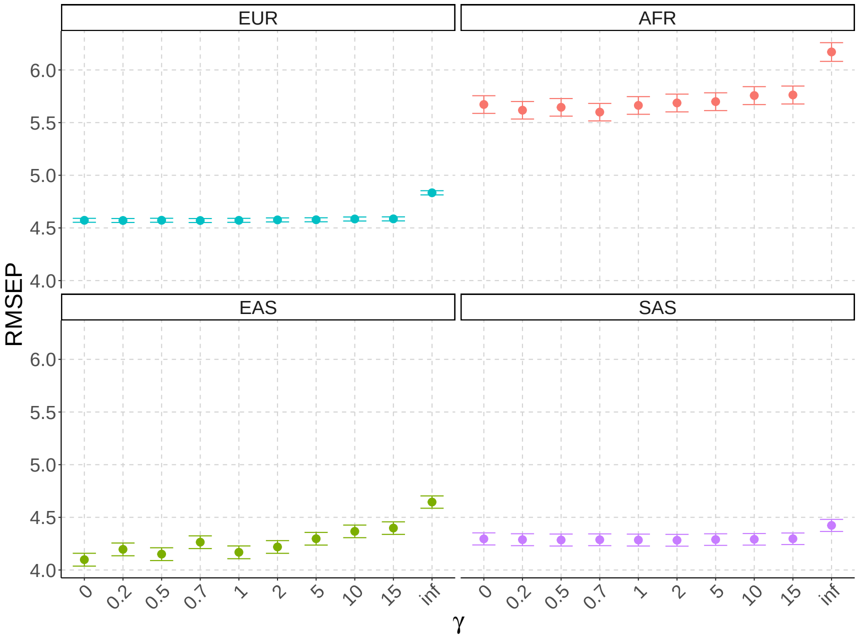


FIGURE 1. Results of anchor regression for BMI data from the UK Biobank via snpboost for varying γ. Models were trained on individuals of European (EUR) ancestry and tested on independent individuals of European (EUR), African (AFR), East Asian (EAS) and South Asian (SAS) ancestry, respectively. RMSEP ± SE on the test data are shown.

$n_{\text{train}} = 191{,}036$ individuals of European ancestry and the stopping iteration of snpboost was tuned on $n_{\text{val}} = 64{,}193$ individuals of European ancestry. The resulting models were then applied on $n_{\text{test,EUR}} = 63{,}639$ individuals of European ancestry, $n_{\text{test,AFR}} = 5{,}397$ individuals of African ancestry, $n_{\text{test,EAS}} = 1{,}924$ individuals of East Asian ancestry and $n_{\text{test,SAS}} = 6{,}781$ individuals of South Asian ancestry, respectively. Prediction performance was assessed via the root mean squared error of prediction (RMSEP) on the test sets.

Results on the prediction performance are shown in Figure 1. For individuals of South Asian ancestry, the prediction performance hardly differed for varying values of γ, while for the other ancestries the RMSEP increased with increasing γ. Overall, the prediction performance was quite stable in all ancestries for small to medium-sized γ. Noteworthy, the optimal γ differs dependent on the ancestry of interest. On another note, the resulting PRS models tended to be sparser for larger values of γ (Figure 2). For $\gamma \to \infty$, the final model only includes 21 SNPs but also yields the highest RMSEP for all ancestries. On the other hand, the edge case $\gamma = 0$ includes 19,011 variants but yields a comparable prediction performance to e.g. the model corresponding to $\gamma = 5$, incorporating almost half as many variants.

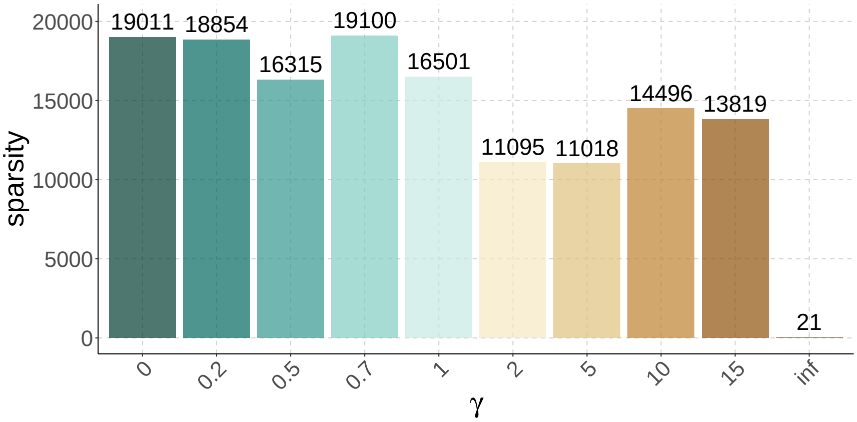


FIGURE 2. Number of included genetic variants (sparsity) of PRS models for BMI fitted on European data from the UK Biobank via anchor regression within the snpboost framework. Numbers on top of the bars indicate the number of variables that form the final model.

4 Discussion

We were able to show that the proposed anchor regression approach can be efficiently applied to derive PRS on large-scale genotype data via the snpboost framework. Our first analysis on UK Biobank indicates stable prediction performance of anchor regression for small to medium-sized γ, while the performance is still varying between ancestry groups and further work on exploring the potential is needed. Specifically, future work should focus on the choice of anchors and investigate whether the results are sensitive to the number of included PCs, or if other anchors (e.g., categorical coding of ancestry) would be more suitable. Additionally, it is of interest to analyse the effect of including individuals with non-European ancestry in the training of the PRS (cf., Tanigawa and Kellis, 2023). As the final prediction models tend to be sparser for larger values of γ with still competitive prediction performance, it may be of particular interest to use anchor regression as a variable selection technique to identify relevant genetic variants with “stable” effects for different populations and potentially integrate a refitting step to optimize prediction performance. Furthermore, we plan to investigate the chosen variants in more detail, e.g., examine if less variants in LD are chosen in the sparser models. With the help of simulated genotype and phenotype data, we will also analyse if increasing the parameter γ and therefore yielding sparser models could reduce the false positive rate. If this was the case, anchor regression might help detecting potentially causal variants and simplify biological downstream analysis.

Besides increasing transferability across ancestries, with the choice of adequate anchors, anchor regression could also help to account for environmental changes, e.g., for combined data from different studies. Noteworthy, for predictions on new test data, the values of the anchors are not needed. Therefore, anchor regression offers the opportunity to account for covariates that are observed in the training data but might be unknown in future test data.

References

Klinkhammer, H., Staerk, C., Maj, C., Krawitz, P. and Mayr, A. (2023). A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Frontiers in Genetics*, **13**.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P. and Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **83**, 215 – 246.

Tanigawa, Y. and Kellis, M. (2023). Power of inclusion: Enhancing polygenic prediction with admixed individuals. *American Journal of Human Genetics*, **110**, 1888 – 1902.

Spatial confounding in gradient boosting

Lars Knieper [1], Thomas Kneib [2], Elisabeth Bergherr[1]

[1] Chair of Spatial Data Science and Statistical Learning, University of Göttingen, Germany
[2] Chair of Statistics, University of Göttingen, Germany

E-mail for correspondence: `lars.knieper@uni-goettingen.de`

Abstract: Collinearity between covariates and spatial effects can lead to a bias in the corresponding fixed effects' estimates known as spatial confounding. Recently, the *Spatial+* approach suggests to regress out the spatial effect in the covariate first, before estimating the model of interest. Drastic spatial confounding is observed in gradient boosting due to its step-wise procedure. In this contribution we apply the suggested two-step approach and confirm its ability to correct spatial confounding for gradient boosting as well.

Keywords: Gradient boosting; Spatial confounding; Spatialplus.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Shape analysis of AF segments for rapid assessment of Mohs layers for BCC presence by AF-Raman microscopy

Alexey A. Koloydenko[1], Ioan Notingher[2], Radu Boitor[2], Jüri Lember[3]

[1] Department of Mathematics, Royal Holloway, University of London, Egham, UK
[2] School of Physics and Astronomy, University of Nottingham, Nottingham, UK
[3] Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia

E-mail for correspondence: alexey.koloydenko@rhul.ac.uk

Abstract: An automated method to detect Basal Cell Carcinoma (BCC) relies on Autofluorescence (AF) imaging guiding Raman microscopy to obtain biochemical information for tissue classification. The guidance is provided via an image segmentation technique aiming to reduce the risk of missing cancer. We present evidence that shape of an AF segment may be useful for 'trimming' 6-8% of non-BCC segments in an essentially coordinate free manner, without compromising BCC detection. By allowing the AF-Raman method to direct the more time consuming Raman analysis toward more relevant regions, the proposed trimming of unnecessary segments should ultimately improve the overall accuracy. The presented shape analysis uses the recently introduced Weighted Euler Curve Transform (WECT). WECT embeds segments in a space of real matrices of fixed dimensions, where a shape is an equivalence class of segments with WECTs matching under a cyclic permutation of columns. The induced rotation invariant distance is non-Hilbertian, which requires special care in using it with kernel methods (e.g. Kernel PCA, Kernel LDA, SVMs). Our currently best results are achieved by L_1 SVMs based on the Laplace 'kernel'.

Keywords: Autofluorescence imaging; Raman spectroscopy; skin cancer; Weighted Euler curve transform; Shape analysis.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

Inferential tools for hidden Markov models with periodic components

Jan-Ole Koslik[1], Carlina C. Feldmann[1], Sina Mews[1], Rouven Michels[1], Roland Langrock[1]

[1] Department of Business Administration and Economics, Bielefeld University, Germany

E-mail for correspondence: r.michels@uni-bielefeld.de

Abstract: State processes of hidden Markov models with periodic components are a special case of inhomogeneous Markov chains. The cyclic patterns exhibited by such chains can be utilised to infer the time-varying unconditional stationary distribution and the overall state dwell-time distribution. We model data from movement ecology to show that the previously used approximation of the true stationary distribution can be biased and that the unrealistic consequence of geometrically distributed state dwell times vanishes for HMMs with periodicity.

Keywords: Markov chain; Seasonality; Stationary distribution; Time series.

1 Introduction

The capability to incorporate periodic trends into hidden Markov models (HMMs) is of tremendous significance in certain application domains. For instance, it enables modelling the volatility of financial market based on seasonal variations or animal behaviour patterns depending on the time of day. An essential aspect in the subsequent interpretation of the effects lies in the state distribution. In inhomogeneous Markov chains, this distribution is non-stationary and thus challenging to depict accurately. In this contribution, we derive properties of *periodically* inhomogeneous HMMs and utilise elephant movement data to demonstrate the necessity of the elaborated inferential tools.

2 Periodic variation in hidden Markov models

An HMM consists of two processes: one state-dependent process $\{X_t\}_{t\in\mathbb{N}}$ (where X_t can be a vector) and one latent state process $\{S_t\}_{t\in\mathbb{N}}$, where $S_t \in \{1,\dots,N\}$ selects from which of the N state-dependent distributions X_t is generated. The unobserved state process $\{S_t\}$ is assumed to be Markovian of order one and fully characterised by its initial distribution and the (potentially time-varying) transition probability matrix (t.p.m.)

$$\boldsymbol{\Gamma}^{(t)} = (\gamma_{ij}^{(t)}), \quad \text{with} \quad \gamma_{ij}^{(t)} = \Pr(S_{t+1} = j | S_t = i), \quad t \in \mathbb{N}.$$

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

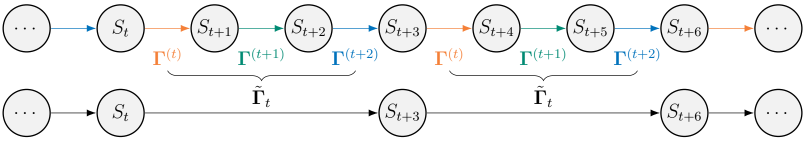


FIGURE 1. Example visualisation of *periodic stationarity* with $L = 3$. The thinned Markov chain $S_t, S_{t+3}, S_{t+6}, \dots$ has constant t.p.m. $\tilde{\mathbf{\Gamma}}_t$.

In contrast, the observations X_t, $t \in \mathbb{N}$, are assumed to be conditionally independent of each other, given the current states.

In many settings, modelling variation over time in the state process is deemed necessary. For instance, in ecological applications, S_t proxies the behavioural mode of an animal at time t, such as resting, foraging or travelling, which crucially depends on the time of the day. When accounting for diel rhythms, it is straightforward to see that $\mathbf{\Gamma}^{(t)} = \mathbf{\Gamma}^{(t+L)}$, for $t \in \mathbb{N}$, with L denoting the length of a cycle.

However, in general, interpreting these transition probabilities with respect to the time is not straightforward, especially not for $N > 2$. One way to circumvent this problem is to consider the hypothetical stationary distribution that the Markov chain would converge to if the process followed $\mathbf{\Gamma}^{(t)}$ constant over time (for given t), i.e. the solution to

$$\boldsymbol{\rho}^{(t)} = \boldsymbol{\rho}^{(t)}\mathbf{\Gamma}^{(t)}$$

for each $t = 1, \dots, L$, subject to $\sum_{i=1}^{N} \rho_i^{(t)} = 1$ (Patterson et al., 2009). However, this assumption will lead to biased estimates of the *true* unconditional distribution of the state process as the preceding dynamics of the state-switching probability matrix are neglected. Fortunately, in case of periodicity, we can use that $\mathbf{\Gamma}^{(t)} = \mathbf{\Gamma}^{(t+L)}$ and consider the thinned Markov chain $\{S_{t+kL}\}_{k\in\mathbb{N}}$, for fixed t, which is homogeneous with a constant t.p.m.

$$\tilde{\mathbf{\Gamma}}_t = \mathbf{\Gamma}^{(t)}\mathbf{\Gamma}^{(t+1)} \dots \mathbf{\Gamma}^{(t+L-1)},$$

see Figure 1 for a visual representation. This thinned Markov chain (given it is irreducible) has a unique stationary distribution $\boldsymbol{\delta}^{(t)}$ (Kargapolova and Ogorodnikov, 2012) being the solution to

$$\boldsymbol{\delta}^{(t)} = \boldsymbol{\delta}^{(t)}\tilde{\mathbf{\Gamma}}_t.$$

In addition to the stationary distribution, we want to obtain an *overall* state dwell-time distribution as a simpler inference tool describing the distribution of the dwell-time in each state. Thus, we derive the probability mass function of the overall dwell-time distribution in state i for a periodically stationary Markov chain defined by $\mathbf{\Gamma}^{(t)}$, $t = 1, \dots, L$, as

$$d_i(r) = \sum_{t=1}^{L} w_i^{(t)} d_i^{(t)}(r), \qquad r \in \mathbb{N},$$

where $w_i^{(t)}$ denote mixture weights depending on the previously derived periodically stationary distribution and transition probabilities and

$$d_i^{(t)}(r) = \left(1 - \gamma_{ii}^{(t+r-1)}\right) \prod_{j=1}^{r-1} \gamma_{ii}^{(t+j-1)}, \; r \in \mathbb{N}.$$

3 Application to elephant data

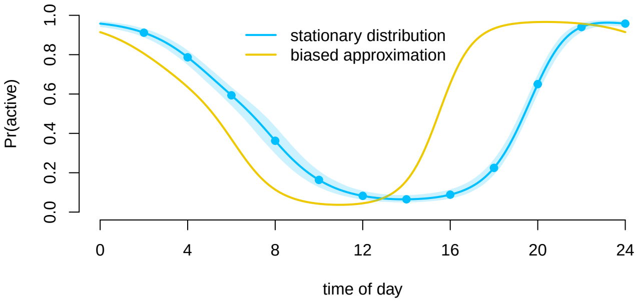


FIGURE 2. Periodically stationary distribution in the elephant example.

In the following case study, we use a dataset of a movement track of an elephant from the Ivory Coast consisting of 12,170 observations of longitude and latitude, recorded at 2-hour intervals. The data is accessible via the Movebank for animal tracking data (Movebank-ID: 2736765655). These observations were then converted into step lengths and turning angles to model them by gamma and van Mises distributions, respectively, in a 2-state HMM where the transition probabilities were modelled via trigonometric functions comprising one sine and one cosine term to account for the cyclic nature. Applying the inferential tools developed in Section 2, Figure 2 displays the time-dependent unconditional state distributions as well as its approximation. It reveals the aforementioned bias in the latter one as the correct stationary distribution is shifted, caused by the ignoration of the preceding dynamics in the approximative version.

Additionally, Figure 3 demonstrates the overall dwell-time distribution in both states. It can easily be seen that the arising distribution substantially deviates from a geometric distribution, which is the standard case implied by the Markov property in HMMs with homogeneous transition probabilities. Thus, this undesirable consequence of the Markov property can be avoided by including periodic components in the state process.

4 Discussion

This paper derives important properties of hidden Markov models in cases of periodic variation. Specifically, we demonstrated that for state processes including cyclic components, the periodic structure present in the inhomogeneity can

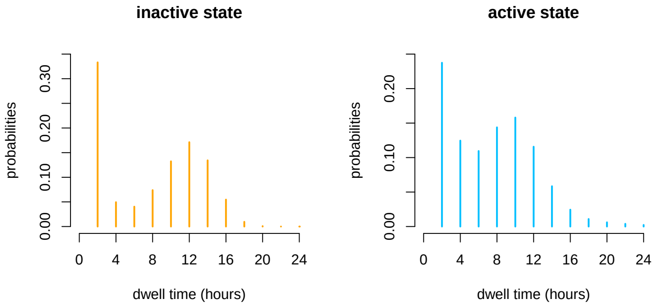


FIGURE 3. Visualisation of the overall dwell-time distribution in both states.

be exploited to derive the periodically varying unconditional state distribution. Based on this, an overall dwell-time distribution can also obtained which can — implied by the periodicity — deviate rather substantially from a geometric distribution. However, this only holds true for periodic HMMs. To circumvent this problem also in cases without periodicity, hidden semi-Markov models are a natural alternative. Elaborating more on this model class constitutes an exciting path for future research.

References

Kargapolova, N. A. and Ogorodnikov, V. A. (2012). Inhomogeneous Markov chains with periodic matrices of transition probabilities and their application to simulation of meteorological processes. *Russian Journal of Numerical Analysis and Mathematical Modelling*, **27**, 213 – 228

Patterson, T. A., Basson, M., Bravington, M. V. and Gunn, J. S. (2009). Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *Journal of Animal Ecology*, **78**, 1113 – 1123

Wikelski, M., Davidson S.C. and Kays R. (2024). Movebank: archive, analysis and sharing of animal movement data. Hosted by the *Max Planck Institute of Animal Behavior*. www.movebank.org, ID: 2736765655.

Bayesian ordinal regression for crowd-sourced fact-checking

Michele Lambardi di San Miniato[1], Michela Battauz[1], Ruggero Bellio[1], Paolo Vidoni[1]

[1] Department of Economics and Statistics, University of Udine, Italy

E-mail for correspondence: michele.lambardi@uniud.it

Abstract: Fact-checking can be done in crowd-sourcing mode by aggregating judgments provided by many workers. We propose a Bayesian approach to carry out this aggregation, considering the truthfulness of the statements as a latent variable that underlies the ratings. We illustrate this approach and test it against alternative methods using a publicly available dataset.

Keywords: Bayesian statistics; Cross-validation; Fact-checking; Crowd-sourcing.

1 Introduction

Fact-checking is crucial to improve the quality of public discussions. In the internet age, statements and claims flow in at overwhelming rates, but only a handful of experts will check them. One possibility is to surrogate each expert judgment with an aggregation of many workers' judgments; this is called crowd-sourcing. Considering the truthfulness of the statements as a latent variable to be assessed and the ratings given by workers as repeated measurements of the latent construct, these data can be analyzed within the framework of Item Response Theory (IRT; van der Linden, 2016). An ordinal regression model can include worker-specific parameters to account for their personal scaling. Here, we present a Bayesian ordinal regression to do fact-checking in crowd-sourcing mode and compare it with some alternative methods. We illustrate the methodology with the SIGIR data presented by Roitero et al. (2020).

2 Proposal

Let there be n statements indexed by $i = 1, \dots, n$ and m workers indexed by $j = 1, \dots, m$. An expert may rate the i-th statement as Y_i, while the j-th worker may rate it independently as X_{ij}. The aim is to surrogate the Y's via the X's so

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

that, after suitable training, it is possible to rate statements via crowds instead of experts.

We assume that the Y's and X's are on an ordinal scale with K levels indexed by $h = 1, \dots, K$. In the ordinal regression framework, it is natural to assume that Y_i and X_{ij} have latent continuous counterparts Y_i^* and X_{ij}^*. Then, the observation model is

$$Y_i = h \iff Y_i^* \in \mathcal{I}_h^Y \,, \quad X_{ij} = h \iff X_{ij}^* \in \mathcal{I}_h^X \,,$$

depending on two partitions of $\mathbb{R}$ into intervals $\mathcal{I}_h^Z =]\gamma_{h-1}^Z, \gamma_h^Z]$ with ordered cut-points γ_h^Z constrained as $\gamma_0^Z = -\infty$, $\gamma_K^Z = +\infty$, for $Z = X, Y$.

Let $\xi_i \sim \mathcal{N}(0,1)$ denote the latent truthfulness value of the i-th statement; assuming a standardized distribution is common in the IRT framework. The expert ratings are modelled as

$$Y_i^* = \sigma_\xi \xi_i + \epsilon_i \,, \quad \epsilon_i \sim \mathcal{N}(0,1) \,,$$

where $\sigma_\xi > 0$ implies the expert's reliability $\rho = \sigma_\xi^2/(1+\sigma_\xi^2)$. Workers' rating behaviour is described by parameters $\alpha_j \sim \mathcal{N}(0, \sigma_\alpha^2)$ and $\beta_j \sim \mathcal{N}(0, \sigma_\beta^2)$ according to the following model

$$X_{ij}^* = \alpha_j + \beta_j \xi_i + \eta_{ij} \,, \quad \eta_{ij} \sim \mathcal{N}(0,1) \,. \tag{1}$$

Treating α_j, β_j, and ξ_i like latent variables allows for correlation within workers and within statements. The training of the model relies crucially on statements already rated by the expert: for these statements, information is available on ξ_i and, when the j-th worker rates them, some information is gathered on their α_j and β_j. The same worker can then help provide ratings on statements for which the expert rating is missing.

The model in Equation (1) can help classify the rating behaviour of workers. In the long run, workers may turn out to be spammers, with $\beta_j < 0$, if they systematically disagree with the expert; they may simply give random answers, if $\beta_j \approx 0$; however, hopefully, workers may be capable to discern and communicate the truthfulness of statements, since $\beta_j > 0$. Apart from this distinction, it is still possible to be generally in favour of $(\alpha_j > 0)$ or against $(\alpha_j < 0)$ the statements.

The model may be estimated more naturally within the Bayesian framework. One can efficiently simulate draws from the posterior distribution with available software, such as the `R` interface to the `Stan` probabilistic programming language (Stan Development Team, 2024).

The Bayesian approach has some advantages since it allows the smooth handling of several latent variables. The prior is crucial to ensure identifiability and regularization, which are also central in IRT. We adopt an informative prior for σ_ξ^2, one that allows us to achieve experts' reliability $\rho \approx 99\%$. The prior on the other parameters, namely, α_j and β_j, should instead be weakly informative (see, for instance, Gelman et al., 2013, Ch. 5) as it is potentially unknown whether a new worker can be relied upon. In practical crowd-sourcing, one may delegate rating tasks over online platforms where a wide variety of people can participate in many projects, but reliability reports from previous participations may be unavailable.

3 Comparison with alternative methods

The proposed model can be compared with other methods, which seem natural for this field of applications, based on the predictions it provides on new statements.

We surrogate this assessment via Leave-One(-Y_i)-Out cross-validation (LOO), which involves hiding the expert rating for the generic i-th statement and estimate the distribution of Y_i only based on Y_{-i} and all the X_{ij}. For ease of assessment, our Bayesian model is such that all the Y_i's are conditionally independent given the parameters, so the conditions set by Vehtari et al. (2017) are satisfied, which allows to approximate LOO with a single fit of the model based on the full dataset. We thus approximate LOO via Pareto-smoothed importance sampling, which only needs the likelihood terms $\Pr(Y_i = y_i \mid \theta)$ as additional quantities during the model fitting step.

Maximum a posteriori is a common way to summarize predictive distributions, so it was chosen as a prediction to be compared across methods. Alternatives to our proposal include the median and Naive Bayes (NB, Hastie et al., 2009). The median prediction uses the sample medians of workers' ratings within each statement as surrogates for the expert's ratings. Naive Bayes may be stratified by workers or their covariates. The median accounts properly for the ordinal nature of Y, while NB treats all variables as categorical. Our proposal constitutes an improvement over both these methods, as it incorporates information on Y and assumes some structure for the workers' rating behaviour.

For each method, we compare the LOO prediction with the observed values of Y, which implies a misclassification matrix. These matrices can be summarized further, based on a few metrics: the mean absolute error (MAE), the root mean square error (RMSE), and weighted kappas with linear (κ_1) and quadratic weights (κ_2; Lin et al., 2012), the misclassification rate (MR) and a variant that tolerates errors up to one rank (MR2). Kappas are actually a natural method to assess the agreement between raters, while rates are popular in machine learning and artificial intelligence applications. The optimum is achieved when all the kappas are high and all the other metrics are low.

4 Analysis

For illustration, we present the method based on the SIGIR data, the most challenging dataset available at the following repository:
`https://github.com/KevinRoitero/crowdsourcingTruthfulness`
The SIGIR data contain $n = 122$ statements made by American politicians. Each statement was rated independently by the American fact-checking agency PolitiFact, which serves as the expert, and by 10 out of $m = 200$ crowd workers. Truthfulness is rated on an ordinal scale with $K = 6$ levels, encoded by PolitiFact as

$$1 = \text{pants on fire} < \text{false} < \text{mostly false} < \text{half true} < \text{mostly true} < \text{true} = 6\,.$$

The baseline methods have a low computational footprint, while the Bayesian estimation of our model is more demanding. However, the prediction performance outweighs the costs, as shown in the following.

Figure 1 reports the misclassification matrices for each method, with different observed and predicted values for each row and column, respectively. The median provides predictions that look shrunk too much towards the middle of the scale, as extensively pointed out by Roitero et al. (2020).

Naive Bayes was tested in three variants, one using a single misclassification matrix to predict Y, and the other two assuming a distinct matrix for each group

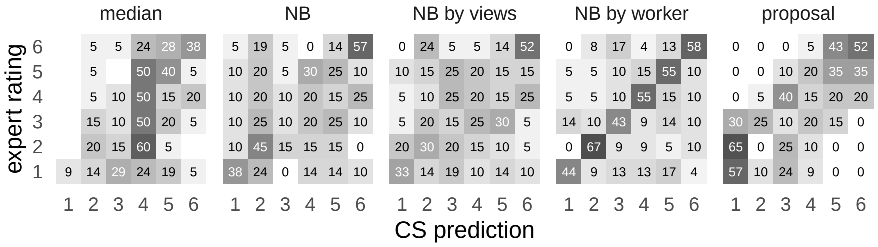


FIGURE 1. LOO misclassification probabilities (%); rows sum to 100.

	median	NB	NB by views	NB by worker	proposal
MAE	1.287	1.402	1.443	1.023	**0.943**
RMSE	1.706	1.961	1.946	1.684	**1.198**
MR	0.721	0.672	0.721	**0.465**	0.713
MR2	0.385	0.385	0.410	0.302	**0.213**
κ1	0.195	0.293	0.256	0.473	**0.535**
κ2	0.298	0.365	0.347	0.512	**0.774**

TABLE 1. Performance metrics; darker highlights are better, bold for best.

of workers, based on their self-declared political views (liberal, conservative, etc), or a distinct matrix for each worker. Such matrices can be rather sparse, so all the counts of events are augmented by a small constant, in order to avoid zero-probability estimates. Based on the misclassification matrices, NB works best when assuming distinct misclassification matrices for each worker, which is cumbersome and still makes coarse prediction errors, so it can be improved upon. Our model is an improvement over the median and NB, both from a modelling standpoint and in terms of performance. The model may miss the target more frequently than the alternatives, but the prediction error is much more under control in the case of misclassification. The model would not label a pants-on-fire statement as true, as the other methods can still do. This fact can be appreciated in terms of metrics, which are reported in Table 1. The proposal is the best performer according to all metrics but MR. However, such a metric is rather questionable in this case, as some levels in the response scale are hardly distinguishable, such as *pants on fire* and *false.*

5 Final remarks

To sum up, when surrogating expert fact-checking with crowd-sourced judgments, it is crucial to account for the workers' misclassification behaviour. Our model assumes a reasonable and intelligible structure over this behaviour, so it is more parsimonious in this sense and scales better with the number of workers. Analyzing the predictive capabilities of the methods is crucial in choosing one that does not contradict the expert excessively, as this may cast shadows over the fact-checking service.

Ongoing research includes testing our proposed solution in a plurality of cases. Future work may address more than one dimension beyond the overall truthfulness of statements, such as the relevance of claims, the perceived damage of trusting the statements, and so on. Further advances may consist of replacing estimation via MCMC with fast variational approximations (Gelman et al., 2013, Ch. 13). Necessarily, reliable algorithms are needed for this possibility to be actively used in practical applications.

References

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis (3rd ed.).* Chapman and Hall/CRC.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning (2nd ed.).* Springer.

Lin, L. , Hedayat, A. S. and Wu, W. (2012). *Statistical Tools for Measuring Agreement (1st ed.).* Springer.

Roitero, K., Soprano, M., Fan, S., Spina, D., Mizzaro, S. and Demartini, G. (2020). Can the crowd identify misinformation objectively? The effects of judgment scale and assessor's background. In: *Proceedings of the 43rd ACM SIGIR on R&D in IR*, China, 439 – 448.

Stan Development Team (2024). RStan: the R interface to Stan. R package version 2.32.5. `https://mc-stan.org/`.

van der Linden, W. J. (2016). *Handbook of Item Response Theory.* Chapman and Hall/CRC.

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**, 1413 – 1432.

Non-parametric estimation of net survival under dependence between death causes

Oskar Laverny[1], Nathalie Graffféo[1], Roch Giorgi[2]

[1] Aix Marseille Univ, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France.

[2] Aix Marseille Univ, APHM, INSERM, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Hop Timone, BioSTIC, Biostatistique et Technologies de l'Information et de la Communication, Marseille, France.

E-mail for correspondence: `oskar.laverny@univ-amu.fr`

Abstract: Relative survival analysis deals with a competing risks survival model where the cause of death is unknown. This lack of information occurs regularly in population-based cancer studies. Non-parametric estimation of the net survival is possible through the Pohar Perme estimator, taking other causes of mortality into account. Derived similarly to Kaplan-Meier, it nevertheless relies on untestable independence assumptions. We propose here to relax these assumptions and provide a generalized estimator that works for other dependence structures, by leveraging the underlying counting process and martingales. Our approach provides a new perspective on the Pohar Perme estimator and the acceptability of this assumption. We showcase the difference between the two estimators on population-based colorectal cancer registry, and discuss potential extensions of the methodology.

Keywords: Survival analysis; Net survival; Non-parametric estimators; Copulas.

1 Net survival analysis

Survival analysis produces valuable tools for prognosis of cancer patients. However, in population-based cancer studies, the cause of death – assumed binary, studied cancer or not – is usually unreliable or unavailable. Relative survival analysis takes this particularity into account to evaluate the excess mortality – due to cancer – with respect to population life tables.
Let E, P and $O = E \wedge P$ be random times to death from (resp) the Excess, Population and Overall mortalities. Let $\mathbf{X}$ be a vector of covariates, C the time to censorship, and denote $T = O \wedge C$ and $\Delta = \mathbf{1}\{T \leq C\}$. Only $(\mathbf{X}, T, \Delta)$ is observable. In particular, we do not observe a potential ordering indicatrix

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

$\mathbf{1}\{E \geq P\}$. Alike standard approaches, we suppose the distribution of $P|\mathbf{X}$ known from population life tables. To simplify our exposition, assume that the rest does not depend on covariates.
Let $(\mathbf{X}_i, T_i, \Delta_i)_{i=1,...,n}$ be an observed n-sample and $(\Omega, \mathcal{A}, \{\mathcal{F}_t, t \in \mathbb{R}_+\}, \mathbb{P})$ the associated filtered probability space, with

$$\mathcal{F}_t = \sigma\{\mathbf{X}_i, (T_i, \Delta_i) : T_i \geq t, \ \forall i \in 1, .., n\}.$$

The mutual independence of (E, P, C) is a central assumption in survival literature (see e.g., Czado & Van Keilegom (2023) for recent discussion), even if the epidemiological interpretation makes it hard to justify. We here suppose only independent censorship, which, denoting $\mathcal{C}$ the survival copula (see Nelsen (2006)) of the random vector (E, P), writes:

$$(\mathcal{H}_\mathcal{C}): \ S_{P \wedge E}(t) = \mathcal{C}(S_P(t), S_E(t)).$$

In particular, the standard independence assumption writes simply $(\mathcal{H}_\Pi)$ for $\Pi(u_1, u_2) = u_1 u_2$. There are a few standard non-parametric estimators under $(\mathcal{H}_\Pi)$, from Ederer (1961), Hakulinen (1982), to more recently Pohar Perme & al (2012), but none under $(\mathcal{H}_\mathcal{C})$. If Adatorwovor & al (2023) provides parametric estimations under $(\mathcal{H}_\mathcal{C})$, we propose here a non-parametric estimator that generalizes the Pohar Perme estimator from $(\mathcal{H}_\Pi)$ to $(\mathcal{H}_\mathcal{C})$.

2 Estimation of the net survival under $(\mathcal{H}_\mathcal{C})$

Let's define the following stochastic processes:

$$\begin{aligned}
N(t) &= \mathbf{1}\{O \leq t, O \leq C\} && \textit{(Uncensored deaths process)} \\
Y(t) &= \mathbf{1}\{O \geq t, C \geq t\} && \textit{(At-risk process)} \\
N_E(t) &= \mathbf{1}\{E \leq t, E \leq C\} && \textit{(Excess uncensored deaths process)} \\
Y_E(t) &= \mathbf{1}\{E \geq t, C \geq t\} && \textit{(Excess at-risk process)}
\end{aligned}$$

Unfortunately, $(N_{E_i}, Y_{E_i})_{i \in 1,...,n}$ are not observable, but we show that

$$\begin{aligned}
\partial N_E(t) &= \frac{1}{a_t}\mathbb{E}(\partial N(t)|E, C) - \frac{b_t}{a_t c_t}\mathbb{E}(Y(t)|E, C), \\
Y_E(t) &= \frac{1}{c_t}\mathbb{E}(Y(t)|E, C),
\end{aligned}$$

where $a_t = \mathbb{P}(P \geq t | E = t)$, $b_t = \mathbb{P}(P = t | E \geq t)$ and $c_t = \mathbb{P}(P \geq t | E \geq t)$.
Assuming that (P, E) is an absolutely continuous random vector, a_t, b_t, c_t can be computed from partial derivatives $\mathcal{C}_i(\mathbf{u}) = \frac{\partial \mathcal{C}}{\partial u_i}(\mathbf{u}), i \in 1, 2$ of $\mathcal{C}$. Remark that under $(\mathcal{H}_\Pi)$, $a_t = c_t = S_P(t)$ and $b_t = -\partial S_P(t)$ do not depend on (unknown) $S_E(t)$, while they generally do under $(\mathcal{H}_\mathcal{C})$. However, like in classical survival analysis, we show under $(\mathcal{H}_\mathcal{C})$ that $\mathbb{E}(\partial N_E(t)) = \mathbb{E}(Y_E(t)\partial\Lambda_E(t))$, and moreover that $N_{E_i} - \int_0^t Y_{E_i}\partial\Lambda_{E_i}$ are $\mathcal{F}_t$-martingales. A natural estimator for $\partial\Lambda_E(t)$ can therefore be simply constructed as

$$\widehat{\partial\Lambda_E}(t) = \frac{\frac{1}{n}\sum_{i=1}^n \partial N_{E_i}(t)}{\frac{1}{n}\sum_{i=1}^n Y_{E_i}(t)}.$$

However, $(\partial N_{E_i}, Y_{E_i})$ are not directly observable and need to be estimated. For that, replace first unobservable conditional expectations by their stochastic counterpart: $\mathbb{E}\left(\partial N_i(t)|E_i, C_i\right)$ by $\partial N_i(t)$ and $\mathbb{E}\left(Y_i(t)|E_i, C_i\right)$ by $Y_i(t)$. This plug-in is enough to make the estimator computable under $(\mathcal{H}_\Pi)$ (it is the Pohar Perme estimator), but not under $(\mathcal{H}_\mathcal{C})$. We call *generalized Pohar Perme estimator* a solution of the differential equation:

$$\partial\widehat{\Lambda_E}(t) = \frac{\sum_{i=1}^n \frac{1}{\widehat{a_{i,t}}}\partial N_i(t) - \frac{\widehat{b_{i,t}}}{\widehat{a_{i,t}}\widehat{c_{i,t}}} Y_i(t)}{\sum_{i=1}^n \frac{1}{\widehat{c_{i,t}}} Y_i(t)}, \tag{1}$$

where for all $i \in 1, ..., n$,

$$\begin{aligned}
\widehat{a_{i,t}} &= \mathcal{C}_2\left(S_{P_i}(t), e^{-\widehat{\Lambda_E}(t)}\right), \\
\widehat{b_{i,t}} &= -\mathcal{C}_1\left(S_{P_i}(t), e^{-\widehat{\Lambda_E}(t)}\right)\partial S_{P_i}(t) e^{\widehat{\Lambda_E}(t)}, \\
\widehat{c_{i,t}} &= \mathcal{C}(S_{P_i}(t), e^{-\widehat{\Lambda_E}(t)}) e^{\widehat{\Lambda_E}(t)}.
\end{aligned}$$

Unfortunately, the differential equation 1 is now non-separable, and a non-linear equation in $\partial\widehat{\Lambda_E}(t)$ needs to be solved at each time step. Alike previous estimators under $(\mathcal{H}_\Pi)$, the obtained $\partial\widehat{\Lambda_E}$ process is piecewise continuous, with jumps at event times $T_1, ..., T_n$. It is moreover always negative, except at jump points, making the produced survival curve increasing between jumps. The solving scheme must therefore be performed on a very dense mesh $t_1, ..., t_N$ that includes observed times $T_1, ...T_n$. These characteristics were already present under $(\mathcal{H}_\Pi)$.

3 Illustration

We use data on colorectal cancer patients extensively described in Wolski & al (2020). Population is separated along the primary tumor location (left or right). Using several dependence structures, we obtain net survival curves from Figure 1. If Wolski et al. (2020) found the overall survival to be significantly different between left and right, net survival might not be if we take into account the uncertainty in $\mathcal{C}$. Further analysis to derive proper log-rank-type tests under $(\mathcal{H}_\mathcal{C})$ is possible.

References

Adatorwovor, R., Latouche, A. and Fine, J. P. (2023). A parametric approach to relaxing the independence assumption in relative survival analysis. *The International Journal of Biostatistics*, **18** , 577 – 592.

Czado, C. and Van Keilegom, I. (2023). Dependent censoring based on copulas. *Biometrika* **110**, 721,- 738.

Ederer, F. (1961). The relative survival rate: a statistical methodology. *JNCI Monographs*, **6**, 101 -- 121.

Hakulinen, T. (1982). Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, **38**, 933 – 942.

Nelsen, R. B. (2006). *An Introduction to Copulas.* Springer.

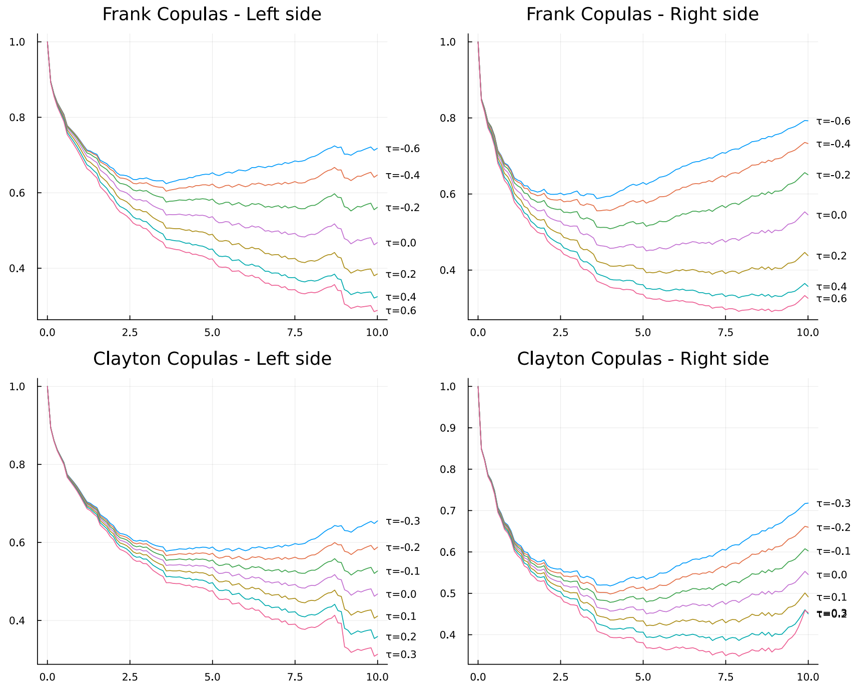


FIGURE 1. Obtained $\widehat{S_E}$ for several $(\mathcal{H}_\mathcal{C})$. Data was split w.r.t. tumor location (left or right) as in Wolski & al (2020), and several runs were done on several copulas: Frank copulas on the top and Clayton copulas on the bottom, with varying parameters θ. In both cases, $\theta = 0 \iff \mathcal{C} = \Pi$, and this curve represents the Pohar Perme estimator.

Pohar Perme, M., Stare, J. and Estève, J. (2012). On estimation in relative survival. *Biometrics*, **68**, 113 – 120

Wolski, A., Grafféo, N., Giorgi, R. and the CENSUR working survival group (2020). A permutation test based on the restricted mean survival time for comparison of net survival distributions in non-proportional excess hazard settings. *Statistical Methods in Medical Research*, **29**, 1612 --1 623

Context matters: including global covariates in relational event models

Melania Lembo[1], Rūta Juozaitienė[2], Veronica Vinciotti[3], Ernst C. Wit[1]

[1] Università della Svizzera italiana, Switzerland
[2] Vytautas Magnus University, Lithuania
[3] University of Trento, Italy

E-mail for correspondence: melania.lembo@usi.ch

Abstract: Relational event models have become popular in the network science literature, as a number of phenomena in various applied fields, such as sociology, ecology and finance, can be described via a network of entities interacting over time. A relational event model allows to describe the formation of instantaneous links over time and to identify its driving factors. Traditional inferential techniques, involving Cox's partial likelihood, can estimate the effects of covariates that are node-specific, such as age or in-degree, or dyadic, such age difference of pairs of nodes or reciprocity. However, the partial likelihood cannot account for global covariates, i.e., factors that are constant for all pairs. Indeed, these covariates, being only time-dependent, drop out from the partial likelihood. Nevertheless, these factors, such as weather or time of the day, are often important in capturing and explaining the temporal nature of the studied events. In this paper, we address this challenge with the use of nested case-control sampling on a time-shifted version of the event process. This will result in a partial likelihood of a degenerate logistic generalized additive model from which we are able to recover effects of all kinds of covariates, including global ones.

Keywords: Relational event model; Generalized additive model; Partial likelihood; Risk set sampling; Dynamic network.

1 Relational event models with global covariates

In a relational event model, the events are *directed interactions* $(s \to r)$ between a *sender* s and a *receiver* r, occurring at specific points in time (Bianchi et al. (2024)). In order to define the model, consider a multivariate counting process $\mathbf{N} = \{N_{sr}\}_{(s,r)\in\mathcal{S}\times\mathcal{C}}$, where $N_{sr}(t)$ counts the number of occurrences of interaction (s, r) in $[0, t]$, and denote with $\mathcal{H}_t$ the history of the process. The underlying dynamics and driving factors of such a process are modelled through its stochastic intensity process λ_{sr}, intuitively representing the probability of an event occurring in an infinitesimal time interval $[t, t+\mathrm{d}t)$ conditioned on everything that has been observed prior to t. This is modelled via a Cox proportional hazards model

$$\lambda_{sr}(t|\mathcal{H}_{t-}) = Y_{sr}(t)\lambda_0(t)\text{exp}\left\{\sum_{l=1}^{q} f_l\left(\mathbf{x}_{sr}^{(l)}(t)\right) + \sum_{l=1}^{w} g_l\left(\mathbf{x}^{(h)}(t)\right)\right\}, \qquad (1)$$

with $\lambda_0(t)$ an arbitrary non-negative baseline hazard and $Y_{sr}(t)$ a binary indicator of whether the pair (s, r) is at risk of occurring at time t or not.

The parametric part of the model describes the effect of covariates on the formation of links over time. This includes the more traditional node-specific or edge specific covariates, denoted with $\mathbf{x}_{sr}^{(l)}(t)$ and whose effect is expressed through the arbitrary functions f_l, as well as global covariates, which are constant for all nodes and pairs. The latter are denoted with $\mathbf{x}^{(h)}(t)$ and associated to arbitrary functions g_l. The functions f_l or g_l can be taken either as linear or smooth non-linear functions of the covariates.

2 Nested case-control sampling on a shifted process

The inclusion of global covariates in the model (1) adds an additional layer of complexity in the estimation of their effects. Indeed, the terms involving these variables, being only time-dependent, drop out from the traditional partial likelihood (Cox (1975)), as they cancel out in the multinomial probabilities of observing a specific interaction against pairs of all interactions in the risk set at a same time t. The same applies to the more efficient extensions of these methods based on nested case-control sampling (Borgan et al. (1995)), where a certain number of non-events is uniformly sampled from the risk set at a specific event time. In this paper we propose a time-shifted version of the original counting process $\mathbf{N}$, from which we are able to estimate also the effects of global covariates.

To this end, let $\mathbf{T} = \{T_{sr}\}_{(s,r)\in\mathcal{S}\times\mathcal{C}}$ be a process such that $T_{sr} \subset [0, \tau]$ is the countable set of event times corresponding to the interaction $(s \to r)$. Consider then another process $\mathbf{H} = \{H_{sr}\}_{(s,r)\in\mathcal{S}\times\mathcal{C}}$, independent of $\mathbf{T}$ with $H_{sr} \geq 0$. We shift the event times of each interaction by the value of $\mathbf{H}$ for the corresponding pair. Thus, we define $T^e = \{H_{sr} + T_{sr,k} \mid \forall(s, r) \in \mathcal{S} \times \mathcal{C};\ k \geq 1\}$ as the set of shifted event times, where $T_{sr,k} \in T_{sr}$ is the k-th occurrence of the dyad (s, r). In this way, we obtain a shifted marked point process $M^e = \{(T_j^e, s_j, r_j)\}_{j\geq1}$ where T_j^e is the j-th order statistic of T^e and the mark (s_j, r_j) represents the specific interaction that occurred at that time. The risk set composition for M^e is inherited from $\mathbf{N}$ according to an indicator process $Y_{sr}^e(t)$, that is equal to $Y_{sr}(t - H_{sr})$ for $t \in [H_{sr}, H_{sr} + \tau]$ and 0 outside this time interval.

We propose to perform nested case-control sampling on this shifted process. In particular, at each shifted event time T_j^e, we sample a non-event pair $(s_j^*, r_j^*) \neq (s_j, r_j)$ among the pairs such that $Y_{sr}^e(T_j^e) = 1$. Then, having observed $\mathbf{H}$ and using the independence between $\mathbf{H}$ and $\mathbf{T}$, the probability of (s_j, r_j) occurring, given that there is an event at T_j^e and that either (s_j, r_j) or (s_j^*, r_j^*) could have happened, follows a Binomial distribution that depends on the original intensity process λ_{sr} evaluated at $T_j^e - H_{sr}$. Considering a realization of n events for the process M^e, the resulting partial likelihood is therefore given by

$$\mathcal{L}(\lambda_0, \mathbf{f}, \mathbf{g}) = \prod_{j=1}^{n} \frac{\lambda_{s_j r_j}(t_j^e - h_{s_j r_j})}{\lambda_{s_j r_j}(t_j^e - h_{s_j r_j}) + \lambda_{s_j^* r_j^*}(t_j^e - h_{s_j^* r_j^*})}, \tag{2}$$

with λ_0, $\mathbf{f}$ and $\mathbf{g}$ the baseline hazard, dyadic and global functions in (1), respectively. Crucially, the global terms $\mathbf{g}$ and λ_0 do not cancel out in (2), as the event and the sampled non-event will have received different shifts.
It can be shown that (2) is the likelihood of a degenerate logistic generalized additive model with covariates given by the difference between the ones of the event and the corresponding sampled non-event, respectively. We can then perform estimation using existing techniques for this class of models (Wood (2017)), based on a spline approximation for the smooth terms and an optimization of a penalized version of (2). These are implemented in the `R` package `mgcv` and return estimates of all effects, including those from global covariates.

3 Bike sharing in Washington D.C.

We aim to use the proposed methodology on bike sharing data from Washington D.C., collected over the course of July 2023 and available at `https://www.capitalbikeshare.com/system-data`. The goal is to investigate how global covariates, such as temperature or precipitation, as well as node-specific or dyadic covariates, such as the distance between stations, affect the rate of bike sharing in D.C.

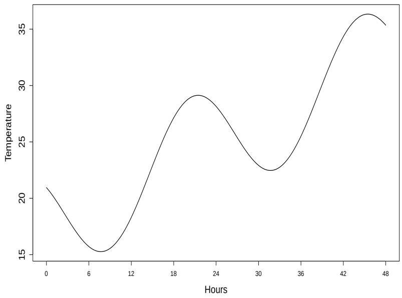


FIGURE 1. Temperature over time, $T(t) = 0.3(t+3\pi)+5\sin(\frac{\pi}{12}(t+3\pi))+15$

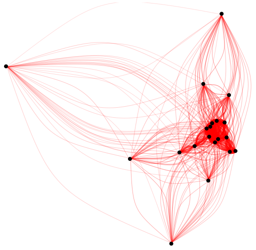


FIGURE 2. Closer stations experience a larger number of simulated events.

In order to evaluate the proposed method, we consider a simple simulation study, which mimics the setting just described. In particular, we generate data on bike shares across two days among 20 bike stations selected from the real data. We consider a rate of events that depends on the temperature, simulated as in Figure 1 and assumed to have a smooth quadratic effect on bike riding, with 23°C as the ideal temperature to ride. We also assume that the distance between stations has a negative fixed effect (equal to -0.5) on the rate. Figure 2 shows that close-by stations (represented through relative position of the nodes in the figure) are associated to a higher number of simulated events among them (wider and darker

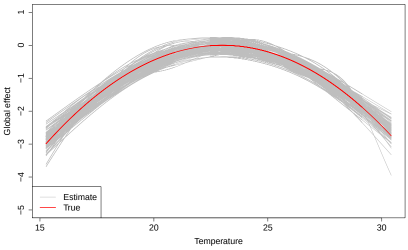


FIGURE 3. Temperature smooth effect.

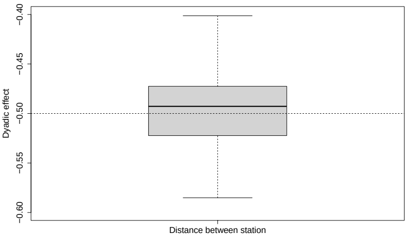


FIGURE 4. Distance fixed effect.

links). We generate shifts from an exponential distribution of mean equal to the average event time of the simulated process and apply our proposed method. The results, based on 100 replications, show that the approach correctly recovers both the smooth global effect associated with temperature (Figure 3) and the dyadic fixed effect of the distance (Figure 4). Given the effectiveness shown by this small example, we will next analyze the full dataset on bike sharing.

References

Bianchi, F., Filippi-Mazzola, E., Lomi, A. and Wit, E.C. (2024). Relational event modeling. *Annual Review of Statistics and Its Application*, **11**, 297 – 319.

Borgan, Ø., Goldstein, L. and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics*, **23**, 1749 – 1778.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269 – 276.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition (2nd ed.)*. Boca Raton: Chapman and Hall/CRC.

Sparse intrinsic Gaussian processes for prediction on manifolds: extending applications to environmental contexts

Yuan Liu[1], Mu Niu[1], Claire Miller[1]

[1] School of Mathematics & Statistics, University of Glasgow, United Kingdom

E-mail for correspondence: y.liu.10@research.gla.ac.uk

Abstract: Traditional Gaussian Processes are limited in their application by complex boundaries and intricately structured manifolds, such as when predicting water quality in the Aral Sea. Intrinsic Gaussian Processes adequately accommodate these complex conditions. To address the computational complexity of Intrinsic Gaussian Processes, we employ the sparse approximation method known as Deterministic Inducing Conditionals (DIC). The Sparse Intrinsic Gaussian Processes approach we propose offers effective prediction over such manifolds.

Keywords: Sparse intrinsic Gaussian processes; Manifold; Deterministic inducing conditionals.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

Number of jobs during working life based on SHARE data

Ivana Malá[1]

[1] Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, Czech Republic

E-mail for correspondence: malai@vse.cz

Abstract: The number of jobs (as well as the number of careers) held up in career history is an important feature of people's labour market behaviour, reflecting not only the economic situation and the attitudes of the population. At the same time, it is changing very rapidly, with a significant increase in the number of jobs over time and a successive decrease in the length of time spent in one job. This paper examines the work history of a generation of European residents up to the year of birth 1967, using data from the Job Episode Panel based on the SHARE survey project for the modelling. Because of the data, work history to the age of 50 is of interest. The aim is to assess the development of the number of jobs over time and the convergence of values for EU members with and without a communist history. The target population was aged over 23 in 1990. Mixtures of two and three Poisson regressions were used to assess the effect of characteristics on the target variable, and AIC was used to compare the models. For artificial components found, the estimated (joint) membership tables are presented for models with and without interactions and for two and three components are presented.

Keywords: Job episodes panel; SHARE; Finite mixture of GLM; Poisson regression.

1 Introduction

The European Survey of Health, Ageing and Retirement in Europe provides a huge database of data connected to the European population aged over 50 (SHARE, 2023). The retrospective Job Episodes Panel (Brugiavini & all., 2019) based on this project describes the whole life and working history of respondents from their birth through education and active economic life to retirement to end of life in a yearly panel data. The last year of observation is 2017 (for those alive and not lost) or year of death. Respondents were at least 50 years old in 2017, so their year of birth is 1967 or earlier. The aim is to model the number of jobs

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

up to the age of 50 and to assess the differences between the original European Union countries and European Union members with experience of communist regimes. Information on gender, education (described by the number of years in school) and children raised are used as additional independent variables. In order to assess the convergence of the behaviour of respondents from both groups of countries over time, a variable was constructed indicating the decade in which the respondent turned 50.

The Poisson GLM regression is used to model the dependence between the number of jobs (not careers) and covariances. Our modelling aim is to find groups of observations with similar regression coefficients and to identify more homogeneous subpopulations in the populations of interest - two subgroups of European countries. In the model, not only main effects are included, but also interactions are added to test for differences in the impact of independent variables on various subpopulations.

2 Data, models and results

We have 5,813,133 years of life of respondents from all EU countries (excl. Malta and Cyprus) and Switzerland in the database. This means 85,935 unique respondents (44 % men and 56 % women), 36,046 still economically active respondents (42 %) and 49,889 retired respondents (58 %).

Variables included in covariates $\mathbf{x}$ in the model ((1) and (2), base categories in italics):

- $gender, x_1$ (*male*, female) $years\,in\,school, x_2$
- $oldEUcountry, x_3$ (*yes*, no) $children, x_4$ (*no*, yes)
- $year\,group\,50, x_5$ (1980-, 1981-1990, 1991-2000, *2001-2010*, 2011+)

The target variable Y is a number of jobs up to the age of 50. Only one job during the study period is the mode for the empirical distributions of all populations studied with mean number being between 2 and 4 jobs. The subpopulations differ in the probability that the respondent has never worked (as expected, gender is a strong predictor of this value, as is the year of birth). The maximum value in the sample is 17 occupations. For such a data, normal regression model is not suitable, for this reason, the Poisson regression is applied using a large spectrum of methods and implementations for GLM models.

The Poisson GLM models (model 1 with (two-way) interactions for x_1, x_3, x_5 and a nested model 2 without interactions, where only main effects of x_1 to x_5 are included) are of the form:

$$E(Y|\mathbf{x}) = exp\left\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right\}, Var(Y|\mathbf{x}) = exp\left\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right\}. \tag{1}$$

We further hypothesize that there are two or more (K) distinct, artificial unobserved groups of respondents in the population that differ in their behaviour in the labour market. Moreover, this approach enables us to fix problems with overdispersion; zero value of our target variable refers to those, who have no job in their life history.

Finite mixture model of K Poisson regressions (models (1) fitted to a selected groups by the EM algorithm, see Faria, Rodrigues Gonçalves, 2013, Grün, Leisch, 2008)

$$P(y;\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K,\pi_1,...,\pi_{K-1}) = \sum_{j=1}^{K} \pi_j P_j(y;\boldsymbol{\beta}_j), y = 0,1,2,..., \qquad (2)$$

where K is the number of components (K=2 and 3 in this text), $P_j\ j = 1,2,...,K$ are Poisson probability functions from component Poisson regression models, $\boldsymbol{\beta}_j$ are parameters from (1), $0 < \pi_j < 1$, and $\sum_{j=1}^{K} \pi_j = 1$. Package Flexmix in the freeware R (R Core Team, 2021) was used, see (Grün, Leisch, 2008). Three components seem to be a sufficient number, the analysis of the number of components was unable to find larger number of identifiable components. Our task is to find out homogenous subgroups, in this case large number of components is not interpretable.

The AIC criterion was used to compare models. Due to large sample, all test are very strict and almost all individual test of parameters are significant and even a small decline in AIC is tested to be significant. For this reason, we prefer just AIC for a descriptive comparison. We present AIC values of all models in Table 1 illustrating the decrease in AIC in both models and K =1, 2, 3.

We applied only models with up to 3 components, in case of more components a too small components were found with weights under 1% and there have been problems with identifiability of parameters. Estimated cluster membership (together with component weights) for both models (1 and 2) (Table 2) and 2 and 3 selected components (Table 3) are given to look whether the hidden subpopulations are present, stable using impact of our covariates and similarly identified by both models. We order components according to their estimated weights.

In order to estimate standard errors of estimated parameters, resampling was applied (procedure included in the package Flexmix).

TABLE 1. Values of AIC criterion.

model 1	$k = 2$	$k = 3$	model 2	$k = 2$	$k = 3$
319,294	315,649	314,138	319,813	314,800	314,720

TABLE 2. Distribution of respondents in artificial components. Impact of interactions in the model, comparison of results for model 1 and model 2.

models 1/ 2		1	2		1	2	3
	weight	0.887	0.113	weight	0.470	0.445	0.085
1	0.828	79 811	1 298	0.779	51 181	29 284	589
2	0.172	1 841	2 985	0.153	1 039	2	1 700
3	-	-	-	0.068	2	8	2 130

In the case of the two components, both models found similar components in terms of estimated weights and (estimated) component membership. For the three component models, the components found by this model have been approximately

TABLE 3. Distribution of respondents in artificial components. Joint membership for 2 and 3 components.

	model 1	1	2	3	model 2	1	2	3
	weight	0.779	0.153	0.068	weight	0.470	0.445	0.085
1	0.828	79600	1033	476	0.887	51 986	29 294	372
2	0.172	1454	1708	1664	0.113	236	0	4 407

distributed to those in two components. If there are really unobserved homogeneous groups in the population, both models should approximately find them. We consider the similarity in classification (not all analyses included in this text) acceptable. By analysing the members of the subpopulations concerning country, gender and birth group, we can look at the representation of the observed subpopulations. We compared a structure for categorical variables (main effects only) and means for the quantitative variable years in school and signs of the corresponding parameters.

In the main effects model 2, there are opposite signs in the estimated parameter for years in school, resulting in different subgroup means of this variable and an opposite direction of relationship. There is a similar difference in the dependence on having children, with a higher representation of childless respondents than in the whole dataset in the major component. In the model with three components, the direction of the partial relationship differs also for gender.

In the complex model 1 with interactions, the opposite sign for the parameters is again observed for years in school and children, with more differences appearing for the effect of birth category. The differences are reflected in the parameter estimates for the interactions.

3 Conclusion

The Poisson regression model is an acceptable description of the behaviour of the respondents, the addition of the hidden component provides a decrease in the AIC and an interesting look at the estimated component membership representing homogeneous unobserved subgroups of respondents. It also allows a discussion of the differences in the estimated parameters (their sign and magnitude) to identify the source of the differences.

We can expect complete data on the respondents in the study within the next ten or fifteen years. However, from the construction of samples at each wave, additional samples are drawn to maintain or increase the size of the samples. To model the number of jobs at retirement for the generation of EU residents born before 1967 based on our data, it is possible to use a Poisson regression model estimated from right-censored data.

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.

This paper uses data from the generated Job Episodes Panel (DOI: 10.6103/ SHARE.jep.600), see Brugiavini et al. (2013) and Antonova et al. (2014) for methodological details. The Job Episodes Panel release 6.0.0 is based on SHARE Waves 1, 2 and 3 (SHARELIFE) (DOIs: 10.6103/SHARE.w1.600, 10.6103/SHARE.w2.600, 10.6103/SHARE.w3.600).

References

Antonova, L., Aranda, L., Pasini, G. and Trevisan, E. (2014). Migration, family history and pension: the second release of the SHARE Job Episodes Panel. Working paper 18

Brugiavini, A., Orso, C.E., Genie, M.G., Naci, R. and Pasini, G. (2019). Combining the retrospective interviews of wave 3 and wave 7: the third release of the SHARE Job Episodes Panel1. Working Paper 36.

Grün, B. and Leisch, F. (2008). FlexMix Version 2: Finite mixtures with concomitant variables and varying and constant Parameters. *Journal of Statistical Software*, **28**, 1 − 35

SHARE (2023). SHARE. The largest European social science panel study. [Accessed 01.03.2023]. Available: https://share-eric.eu/

Faria, S. and Rodrigues Gonçalves, F. I. (2013). Financial data modeling by Poisson mixture regression, *Journal of Applied Statistics*, **48**, 2150 – 2162

R Core Team (2021). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing,Vienna, Austria,
https://www.R-project.org/.

Modelling age-space mortality dynamics in small areas

Jacob Martin[1,2], Carlo Giovanni Camarda[1]

[1] Institut national d'études démographiques, Aubervilliers, France
[2] Universidad del País Vasco, Bilbao, Spain

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

Abstract: Large stochastic fluctuations due to small death counts limit understanding the true extent of geographic differences in mortality between small areas. We propose a model that borrows strength across age and space to produce reasonable estimates of the age-specific mortality schedule in small populations. We apply our model to mortality data for 50 Spanish provinces.

Keywords: Age-space interaction; Mortality modeling; Small area estimation; Smoothing.

1 Motivation

The quest to understand fine-grain mortality inequalities has spurred the creation of models tailored for small areas. The challenge lies in estimating age-specific death risks in small populations, given limited observed deaths leading to large fluctuations in mortality rates, primarily due to stochastic factors rather than substantial differences in mortality conditions. Despite this, genuine public health differences, such as exposure to environmental hazards, may exist. While existing models use prior knowledge, they grapple with issues like selecting a standard schedule, neglecting uncertainty in its choice, and not fully leveraging the spatial structure of the data (see, Gonzaga and Schmertmann, 2016; Alexander et al. 2017). In response, our proposed data-driven model for small areas avoids external standards, incorporating information across ages and space. This approach results in a properly specified stochastic model, featuring a nonparametrically estimated standard schedule based on the data and incorporating details about the total population.

To illustrate our results, we took age-specific death and popoulation data for 50 provinces of Spain (excluding the Canary Islands) from the Instituto Nacional de Estadística (INE). We have deaths by single year of age up to age 110 and calculated population exposures by single year of age from the death and population data for each province. We apply our model to males in the year 2019.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

2 The age-space model

Data for our model consists of two $m \times n$ matrices, $\mathbf{Y}$ (deaths) and $\mathbf{E}$ (exposures), over age i and region j. To incorporate structure in the age-pattern, we include total population data as the first region (resulting in $n+1$ regions, adding the sum of all regions as an additional row to $\mathbf{Y}$ and $\mathbf{E}$). Spatial information is based on the centroids of each territorial unit. We assume that y_{ij} is Poisson distributed with expectation $\mu_{ij}e_{ij}$, where μ_{ij} is the force of mortality we aim to model in a log-scale.

We model each region as the sum of three components: a smooth standard age-pattern ($\boldsymbol{\eta}^0$), a spatially varying smooth age deviation ($\boldsymbol{\delta}_j$), and a scalar allowing for regional differences in mortality level that is not smoothed spatially (γ_j). To ensure convergence and allow interpretability of each component, the standard $\boldsymbol{\eta}^0$ is estimated from the mortality of the total population, and we add a ridge penalty to the γ_j. Vectorizing the linear predictor our model is expressed as:

$$\ln(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\,\boldsymbol{\theta} = \left[\begin{array}{c|cc} \mathbf{1}_{n+1} \otimes \mathbf{B}_a & \multicolumn{2}{c}{\mathbf{0}_{m,k_{\text{as}}+n}} \\ \cline{2-3} & \mathbf{B}_s \otimes \mathbf{B}_a & \multicolumn{1}{|c}{\mathbf{I}_n \otimes \mathbf{1}_m} \end{array}\right] \boldsymbol{\theta}\,, \tag{1}$$

where $\mathbf{B}_a$ is a $m \times k_a$ B-spline basis over age modified in order to take into account the sharp descent of the level of mortality after the first year of life (Camarda, 2019). The matrix $\mathbf{B}_s$ captures the spatial dimension, constructed as the row tensor product of bases for each individual spatial dimension. Let $\mathbf{B}_{\text{lon}}$ be a $n \times k_{\text{lon}}$ B-spline basis over the longitude coordinates of the centroids, and $\mathbf{B}_{\text{lat}}$ be a $n \times k_{\text{lat}}$ basis over the latitude coordinates. Then $\mathbf{B}_s$ is the $n \times k_{\text{lat}}k_{\text{lon}}$ basis given by

$$\mathbf{B}_s = \mathbf{B}_{\text{lat}} \Box \mathbf{B}_{\text{lon}} = (\mathbf{B}_{\text{lat}} \otimes \mathbf{1}^{\mathrm{T}}_{k_{\text{lon}}}) \odot (\mathbf{1}^{\mathrm{T}}_{k_{\text{lat}}} \otimes \mathbf{B}_{\text{lon}}). \tag{2}$$

Space interacts with ages by creating the $mn \times k_{\text{as}}$ design matrix $\mathbf{B}_s \otimes \mathbf{B}_a$. The supplementary lower-right corner of matrix $\mathbf{X}$ corresponds to the regional mortality level γ_j (the identity matrix is present since we do not consider spatial structure here). The zeros in the top right of the matrix are present because the total population has no spatial or region-specific components.

The coefficients vector $\boldsymbol{\theta}$ can be seen as the combination of three sets of coefficients, denoted as $\boldsymbol{\theta} = [\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}]^{\mathrm{T}}$, each with lengths k_a, k_{as}, and n, respectively. As per (1), the mortality age-pattern for the total population serves as the standard for each region and it is concurrently estimated within the same framework: $\boldsymbol{\eta}^0 = \mathbf{B}_a\boldsymbol{\alpha}$. The coefficients $\boldsymbol{\beta}$ are then estimated to characterize age-space interaction, specifically, the age-regional specific deviation $\boldsymbol{\delta}_j$ that varies across space.

Following a P-splines approach, we ensure smoothness across age and space by introducing a discrete penalty $\mathbf{P}$ to the likelihood associated with (1), that is difference penalties on the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and a ridge penalty on $\boldsymbol{\gamma}$.

We estimate the model using Iteratively Reweighted Least Squares within a GLAM framework (Currie et al., 2006). Smoothing parameters associated with the standard $\boldsymbol{\eta}^0$, each $\boldsymbol{\delta}_j$, the smooth variations of these deviations across space, and the magnitude of the ridge penalty are optimized by minimizing the Bayesian Information Criterion. Once the model is estimated, variance-covariance matrix of the parameters can be evaluated, allowing us to quantify uncertainty regarding each model component and the fitted age-regional-specific mortality patterns.

3 Application to Spanish provinces

Figure 1 shows fitted mortality rates and observed counts for Soria and Madrid provinces. We can observe that for the large population of Madrid where there are very few counts the model fits closely the data, while for the small province of Soria where there are zero deaths at almost every age before 30, the model is able to reconstruct a reasonable mortality schedule at all ages. The confidence intervals around the estimated rates do not contain most of the observed mortality rates, especially for the smaller province of Soria. However, in contrast to prediction intervals, we do not expect the observed rates to fall inside these confidence bands. Figure 2 shows the values for $\boldsymbol{\delta}$ averaged over age, and the values for $\boldsymbol{\gamma}$ by province. We can observe that $\boldsymbol{\delta}$ captures the spatial nature of mortality, while the $\boldsymbol{\gamma}$ allow for smaller deviations that avoid imposing too rigid a smooth structure.

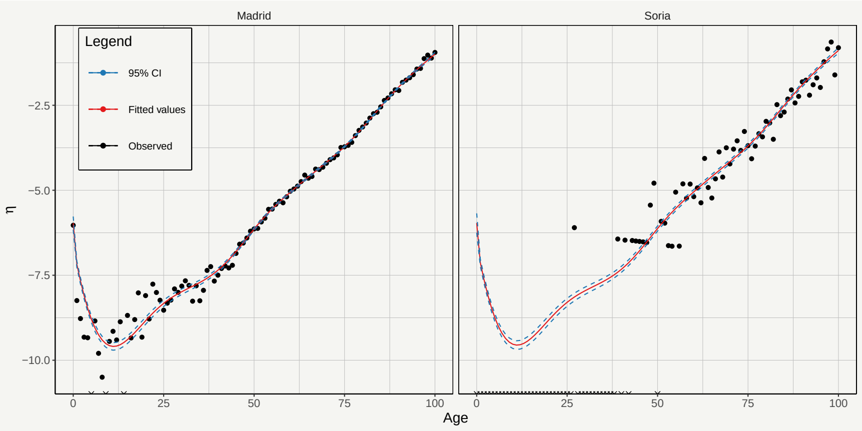


FIGURE 1. Observed log-death rates and fitted values for Madrid and Soria.

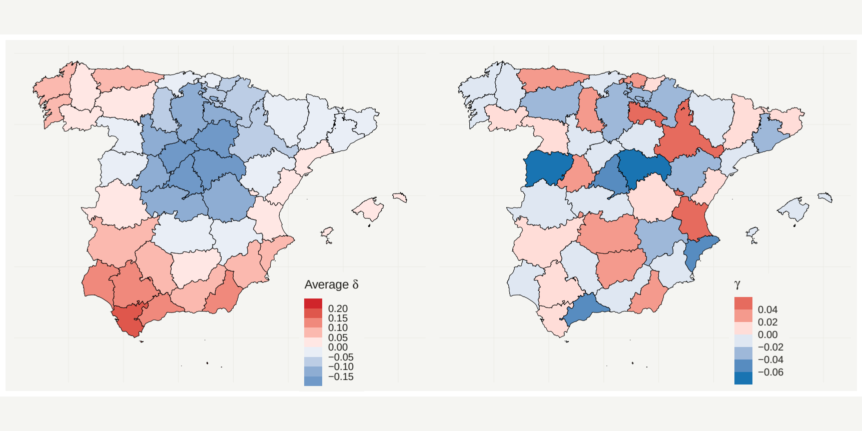


FIGURE 2. Average values of $\boldsymbol{\delta}$ and values of $\boldsymbol{\gamma}$ by province.

4 Conclusion

We introduce a model that expresses regional mortality as the sum of a standard, age-specific deviations smooth in space, and unsmooth regional effects. Our model borrows strength across age and space, with the flexibility that allows for breaks from a perfectly smooth spatial pattern of mortality. We illustrate our model with an application to provincial data, but our model could be also used at the municipal or sub-municipal level.

Further improvements could include adding hierarchical structure to the model, as small areas are often embedded in larger territorial units, and modeling overdispersion. We also plan to embed our model in a Composite Link Model framework, since mortality data for small areas often comes in coarse age groups.

References

Alexander, M., Zagheni, E. and Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography*, **54**, 2025 – 2041.

Camarda, C.G. (2019). Smoothed constrained mortality forecasting. *Demographic Research*, **41**, 1091 – 1130.

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

Lee, D. and Durbán, M. (2011). *P*-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling*, **11**, 49 – 69

Schmertmann, C. and Gonzaga, M. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*. **55**. 1363 – 1388.

Uncorrelatedness in high-dimension: hypothesis testing of independence in presence of outliers

Nirian Martín[1], J. Miguel Marín[2]

[1] Complutense University of Madrid, Madrid, Spain
[2] Carlos III University of Madrid, Getafe, Spain

E-mail for correspondence: nirian@mat.ucm.es

Abstract: Rao's score tests, although important member of classical statistical testing procedures alongside likelihood ratio tests (LRT), are not commonly associated with multivariate and high-dimensional contexts. This paper presents a Rao's score test designed for testing a fixed correlation matrix, with a particular focus on complete independence in normal data scenarios. An expression of the test statistic and the corresponding (n, p)-asymptotic distribution is derived that aligns with Schott's (2005) results in high-dimensional contexts, albeit through a distinct method. The proposed Rao's test-statistic, while primarily intended for normally distributed data, is extended to elliptical distributions. A simulation study is conducted using observations following a normal scale mixture distribution (Muirhead and Waternaux, 1980) and comparing Rao's score test and LRT with different scenarios of sample and dimension sizes, n and p. We assess the performance of the classical Rao's score test for correlation matrices against the traditional LRT (for scenarios where $p < n$) and other competitive tests in high-dimensional environments (where $p \geq n$). This study confirms that the classical Rao's score test is effective not only under the usual dimensional constraints of $p < n$, but also in the more complex high-dimensional context where $p \geq n$.

Keywords: Rao's score test; High-dimensional data; Correlation matrix.

1 Introduction

We focus initially on a multivariate normal distribution, $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where p indicates the dimension for a positive definite matrix $\boldsymbol{\Sigma}$, but the methodology is later adapted to elliptical distributions. Denoting $\boldsymbol{R} = \text{diag}^{-\frac{1}{2}}\{\boldsymbol{\Sigma}\}\boldsymbol{\Sigma}\text{diag}^{-\frac{1}{2}}\{\boldsymbol{\Sigma}\}$, to study complete independence of p normal variables, i.e. for testing

$$H_0 : \ \boldsymbol{R} = \boldsymbol{I}_p \text{ versus } H_1 : \ \boldsymbol{R} \neq \boldsymbol{I}_p, \tag{1}$$

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

we count with a random sample of n individuals, $\boldsymbol{X}_h$, for $h = 1, \ldots, n$. The traditionally mostly used LRT is only valid as a n-asymptotical result when the dimension p is smaller than the sample size n. When p is greater than or equal to n, requiring an (n, p) asymptotical test-statistic, the $p \times p$ sample covariance matrix, represented by $\boldsymbol{S}_{n,p} = \frac{1}{n}\sum_{h=1}^{n}(\boldsymbol{X}_h - \bar{\boldsymbol{X}}_{n,p})(\boldsymbol{X}_h - \bar{\boldsymbol{X}}_{n,p})^\top$, with $\bar{\boldsymbol{X}}_{n,p} = \frac{1}{n}\sum_{h=1}^{n}\boldsymbol{X}_h$, is singular and this fact makes useless most of the LRT related methods. In particular, the traditional LRT for complete independence, (1), becomes invalid. This paper addresses these issues by demonstrating the applicability of Rao's score tests. It is shown that the test of complete independence, as proposed by Schott (2005), is in essence a Rao's score test. We point out that a modification of this test makes also possible to create a test for absence of correlation in elliptical distributions.

The structure of this work begins with the presentation of results pertaining to classical LRTs in Section 2. This is followed by the delineation of the paper's two primary objectives. The initial objective, elaborated in Section 3, is to illustrate the use of Rao's score test for the multivariate normal distribution in assessing a fixed correlation structure when variances and means are not known. This methodology is applicable in both traditional ($p < n$) and high-dimensional ($p \geq n$) settings, given that the prerequisites for employing Rao's score test for the multivariate normal distribution are met. The subsequent objective, outlined in Section 4, seeks to broaden the previous one from multivariate normal to multivariate elliptical distribution. Lastly, a simulation study and real data application is discussed in Section 5.

2 Difficulty of the LRT in high dimensional setting

It is well-known that the n-asymptotic distribution of the LRT for testing (1), valid for $p < n$, is given under H_0 by

$$-2\log\lambda_p(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = -n\log(\det(\boldsymbol{R}_{n,p})) \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \chi^2_{d(p)}, \tag{2}$$

where $d(p) = \frac{1}{2}p(p-1)$ and

$$\boldsymbol{R}_{n,p} = \mathrm{diag}^{-\frac{1}{2}}\{\boldsymbol{S}_{n,p}\}\boldsymbol{S}_{n,p}\mathrm{diag}^{-\frac{1}{2}}\{\boldsymbol{S}_{n,p}\}.$$

The expression of the LRT is conformed as

$$\lambda_p(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = \frac{\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\Theta_0}\mathcal{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sup_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\Theta}\mathcal{L}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma})},$$

being Θ_0 the parameter space under H_0 and Θ the whole parameter space. If $p \geq n$ then $\det(\boldsymbol{S}_{n,p}) = 0$ (a.s.), which means that the MLE $(\bar{\boldsymbol{X}}_{n,p}, \boldsymbol{S}_{n,p})$ is not part (a.s.) of Θ compound by $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ being $\boldsymbol{\Sigma}$ any $p \times p$ positive definite matrix. Consequently, for $p \geq n$, the LRT is inapplicable for conducting the test given in (1).

3 Classical Rao's score test in high dimension

Eliminating all strictly supradiagonal elements of $\boldsymbol{\Sigma}$, i.e., the redundant elements, $\mathrm{vech}(\boldsymbol{\Sigma})$ denotes the $\frac{1}{2}p(p+1)$-th order vector that is obtained by stacking the columns one underneath the other. Let $\boldsymbol{G}_p$ denote the so-called duplication matrix. For details about the explicit expression of $\boldsymbol{G}_p$ and its properties, the reader is addressed to Magnus (1988, Chapter 4) and references therein. From McCulloch (1982), the score function with respect to $\boldsymbol{\theta} = (\boldsymbol{\mu}^\top, \mathrm{vech}^\top(\boldsymbol{\Sigma}))^\top$ is

$$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\partial \log f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ \frac{1}{2}\boldsymbol{G}_p^\top \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right)\left[(\boldsymbol{x}-\boldsymbol{\mu}) \otimes (\boldsymbol{x}-\boldsymbol{\mu}) - \mathrm{vec}\,(\boldsymbol{\Sigma})\right] \end{bmatrix},$$

and the Fisher information matrix,

$$\boldsymbol{I}_F(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X})\boldsymbol{s}_{\boldsymbol{\theta}}^\top(\boldsymbol{X})] = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & \boldsymbol{0}_{p\times\frac{p(p+1)}{2}} \\ \boldsymbol{0}_{\frac{p(p+1)}{2}\times p} & \frac{1}{2}\boldsymbol{G}_p^\top\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{G}_p \end{pmatrix}.$$

The estimating equations are given by $\boldsymbol{\theta}$ under the whole parametric space, are given by

$$\boldsymbol{U}_{n,p}\left(\boldsymbol{\theta}\right) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\left(\bar{\boldsymbol{X}}_{n,p}-\boldsymbol{\mu}\right) \\ \frac{1}{2}\boldsymbol{G}_p^\top\left(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{G}_p\mathrm{vech}(\boldsymbol{S}_{n,p}(\boldsymbol{\mu})-\boldsymbol{\Sigma}) \end{bmatrix}$$

It can be proven that the Rao's score (or Lagrange multipliers) test statistic for

$$H_0:\ \boldsymbol{R}=\boldsymbol{R}_0 \quad \text{vs.} \quad H_1:\ \boldsymbol{R}\neq\boldsymbol{R}_0, \tag{3}$$

is given by

$$L_{n,p}(\tilde{\boldsymbol{\theta}}) = n\boldsymbol{U}_{n,p}^\top(\tilde{\boldsymbol{\theta}})\boldsymbol{I}_F^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{U}_{n,p}(\tilde{\boldsymbol{\theta}}) = \frac{n}{2}\mathrm{tr}\left(\left(\boldsymbol{R}_0^{-1}\boldsymbol{R}_{n,p}-\boldsymbol{I}_p\right)^2\right), \tag{4}$$

with $\boldsymbol{R}_0$ being a completely known correlation matrix, $\tilde{\boldsymbol{\theta}}$ is denoting the estimator of $\boldsymbol{\theta} \in \Theta_0$, i.e. under H_0. Its n-asymptotic distribution, for a fixed value of p, is $\chi^2_{d(p)}$, with $d(p) = \frac{1}{2}p(p-1)$. In addition, if we assume that both, n and p increase in such a way that $\frac{p}{n-1}$ tends to a fixed $\gamma \in (0,+\infty)$ and based on similar arguments of Ledoit and Wolf (2002) we can obtain that both

$$T_{n,p}(\tilde{\boldsymbol{\theta}}) = \sqrt{\frac{n+1}{n-2}\frac{d(p)}{2}}\left(\frac{n-1}{n}\frac{L_{n,p}(\tilde{\boldsymbol{\theta}})}{d(p)}-1\right) \tag{5}$$

and $\sqrt{\frac{d(p)}{2}}\left(\frac{L_{n,p}(\tilde{\boldsymbol{\theta}})}{d(p)}-1\right)$, have an (n,p)-asymptotic distribution given by $\mathcal{N}(0,1)$. In particular, for $\boldsymbol{R}_0 = \boldsymbol{I}_p$, the test-statistic proposed by Schott (2005) is obtained, where $L_{n,p}(\tilde{\boldsymbol{\theta}}) = n\sum_{i<j}R_{ij}^2$, with R_{ij} being the (i,j)-th component of $\boldsymbol{R}_{n,p}$. It is important to take into account that in such a case, it holds $\mathrm{E}[T_{n,p}(\tilde{\boldsymbol{\theta}})] = 0$ and $\mathrm{Var}[T_{n,p}(\tilde{\boldsymbol{\theta}})] = 1$ for any value of (n,p), not only for large values.

4 Extension to elliptical distributions

Muirhead and Waternaux (1980) payed special attention on the ϵ-contaminated p-variate elliptical normal distribution. Taking into account that this is a particular case of a p-variate elliptical distribution $\boldsymbol{X} \sim \mathcal{E}_p(\boldsymbol{\mu},\boldsymbol{\Omega},g)$, it is considered the idea of preserving the asymptotic distributions given in Section 3 through a correction factor, $1+\kappa$ or $1+\widetilde{\kappa}$, related to the kurtosis parameter κ of $\boldsymbol{X}$. For testing (3), having a non fully determined elliptical distribution, it is possible to use the same three test statistics replacing $L_{n,p}(\tilde{\boldsymbol{\theta}})$ by $L_{n,p}(\tilde{\boldsymbol{\theta}}, \widetilde{\kappa}) = L_{n,p}(\tilde{\boldsymbol{\theta}}) / \left(1 + \widetilde{\kappa}\right)$, where

$$1 + \widetilde{\kappa} = \frac{1}{p(p+2)} \frac{1}{n} \sum_{i=1}^{n} \left(\widetilde{\Delta}_p^2(\boldsymbol{X}) \right)^2 , \tag{6}$$

with $\widetilde{\Delta}_p^2(\boldsymbol{X}_i) = \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_n\right)^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{n,p}\right)$ and $\widetilde{\boldsymbol{\Sigma}} = \mathrm{diag}^{\frac{1}{2}}\{\boldsymbol{S}_{n,p}\}\boldsymbol{R}_0$ $\times \mathrm{diag}^{\frac{1}{2}}\{\boldsymbol{S}_{n,p}\}$.

5 Real data application and simulation study

To illustrate the proposed approach, we employ a subset of the biochemical data presented in the work of Beerstecher et al. (1950). This dataset comprises 62 distinct measurements for each of the 12 subjects, with 8 serving as controls and the remaining 4 identified as alcoholics. Our focus will be limited to a group of 8 measurements related to blood serum. For both groups, the control and the alcoholic, we intend to examine the hypothesis of complete independence. It is evident that we cannot apply the likelihood ratio test for either of the groups due to the fact that $p = 8$ and $n = 8$ for the control group, while $n = 4$ for the alcoholic group. In the case of the control group, we obtain that $L_{n=8,p=8}(\tilde{\boldsymbol{\theta}}, \widetilde{\kappa}) = 63.7245$ and $\widetilde{\kappa} = 0.1294$. We obtain a value 3.4977 for eq. (5), which has a p-value equals 0.0005. Under the assumption of p-variate ellipticity, this provides us with substantial evidence of some correlation among the eight variables. Shifting our focus to the alcoholic group, we find that $L_{n=4,p=8}(\tilde{\boldsymbol{\theta}}, \widetilde{\kappa}) = 47.2417$ and $\widetilde{\kappa} = -0.1437$, having a p-value equals 0.005 (from 2.8268 as the value for eq. (5)). Thus, we found enough evidence to reject lack of correlation for both the alcoholic group as well as for the control group. Such a result, under the assumption of p-variate ellipticity, does not align with the one presented in Schott (2005) for the same data, alcoholic group, under the assumption of normality, nor with the one in Shi et al. (2024) under a mild distributional assumption, a continuous p-variate random vector. In both papers, there is no clear evidence to reject the hypothesis of no correlation, but the p-value is close to 0.05.

TABLE 1. Estimated significance levels for test-statistic (5) adapted for elliptical distributions

	$n = 5$	$n = 9$	$n = 17$	$n = 33$	$n = 65$	$n = 129$	$n = 257$
$p = 4$	0.0780	0.0597	0.0529	0.0474	0.0493	0.0481	0.0448
$p = 8$	0.0753	0.0534	0.0471	0.0408	0.0428	0.0449	0.0417
$p = 16$	0.0889	0.0578	0.0468	0.0423	0.0386	0.0440	0.0457
$p = 32$	0.1144	0.0670	0.0512	0.0408	0.0456	0.0458	0.0487
$p = 64$	0.1917	0.1078	0.0622	0.0490	0.0436	0.0432	0.0459
$p = 128$	0.2780	0.2456	0.1150	0.0630	0.0491	0.0466	0.0453
$p = 256$	0.2800	0.3885	0.2625	0.1283	0.0721	0.0511	0.0469

Though 10,000 replications we desire to study by simulation the estimated exact significance levels with nominal level $\alpha = 0.05$, when $p, n{-}1 \in \{4, 8, 16, 32, 64, 128, 256\}$.

TABLE 2. Estimated significance levels for $-n\log(\det(\boldsymbol{R}_{n,p}))/\,(1+\widetilde{\kappa})$

	$n=5$	$n=9$	$n=17$	$n=33$	$n=65$	$n=129$	$n=257$
$p=4$	0.1602	0.0714	0.0588	0.0493	0.0508	0.0513	0.0475
$p=8$		0.3478	0.0612	0.0414	0.0384	0.0447	0.0434
$p=16$			0.5960	0.0376	0.0203	0.0265	0.0312
$p=32$				0.7643	0.0132	0.0066	0.0099
$p=64$					0.8935	0.0014	0.0000
$p=128$						0.9710	0.0000
$p=256$							0.9982

The simulation results, given in Tables 1 and 2, consider an ϵ-contaminated p-variate elliptical distribution. This is achieved by simulating the p-variate standard normal distribution and contaminating a percentage of 5% of the data (i.e., $\epsilon = 0.05$) with $\sigma = 2$. This implies that $\kappa = 0.32$. In an extended version of this study, power estimates are also taken into account.

As expected, the observed significance levels approach the exact ones more precisely when increasing (n, p). The overall behavior is similar to the one shown in Schott (2005) for the normal distribution. However, when the distribution is extended to the whole family of elliptical distributions, the precision decreases. The LRT adapted for elliptical distributions (see Table 2) performs poorly when $p = n - 1$. For a fixed p, the approximation to the nominal level clearly improves as n increases. However, for a fixed n, the improvement is not as clear when p increases, particularly for low values of n.

Acknowledgments: Nirian Martín acknowledges the support of the UCM 2022 Call for Research Project Financing Grants under project reference PR3/23-30832. J. Miguel Marín acknowledges financial support from the Spanish State Research Agency under project reference PID2020-115598RB-I00.

References

Beerstecher Jr., E., Sutton, E., Berry, H.K., Brown, W.D., Reed, J., Rich, G.B., Berry, L.J. and Williams, R.J. (1950). Biochemical individuality. V. Explorations with respect to metabolic patterns of compulsive drinkers. *Archives of Biochemistry*, **29**, 27 – 40.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics*, **30**, 1081 – 1102.

Magnus, J.R. (1988). *Linear Structures.* Griffin's Statistical Monographs, London and Oxford University Press, New York.

McCulloch, C.E. (1982). Symmetric matrix derivatives with applications. *Journal of the American Statistical Association*, **77**, 679 – 682.

Muirhead, R.J. and Waternaux, C.M. (1980). Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations. *Biometrika*, **67**, 31 – 43.

Schott, J. (2005). Testing for complete independence in high dimensions. *Biometrika*, **92**, 951 – 956.

Shi, X., Jiang, Y., Du, J. and Miao, Z. (2024). An adaptive test based on Kendall's tau for independence in high dimensions. *Journal of Nonparametric Statistics*, in Press.

DGLMExtPois package: regression models for under-dispersed count data

Ana María Martínez-Rodríguez[1], Antonio Conde-Sánchez[1], María José Olmo-Jiménez[1], José Rodríguez-Avi[1]

[1] University of Jaén, Spain

E-mail for correspondence: `ammartin@ujaen.es`

Abstract: This work presents the DGLMExtPois package, which allows the estimation of hyper-Poisson and COM-Poisson regression models suitable for under-dispersed count data. The package also includes functions for analysing these regression models, as well as performing model diagnostics. To demonstrate the practical utility of the package, it is applied to a real data set, showing its effectiveness in modelling real-world scenarios.

Keywords: Hyper-Poisson; Regression model; Count data.

1 Introduction

The analysis of count data is usually performed using the Poisson or the negative binomial distributions, which are not suitable when there is underdispersion. Therefore, in recent years, models that do allow for underdispersion, such as the hyper-Poisson model or the COM-Poisson model, have been considered and used in regression analysis. This work presents the **DGLMExtPois** package of R, which allows dealing with such regression models (Sáez-Castillo et al., 2022). Such models can simultaneously consider overdispersion and underdispersion as a function of the levels of the covariates.
To demonstrate the practical utility of the package, it is applied to a real-world dataset. In this case, the number of primary schools per municipality in Andalusia (Spain). The most appropriate hyper-Poisson model has been estimated by performing a diagnosis of the model and comparing it with the COM-Poisson model. It should be noted that there is no other package that offers the option of fitting a hyper-Poisson model.

1.1 Hyper-Poisson regression model

Let us consider Y as a count variable of a hyper-Poisson distribution (hP), which has two parameters γ and λ (see details in Sáez-Castillo and Conde-Sánchez

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

(2013)). The parameter γ is known as the dispersion parameter because for $\gamma > 1$ the distribution is over-dispersed and for $\gamma < 1$ is under-dispersed (for $\gamma = 1$ we have the Poisson distribution).

In a $hP_{\mu,\gamma}$ GLM model, Y follows a hP distribution whose mean, μ, and dispersion parameter, γ, are functions of the covariates. Also, log-linear relationships are always considered, i.e.,

$$\begin{aligned}\mu_i &= \exp\left(\mathbf{x_i'}\boldsymbol{\beta}\right)\\ \gamma_i &= \exp\left(\mathbf{z_i'}\boldsymbol{\delta}\right)\end{aligned}$$

This allows for the possibility of over- and/or under-dispersion depending on the values of the covariates. Also, the parameter λ is solution of

$$\mu = \exp(\mathbf{X}\boldsymbol{\beta}) = \sum_{y=0}^{\infty} y \times P[Y = y]. \tag{1}$$

1.2 COM-Poisson regression model

The COM-Poisson distribution also has two parameters, θ and, ν (Huang, 2017). Here, $\nu < 1$ corresponds to over-dispersion and $\nu > 1$ to under-dispersion. In this distribution, θ is a location parameter with a similar role to λ for the hP distribution.

In the regression model proposed by Huang (2017), $CMP_{\mu,\nu}$, the covariates also determine the mean and dispersion, unlike previous models, where the covariates are introduced in expressions of the parameters (Sellers, 2010). In addition, the parameter ν is solution of an equation similar to (1).

2 DGLMExtPois package

The R package **DGLMExtPois** allows the estimation of $hP_{\mu,\gamma}$ and $CMP_{\mu,\nu}$ regression models, as well as the diagnosis of the estimated models. This package uses a procedure for estimating the regression coefficients within the GLM framework, through a gradient-based algorithm, solving a non-linear constrained optimisation problem. **DGLMExtPois** has been created trying to reproduce the syntax of GLM fits.

Thus, functions `glm.hP` and `glm.CMP` provide the corresponding fits for $hP_{\mu,\gamma}$ and $CMP_{\mu,\nu}$, respectively. There are also `print` and `summary` functions to visualize the fitted models and `AIC`, `confint`, `predict`, `residuals` and `plot` functions to evaluate goodness of fit, obtain confidence intervals of $\boldsymbol{\beta}$ regression coefficients, predictions, residuals and some associated plots, such as QQ-plots and simulated envelopes, respectively. The package additionally incorporates `expected`, a function which calculates the marginal probabilities of the Y variable, and `lrtest` to perform the likelihood ratio test in nested models. Other functions are provided, for example, to work with probabilities of CMP and hP distributions. A brief description of the main functions in the package is provided in Table 2.

TABLE 1. Main functions in the DGLMExtPois package.

Function	Description
`glm.hP`	Fits a $hP_{\mu,\gamma}$ model
`glm.CMP`	Fits a $CMP_{\mu,\nu}$ model
`summary`	Computes and returns a list of summary statistics
`print`	Prints estimated model
`lrtest`	Likelihood ratio chi-squared test
`residuals`	Extracts model residuals (Pearson, response and quantile)
`plot`	Plot of residuals against fitted values and a Q-Q plot
`predict`	Predictions
`confint`	Confidence intervals for $\boldsymbol{\beta}$ regression coefficients
`AIC`	Returns AIC
`hP_expected`	Expected frequencies for $hP_{\mu,\gamma}$ model
`CMP_expected`	Expected frequencies for $CMP_{\mu,\nu}$ model

3 Application

This work analyses the factors that influence the number of primary schools in Andalusian municipalities (range: 0 to 94). Since this variable is a count variable, the Poisson regression model has been considered, as well as the *hP* and *CMP* regression models.

The factors considered are: population, average age, unemployment rate, number of population centres, distance to the capital, and province, which is a categorical variable with 8 provinces (Almería, Cádiz, Córdoba, Granada, Huelva, Jaén, Málaga and Sevilla). The data are for the year 2021 (785 observations) and have been collected from the Andalusian Institute of Statistics (IEA).

A *hP* model has been fitted ($AIC = 1381$), observing that there is underdispersion and that its fit is much better than that obtained with the equivalent Poisson ($AIC = 1947$) and *CMP* ($AIC = 1476$) models. The results are shown in the Table 2, where we have eliminated those variables that are not significant: average age and distance to the capital city. In addition, a logarithmic transformation of the population has been considered, as well as its square and that of the variable number of population centres. The provinces of Jaén and Córdoba, on the one hand, and the rest of the provinces on the other (being the base category) have been grouped, also considering the interaction with the population variable.

It has also been found that none of the factors analysed influences the dispersion parameter, so that a model with constant dispersion has been considered (for the *hP* model):

$$\gamma = \exp\left(\boldsymbol{\delta_0}\right)$$

The estimation of $\widehat{\delta_0} = -7.950$ in the hP model leads to an underdispersed model, since $\widehat{\gamma} = 0.000353 < 1$. The same is true for the COM-Poisson model (where $\widehat{\nu} > 1$). The presence of under-dispersion in the distribution analysed indicates that the variability is small once the factors that influence the behaviour of the response variable have been considered.

It can be seen that the variable that has the greatest influence is population, with differences also being found between the two groups of provinces, so that in Jaén and Córdoba the number of provincial centres is greater than in the rest of the provinces, with this difference increasing as the population grows.

TABLE 2. Coefficient estimates and standard errors of *CMP* and *hP* fitted model.

	Mean model coefficients			
	CMP		*hP*	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	-0.78227	0.062679 ***	-0.64151	0.079525 ***
P.Centers	0.02190	0.002496 ***	0.02920	0.004316 ***
P.Centers^2	-0.00033	0.000054 ***	-0.00046	0.000094 ***
log(Pob)	0.51252	0.023870 ***	0.46446	0.028276 ***
log(Pob)^2	0.03785	0.003198 ***	0.04321	0.004595 ***
prov	0.28088	0.076718 ***	0.28151	0.052336 ***
Unemployment	0.01331	0.002295 ***	0.00932	0.003262 **
log(Pob)*prov	-0.16373	0.062599 **	-0.19614	0.063656 **
log(Pob)^2*prov	0.03069	0.012214 *	0.04094	0.015365 **
	Dispersion model coefficients			
(Intercept)	1.649	0.056 ***	-7.950	1.156 ***
. p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001				

A diagnosis of the *hP* model has also been performed, through the residual plots shown in Figure 1, which indicate that the diagnosis is satisfactory. In addition, the observed and expected frequencies for the response variable (up to the value 20) have been compared and represented graphically by means of a barplot (Figure 2).

TABLE 3. Goodness-of-fit measures of *CMP* and *hP* fitted model.

	CMP	*hP*
AIC	1476	1381
Dif	235.8146	78.35461
χ^2	155.3135	84.16762

On the other hand, the diagnosis of the *CMP* model is not as adequate, as the QQ-plot of the Figure 3 shows. Also, the expected frequencies are much further away from the observed frequencies (Figure 2). In fact, the difference in absolute value of such frequencies (Dif statistic), as well as the squared differences between these frequencies weighted by the expected frequencies (χ^2 statistic) are much higher than for the *hP* model (Table 3). In conclusion, the *hP* model is better suited than the *CMP* model to fit these data.

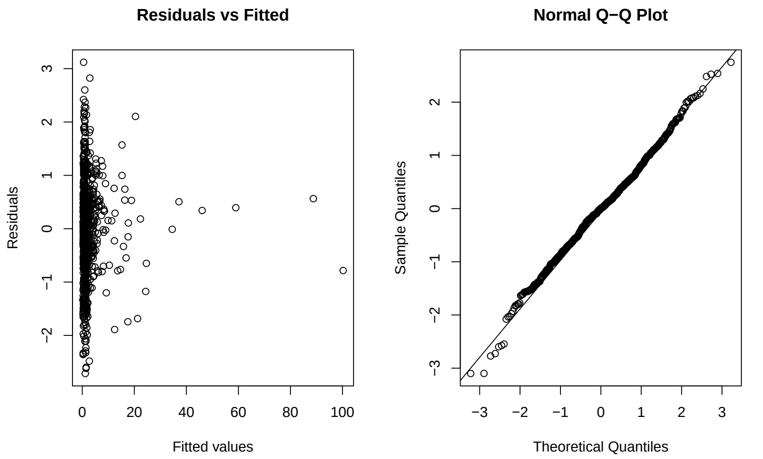


FIGURE 1. Diagnosis plots (hP model).

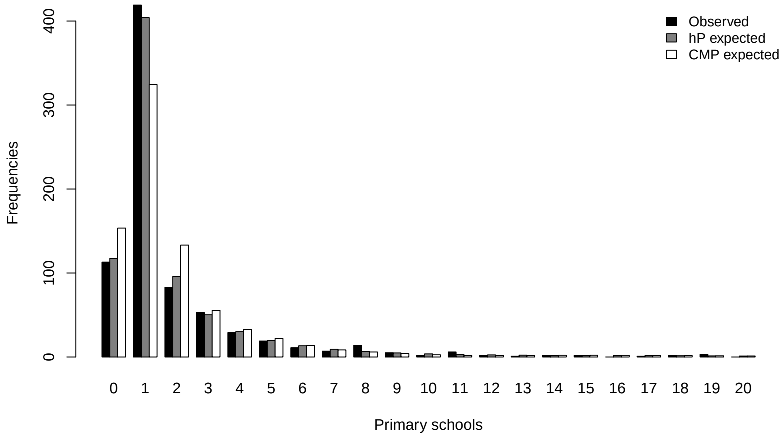


FIGURE 2. Expected and observed frequencies.

References

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*, **17**, 359-–380.

Sáez-Castillo, A.J. and Conde-Sánchez, A. (2013). A hyper-Poisson regression model for overdispersed and underdispersed count data. *Computational Statistics and Data Analysis*, **61**, 148 – 157.

Sáez-Castillo, A.J., Conde-Sánchez, A. and Martínez, F. (2022). DGLMExtPois:

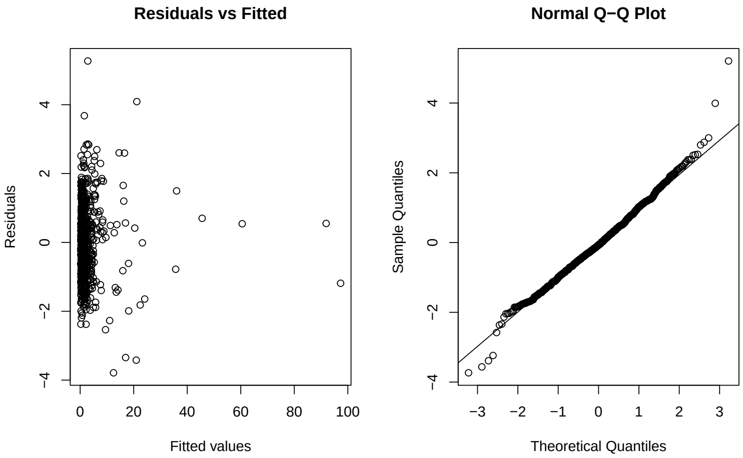


FIGURE 3. Diagnosis plots (CMP model).

Advances in dealing with over and under-dispersion in a double GLM framework. *The R Journal*, **14**, 121 – 140.

Sellers, K.F. and Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics*, **2**, 943 – 961.

An underrated prior distribution for proportions. The logistic–normal for dynamical football predictions

Rui Martins[1]

[1] Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)

E-mail for correspondence: `rmmartins@fc.ul.pt`

Abstract: The result of a football match in terms of Home–Win, Draw or Away–Win can be modelled by considering the observed outcome as a realization of a Multinomial random variable with three mutually exclusive events over a single trial. Most applications consider the Dirichlet distribution to represent the prior uncertainty about the Multinomial's proportion parameters, mainly because of conjugacy and the reduced number of parameters. As alternative we propose to use the Logistic–Normal, a multivariate prior distribution for proportions but to which little attention has been paid. This approach was motivated by the question – Are women's and men's football leagues equally predictable? The models developed are applied to the main Portuguese women's and men's football leagues over seven seasons, starting from 2016–2017 up to 2022–2023. The work also provides estimates of latent team-specific strengths and addresses the variability between and within seasons, along with insights of each team's home advantage.

Keywords: Compositional data; Dynamic models; Football prediction; Logistic-normal prior; Sports comparison.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Multivariate forecasting of operational times using hidden Markov models

Fernando Miguelez[1,2], Josu Doncel[3], Maria Dolores Ugarte[1,2]

[1] Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Spain
[2] Institute for Advanced Materials and Mathematics (InaMat^2), Public University of Navarre, Spain
[3] Department of Mathematics, University of the Basque Country, Spain

E-mail for correspondence: `fernando.miguelez@unavarra.es`

Abstract: Industrial processes generate a massive amount of monitoring data that can be exploited to uncover hidden time losses in the system, leading to enhanced accuracy of maintenance policies and, consequently, increasing the effectiveness of the equipment. In this work, we propose a method for one-step probabilistic multivariate forecasting of time variables based on a Hidden Markov Model with covariates (IO-HMM). These covariates account for the correlation of the predicted variables with their past values and additional process measurements by means of a discrete model and a continuous model. The probabilities of the former are updated using Bayesian principles, while the parameter estimates for the latter are recursively computed through an adaptive algorithm that also admits a Bayesian interpretation. This approach permits the integration of new samples into the estimation of unknown parameters, computationally improving the efficiency of the process. We evaluate the performance of the method using a real data set obtained from a company in the food sector; however, it is a versatile technique applicable to any other data set. The results show a consistent improvement over a persistence model, which assumes that future values are the same as current values, and over univariate versions of our model.

Keywords: Adaptive parameter estimates; Hidden Markov model; Industrial processes; Probabilistic prediction.

1 Background and objectives

In industrial settings, production processes often face inefficiencies that lead to time losses. These time losses can be broadly classified into four categories (Muchiri and Pintelon, 2008): losses due to scheduled stops such as maintenance or cleaning; losses due to unexpected stops such as setup, adjustment, failure,

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

or supply outage; losses due to low production speed and micro-stoppages; and losses due to the production of defective units and rework. One can also derive different production times by successively subtracting each time loss from the total length of the observation period, as well as some important efficiency indexes as ratios of these production times.

In this work, we propose a novel approach to predict time losses by modelling the production process carried out by the equipment as a multi-signal process, where the signals characterize the equipment's current operational mode. Furthermore, the predictive model includes other process features that can have an impact on the model parameters as covariates. To ensure continuous parameter updating using the latest data, we use an adaptive learning algorithm that admits a Bayesian interpretation. The forecasting of time losses in production processes can help to enhance the maintenance strategy's accuracy by identifying areas for improvement.

2 The model

We use an Input-Output Hidden Markov Model (IO-HMM) to model the production process, see (Bengio and Frasconi, 1996) for full details of IO-HMMs. Figure 1 illustrates an IO-HMM diagram. The process goes through K hidden states according to an initial state probability distribution and a transition probability distribution between states. The hidden state of the n-th observation period is denoted by c_n and represents the condition of the production process during that period. Each state gives rise to a different probability distribution of the continuous responses $\mathbf{y}_n$.

In an IO-HMM, the model's probability distributions are affected by an input stream of covariates, denoted by $\mathbf{x}_n$. These covariates may include, among others, calendar variables or the reference produced, and characterize the observation period that is about to begin. Further, we introduce an autoregressive component into the model by allowing the covariates to include past values of the response variables. The covariates that influence the probabilities in the discrete part of the model will be denoted by $\mathbf{z}_n \subseteq \mathbf{x}_n$, while the ones that impact the responses' joint density will be denoted by $\mathbf{w}_n \subseteq \mathbf{x}_n$.

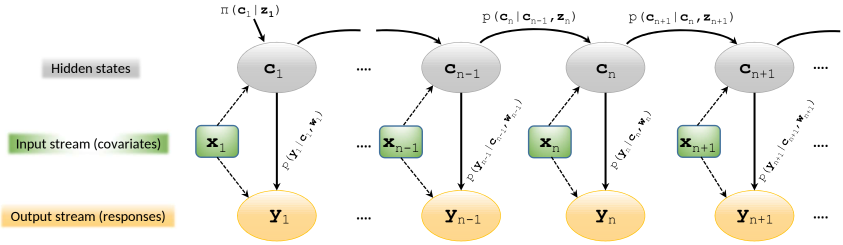


FIGURE 1. Input-Output HMM. Covariates $\mathbf{x}_n$ affect both discrete and continuous processes. Probabilities in the discrete process $\{c_n\}_{n\geq 1}$ are dependent on covariates $\mathbf{z}_n \subseteq \mathbf{x}_n$ and probabilities in the continuous process $\{\mathbf{y}_n\}_{n\geq 1}$ are dependent on covariates $\mathbf{w}_n \subseteq \mathbf{x}_n$.

3 Parameter estimates

We consider that the discrete process $\{c_n\}_{n\geq 1}$ is a Markov chain with K different states, $c_n \in \{1,\ldots,K\}\,,\ n \geq 1$. The probability distributions for the initial state and the transitions between states are dependent on the covariates $\mathbf{z}_n$, which take values in a discrete and finite set of S symbols. The unobserved next state c_n is categorical with parameters $\pi^{(s)}$ if $n = 1$ and $\mathbf{z}_n = s$, or $\mathbf{p}_k^{(s)}$ when $n > 1$ and $\mathbf{z}_n = s$. In turn, $\pi^{(s)}$ and $\mathbf{p}_k^{(s)}$ are Dirichlet with parameters - i.e., counts - updated every time an observation period ends.
On the other hand, we split the responses' joint density function into two conditional Gaussian distributions, namely

$$\begin{aligned} \mathbf{y}_n|\mathbf{w}_n &\sim \mathcal{N}_m(\mathbf{u}_n\mathbf{H}_u, \boldsymbol{\Sigma}_u) \\ \mathbf{y}_n|c_n &\sim \mathcal{N}_m(\mathbf{v}_n\mathbf{H}_v, \boldsymbol{\Sigma}_v), \end{aligned}$$

where $\mathbf{u}_n = \begin{bmatrix}1 \ \mathbf{w}_n^T\end{bmatrix}$, $\mathbf{v}_n = v(c_n)$ for a function $v(\cdot)$, $\mathbf{H}_u, \mathbf{H}_v$ are coefficient matrices and $\boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_v$ are covariance matrices. As soon as a new sample $\mathbf{y}_n$ becomes available, the estimators $(\mathbf{H}_{u,n-1}, \boldsymbol{\Sigma}_{u,n-1}, \mathbf{H}_{v,n-1}, \boldsymbol{\Sigma}_{v,n-1})$ are updated to $(\mathbf{H}_{u,n}, \boldsymbol{\Sigma}_{u,n}, \mathbf{H}_{v,n}, \boldsymbol{\Sigma}_{v,n})$ through an adaptive algorithm described by the multivariate extension of the equations introduced by Alvarez et al. (2021)

$$\begin{aligned} \mathbf{H}_{u,n} &= \mathbf{H}_{u,n-1} + \frac{\mathbf{P}_{u,n-1}\mathbf{u}_n^T}{\lambda_u + \mathbf{u}_n\mathbf{P}_{u,n-1}\mathbf{u}_n^T}\left(\mathbf{y}_n - \mathbf{u}_n\mathbf{H}_{u,n-1}\right) \\ \boldsymbol{\Sigma}_{u,n} &= \boldsymbol{\Sigma}_{u,n-1} - \frac{1}{\gamma_{u,n}}\left[\boldsymbol{\Sigma}_{u,n-1} - \frac{\lambda\left(\mathbf{y}_n - \mathbf{u}_n\mathbf{H}_{u,n-1}\right)^T\left(\mathbf{y}_n - \mathbf{u}_n\mathbf{H}_{u,n-1}\right)}{\lambda + \mathbf{u}_n\mathbf{P}_{u,n-1}\mathbf{u}_n^T}\right] \\ \mathbf{P}_{u,n} &= \frac{1}{\lambda_u}\left(\mathbf{P}_{u,n-1} - \frac{\mathbf{P}_{u,n-1}\mathbf{u}_n^T\mathbf{u}_n\mathbf{P}_{u,n-1}}{\lambda_u + \mathbf{u}_n\mathbf{P}_{u,n-1}\mathbf{u}_{,n}^T}\right) \\ \gamma_{u,n} &= 1 + \lambda_u\gamma_{u,n-1}, \end{aligned}$$

where λ_u is a forgetting factor. The algorithm is initialized with $\mathbf{H}_{u,0} = \mathbf{0},\ \boldsymbol{\Sigma}_{u,0} = \mathbf{0},\ \mathbf{P}_{u,0} = \mathbf{I}$ and $\gamma_{u,0} = 0$. The same updating equations are applied to compute $\mathbf{H}_{v,n}$ and $\boldsymbol{\Sigma}_{v,n}$ with the vector $\mathbf{v}_n$ and the forgetting factor λ_v.

4 Forecasting

At this stage each distribution produces a forecast of the responses, which are then combined using a minimum-variance criterion to obtain the final prediction. In particular, once the parameters are updated at the n-th time step the model computes the final prediction and a measure of its accuracy as

$$\begin{aligned} \hat{\mathbf{y}}_{n+1} &= \mathbf{u}_{n+1}\mathbf{H}_{u,n}\mathbf{D} + \mathbf{v}_{n+1}\mathbf{H}_{v,n}(\mathbf{I} - \mathbf{D}) \\ \hat{\boldsymbol{\Sigma}}_{n+1} &= \mathbf{D}\boldsymbol{\Sigma}_{u,n}\mathbf{D} + (\mathbf{I} - \mathbf{D})\boldsymbol{\Sigma}_{v,n}(\mathbf{I} - \mathbf{D}), \end{aligned}$$

where $\mathbf{D} = \operatorname{diag}(\delta_1, \ldots, \delta_m)$, $\delta_j = \sigma_{v,j}^2 / \left(\sigma_{v,j}^2 + \sigma_{v,j}^2\right),\ j = 1, \ldots, m$, and $\sigma_{v,j}^2$ (respectively $\sigma_{u,j}^2$) is the j-th element in the diagonal of $\boldsymbol{\Sigma}_{v,n}$ (respectively $\boldsymbol{\Sigma}_{u,n}$).

5 Real case study

The proposed model has been employed to predict time losses in the production process of a company that operates in the food industry. To measure the predictions' quality we use the well-known metrics Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The multivariate predictive model with an autoregressive component in the covariates $\mathbf{w}_n$ shows a consistent improvement in the predictions' quality against some benchmark models, including the persistence model, which assumes that future values are the same as current values (i.e., $\hat{\mathbf{y}}_{n+1} = \mathbf{y}_n$), the model with no autoregressive component and the respective univariate versions of our model.

Acknowledgments: This research was supported in part by the Government of Navarre under Smart Zeo Project 0011-1365-2021-000085.

References

Alvarez, V., Mazuelas, S. and Lozano, J.A. (2021). Probabilistic load forecasting based on adaptive online learning. *IEEE Transactions on Power Systems*, **36**, 4, 3668 – 3680.

Bengio, Y. and Frasconi, P. (1996). Input-Ouput HMMs for sequence processing. *IEEE Transactions on Neural Networks*, **7**, 1231 – 1249.

Muchiri, P. and Pintelon, L. (2008). Performance measurement using overall equipment effectiveness (OEE): literature review and practical application discussion. *International Journal of Production Research*, **46**, 3517 – 3535.

Parametric and non-parametric Bayesian imputation for right censored survival data

Shirin Moghaddam[1], John Newell[2], John Hinde[2]

[1] School of Mathematical and Statistical Sciences, University of Galway, Galway, Ireland

E-mail for correspondence: shirin.moghaddam@ul.ie

Abstract: A common feature of much survival data is censoring due to incompletely observed lifetimes. Survival analysis methods have been designed to take account of this and provide appropriate relevant summaries, such as the Kaplan–Meier plot and the median is easily read off this plot. However, a single summary is not really a relevant quantity for communication to an individual patient, as it conveys no notion of variability and uncertainty. The aim of this paper is to consider censored data as a form of missing data and impute them using Bayesian methods. We introduce two novel parametric and non-parametric Bayesian approaches for imputing right censored observations to be used as a complement to formal inferential methods and to allow more interpretable displays to be made for physicians and patients.

Keywords: Survival analysis; Bayesian methods; Imputation methods; Dirichlet process; Censored observation.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Bayesian approaches to model overdispersion in spatio-temporal binomial data

Mabel Morales-Otero[1], Vicente Núñez-Antón[1]

[1] Institute of Data Science and Artificial Intelligence (DATAI), University of Navarra, Pamplona, Spain

E-mail for correspondence: vicente.nunezanton@ehu.eus

Abstract: In this work, we introduce a direct spatio-temporal extension of the spatial conditional overdispersion models for binomially distributed response variables. This proposal incorporates a spatial term similar to the spatial lag of the response variable for each time unit within the linear predictor. These models effectively capture both spatial and temporal correlations inherent in the dataset under study. Furthermore, we introduce temporally varying spatial lag coefficient models, enabling for the possibility of introducing temporal changes in the spatial term. In order to be able to assess the usefulness of our proposals, we apply them to the analysis of low birth weight in Georgia, providing a comparative analysis of the performance of our models to that of the commonly used Knorr-Held's models.

Keywords: Bayesian models; Overdispersion; Spatio-temporal models.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Addressing covariate lack in unit-level small area models using GAMLSS

Lorenzo Mori[1], Maria Rosaria Ferrante[1]

[1] University of Bologna, Department of Statistical Sciences "Paolo Fortunati", Bologna, Italy

E-mail for correspondence: `lorenzo.mori7@unibo.it`

Abstract: The primary goal of this study is to estimate the Theil index using a unit-level Small Area Estimation (SAE) model. This has lead two primary challenges in the unit-level SAE field: the identification of individual covariates and the reduction of computational burden. We propose a unit-level Simplified SAE model based on Generalized Additive Models for Location, Scale and Shape (GAMLSS), which is specified without covariates and is able to reduce variability in comparison with the direct estimator. The performance of the proposed model used to estimate the Theil index is evaluated based on design-based simulations. An application to the Italian Regions, distinguish between Urban, Peri-Urban and Rural areas, conclude the paper.

Keywords: GAMLSS; Atkinson index; Computational burden

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Segmented quantile regression process with τ-varying breakpoints

Vito M.R. Muggeo[1], Gianluca Sottile[1]

[1] Dipartimento di Scienze Economiche, Aziendali e Statistiche, University of Palermo, Italy

E-mail for correspondence: vito.muggeo@unipa.it,gianluca.sottile@unipa.it

Abstract: We introduce the Segmented Quantile Regression Process framework to model the entire quantile region of the conditional response distributions as a segmented relationship with respect to the continuous covariate. Each model parameter, including the breakpoint, is assumed to vary smoothly across the τ. The framework is illustrated on the well-known dataset about maximal running speed and weight in mammals.

Keywords: Quantile regression process; Breakpoint; Segmented regression.

1 Introduction

Segmented quantile regression (segQR) postulates the quantile $Q_Y(\tau|x)$ of the conditional response distribution $Y|x_i$ depends on a continuous covariate via a segmented or piecewise linear relationship. The regression equation is $Q_{Y_i}(\tau|x_i) = \beta_{0\tau} + \beta_{1\tau}x_i + \delta_\tau(x_i - \psi_\tau)_+$, where the subscript refers to the *fixed* probability value of interest. At any value of $\tau \in (0,1)$ the 'left' (i.e. when $x \leq \psi$) slope is $\beta_{1\tau}$, and the 'right', i.e. when $x > \psi$, slope is $\beta_{1\tau} + \delta_\tau$. Estimation of segmented QR models at given τ has been discussed in Li et al. (2011) and Yan et al. (2017), among others: the minimand objective $\rho_\tau(\beta_{0\tau}, \beta_{1\tau}, \delta_\tau, \psi_\tau) = \sum_i |y_i - Q_{Y_i}(\tau|x_i)|w_{i\tau}$ where the $w_{i\tau}$'s are the usual weights equal to τ or $1-\tau$ depending on the sign of residuals.
However, while focus on a specified τ could be relevant in some examples, the entire collection of quantile curves, the so-called Quantile Regression Process QRP, is the big and worthwhile deal of QR analyses: one can estimate parameters more efficiently and also gain insights on the entire response distributions, including the tails where usually data are very sparse and estimates coming from single fits are unstable.
Currently there are two main options to fit linear QRP
Via the LMS method within the GAMLSS tool (Stasinopoulos et al. 2017, chapter 13), whereby: i) one chooses the conditional response distribution depending

on several parameters (usually mean, dispersion, skew and kurtosis); ii) specifies a (flexible, i.e. via splines) sub-regression model for each parameter; iii) fits the model via maximum likelihood; iv) then obtains the quantile curves via the known cumulative distribution function. The main drawback in the LMS/GAMLSS approach is that the model returns coefficients for the different sub-regression equations (mean/dispersion/skew) and it does not return the (meaningful) parameters for the quantile regression, namely slopes and breakpoints.

Alternatively, the Frumento et al. (2016) proposal relies on the 'pure' QR framework, namely they minimize the so-called integrated quantile loss function $\int \rho_\tau(\cdot)d\tau$. While the approach works well for linear models, it does not appear straightforward how to extend it for the breakpoint which is non separable from the covariate.

In this paper we propose the Segmented Quantile Regression Process to extend applicability and of the linear QRP. The rest of the paper is structured as follows: section 2 describes methodology and the estimating algorithm to fit it; section 3 illustrates an example and the last section is devoted to conclusion and discussion.

2 The segmented regression quantile process model: set-up and fitting

We define the segmented QRP, segQRP, via

$$Q_Y(\tau|x_i) = \beta_0(\tau) + \beta_1(\tau)x_i + \delta(\tau)(x_i - \psi(\tau))_+, \tag{1}$$

where the τ's in parentheses, rather than in the subscripts, stress that the overall pattern of each model parameter across the whole probability range is of interest. Interpretation of model parameters is the same of the simple segQR model: $\beta_1(\tau)$ represents the slope when $x_i \leq \psi(\tau)$, while $\beta_1(\tau) + \delta(\tau)$ is the covariate effect when $x_i > \psi(\tau)$.

We remark the aforementioned segQRP has never been proposed. Its main and most noteworthy feature is the breakpoint parameter $\psi(\tau)$ which represents a τ-varying threshold: in fact it can be of scientific interest to estimate how the possible threshold depends on the percentile of the conditional response distribution. But $\psi(\tau)$ is also the most painful point. It is not a trivial matter to fit the segQRP model as the quantile loss function is nonsmooth and nonconvex with respect to the change point which hinders the usual optimizations algorithms. As also sketched in previous section, $\psi(\tau)$ cannot be separated by the covariate which complicates further the settings.

A rough and simple strategy to gain information on the segQRP could be to fit separate segQR models at different τs, but the approach can suffer from several drawbacks, including crossing curves and unstable estimation of the breakpoints with severe loss of efficiency, especially at extreme quantiles.

To fit the segQRP model we rely on the works of Muggeo (2003) to estimate breakpoints in simple mean regression and Muggeo et (2021) which uses a discrete approximation of the integrated loss $\int \rho_\tau(\cdot)\mathrm{d}\tau$ to fit a linear QRP. Using a simple Taylor expansion of the term $(x_i - \psi(\tau))$ we approximate the segmented regression equation around a known value $\tilde{\psi}(\tau)$ into a linear model with coefficients $\beta_0(\tau), \beta_1(\tau), \delta(\tau)$, and $\psi(\tau)$. Then we fix K probability values to build the loss function $\sum_\tau \sum_i \rho_\tau(y_i - Q_i(\tau))$, and express each model parameter via proper B-splines,

$$\beta_0(\tau) = C(\tau)\theta_0 \qquad \beta_1(\tau) = C(\tau)\theta_1 \qquad \delta(\tau) = C(\tau)\theta_2 \qquad \psi(\tau) = C(\tau)\theta_3,$$

where $C(\tau)$ is a, typically but not necessarily the same, B-spline basis on the probability values. Expressing each model parameter via B-splines allows to get smooth effects across the probability values and to achieve an overall loss objective as a function of the $\theta = (\theta_0, \ldots, \theta_3)^T$ only. Such working linear QRP is fitted as discussed in Muggeo et al. (2023) and the procedure is repeated iteratively till convergence.

Remark

To prevent non-crossing proper inequality constraints can be set. For convenience, but without losing generality, we first shift the covariate values in the range $[0, m]$; then, starting from the usual condition to fulfil, $\partial Q(\tau)/\partial\tau \geq 0$, we end up with proper constraints based on the first order differences of the θ's. Such noncrossing constraints translate into linear inequality constraints which can be easily accommodated into optimization algorithms for L_1 norm objectives.
Since the breakpoint is bounded in the covariate range $(0, m)$ say, we re-parametrise it via a logistic function

$$\psi(\tau) = \frac{m \exp \kappa(\tau)}{1 + \exp \kappa(\tau)}$$

where $\kappa(\tau)$ is unbounded and therefore it is straightforward to write $\kappa(\tau) = C(\tau)\theta_3$, rather than $\psi(\tau)$.
The proposed algorithm can be also used if we want to fix the breakpoint, namely $\psi(\tau) = \psi$ for each τ. It suffices to replace the basis $C(\tau)$ by a column of ones in the formula for $\psi(\tau)$.

3 Application

The relationship between maximal running speed (MRS, Km/h) and size (mass, Kg) in land mammals is often expressed by the allometric equation

$$\text{MRC} = \exp(\alpha) \times \text{mass}^\beta \quad \leftrightarrow \quad \log \text{MRC} = \alpha + \beta \log(\text{mass})$$

.

Figure 1 portrays data, body mass (in Kg) and maximal running speed (Km/h) of $n = 107$ land mammals (Garland, 1983); the dataset is named `Mammals` in the R package quantreg. Data, reported on the log scales, suggest that a simple linear relationship is not adequate, as the speed decreases as the mass exceeds some breakpoint, at about $e^{3.5}$ kg. However biologists may be interested in assessing whether such threshold value holds constant both all land mammals, or some difference exists between the slowest and fastest ones. Namely the research question calls for using the segQRP in equation (1), and we fit the model using $K = 11$ probability values with the noncrossing constraints Figure 1 portrays the predictions of one hundred quantile curves with the red line emphasizing the pattern of the τ-varying breakpoint $\hat{\psi}(\tau)$.

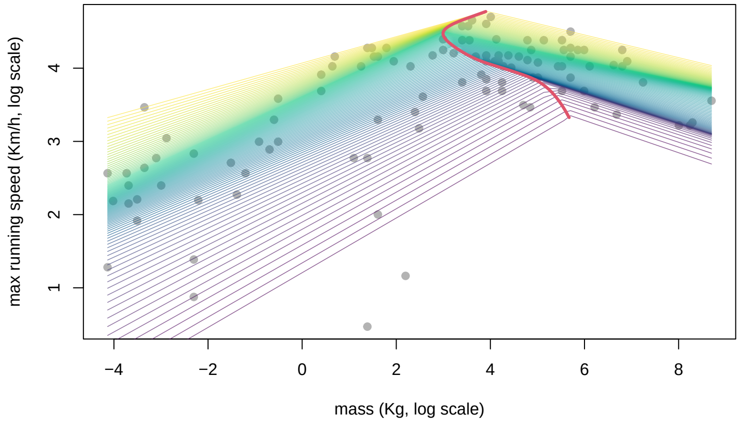


FIGURE 1. The Mammals dataset: data and fitted segQRP lines. The red curved line joins the breakpoint estimates across the different quantile curves.

In Figure 2 we portray the smoothed coefficients of equation (1) along with the estimates coming from the 'naive' fits obtained by assuming several values of τ from time to time. Unsurprisingly, the naive estimates are more wiggly with some abrupt changes which are unreasonable from a biological point of view.

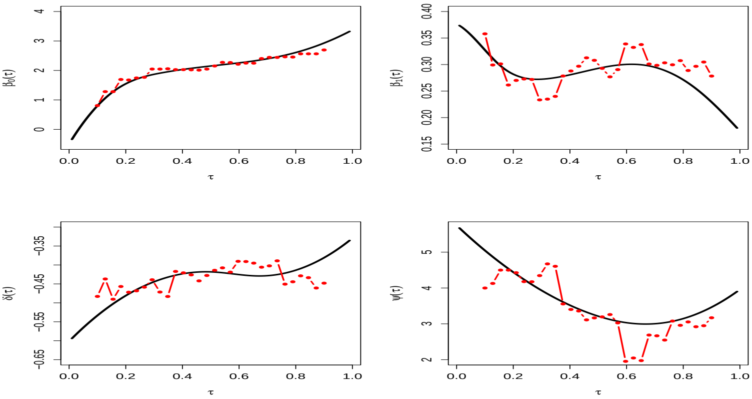


FIGURE 2. The Mammals dataset: Estimated τ-varying pattern of each model parameter of the segQRP; on each panel the dotted points represent the estimates coming from fitting separate model at different values of τ.

4 Conclusion

We have introduced the segmented QRP which models the covariate effect on the conditional quantile via segmented relationship, by returning the smoothed pattern of the threshold parameter across the percentile values. Model fitting is carried out within the recent framework of Muggeo et al (2023) which ensures noncrossing quantile curves via a few linear inequalities. It should be stressed that alternative approaches which could work for *linear* QRP, such as the GAMLSS framework (Stasinopoulos et al. 2017) and the integrated quantile of Frumento and Bottai (2016), here are not usable, as the τ-varying pattern of the breakpoint cannot be modelled via the LMS regression equations and the breakpoints is not separable from the design matrix.

References

Garland, T. (1983). The relation between maximal running speed and body mass in terrestrial mammals. *Journal of Zoology*, **199**, 157 – 170

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

Li, C., Wei, Y., Chappell, R. and He, X. (2011) Bent line quantile regression with application to an allometric study of land mammals' speed and mass, *Biometrics*, **67**, 242 – 249

Muggeo, V.M.R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, **22**, 3055 – 3071

Muggeo, V.M.R., Sottile, G. and Cilluffo, G. (2023). Joint modelling of non-crossing additive quantile regression via constrained B-spline varying coefficients, *Statistical Modelling*, **23**, 540 – 554

Stasinopoulos, D., Rigby, R.A., Heller, G.Z., Voudouris, V. and De Bastiani, F. (2017). *Flexible Regression and Smoothing Using GAMLSS in R*, CRC press.

Yan, Y., Zhang, F. and Zhou, X. (2017). A note on estimating the bent line quantile regression model, *Computational Statistics*, **32**, 611 – 630

Spatio-temporal models for high resolution wind speed maps

Eamonn Organ[1], James Sweeney[1]

[1] Department of Mathematics and Statistics, University of Limerick, Ireland

E-mail for correspondence: `Eamonn.Organ@ul.ie`

Abstract: In this paper a variety of spatio-temporal models to predict wind speeds at unobserved locations are examined. Prediction of wind speeds at unobserved locations is of importance in the wind energy industry to evaluate the potential of new sites for wind farms, a growing source of electricity. These models will include novel non-stationary spatial models that incorporate information from mechanistic physical models. Wind observations are often limited to sparse weather stations, although they offer high temporal resolution. Presently, in order to create wind maps at a high spatial resolution, mechanistic physical models are applied to historical weather data. These datasets are known as reanalysis data. These are still limited in spatial and temporal resolution and require large computational resources. These can also contain biases. For the models presented here, data from weather stations from Ireland's Met Office is used, along with information from mechanistic models to inform the model parameters.

Keywords: Spatio-temporal; SPDE; Wind; Reanalysis.

1 Background and datasets

To address climate change and enhance energy independence, nations are prioritizing the integration of renewable energy sources into their grids. In Ireland, wind energy stands as the primary renewable source, providing 35% of the nation's electricity last year. The Irish government have set a target of 80% renewables by 2030. However, the variability of wind energy poses challenges in both estimating resources and forecasting future values. Our research aims to develop spatial models for creating higher-resolution wind maps to improve wind resource estimation. These can also be generated in real-time from sparse observations. Existing wind speed data is typically sourced from synoptic weather stations maintained by national Met services or reanalysis data such as ERA5 and MERRA5. While synoptic stations offer high temporal resolution, they are spatially limited. Reanalysis data is when numerical weather models are applied retrospectively, essentially using the actual observations from the past to force

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

the model results across a regular grid. Though higher in resolution, these still faces limitations in temporal and spatial granularity. For example, ERA5, which was seen as the most accurate reanalysis dataset in prior studies (see Doddy Clarke, E. et al. (2021)) is at a 0.25° ($\sim$ 20-30km) spatial grid and one hour time resolution. Some previous efforts to create wind maps at sub-kilometre spatial resolution include the European Wind Atlas and the Irish Wind Atlas from the Sustainable Energy Authority Ireland. While high resolution these either provide long term averages, or are not regularly maintained.

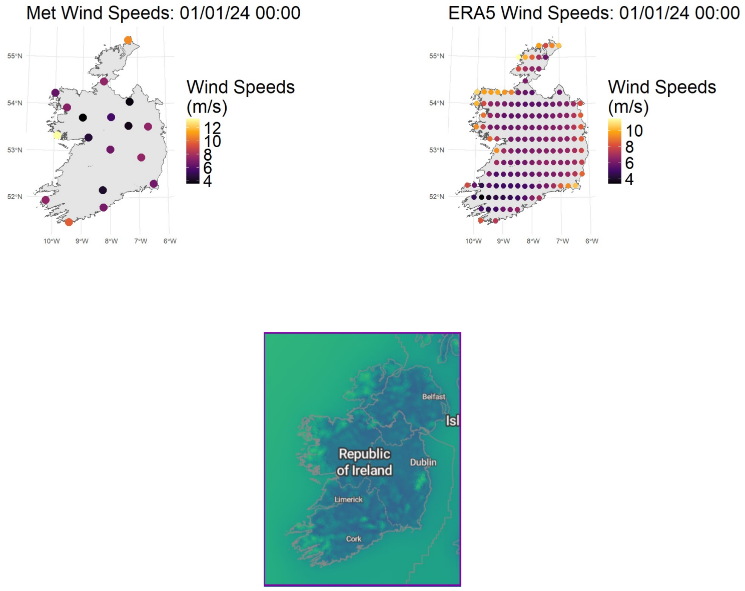


FIGURE 1. Images showing Met data (top left) and ERA (top right) data for a single time point, as well as the New European Wind Atlas (bottom) long term average wind speed.

2 Spatio-temporal models used

The first model introduced considers only data from met station. Later models will describe how reanalysis data can be incorporated to improve the models. The simplest spatio-temporal model is a stationary isotropic Gaussian process. Denoting the wind speed as W, the wind speed is observed over time, with a given time point denoted as t. The wind speed is assumed to be a continuous process over some domain D, and we denote a given location as $s \in D$. The wind speed at a given location and time is modelled as:

$$W_{t,s} = \mu_0 + f(t,s) + \epsilon(t,s), \tag{1}$$

where μ_0 is a constant intercept term. $f(t,s)$ is a stationary,isotropic and separable spatio-temporal Gaussian process. A Gaussian process is a stochastic process where a finite collection of random variables from the process form a multivariate Gaussian distribution. Assuming zero mean, it is fully described by a covariance matrix Σ, which includes the covariance between observed and prediction locations. A separable process means the covariance between space and time can be factored into a spatial covariance function and a temporal covariance function, while stationary isotropic describes models where the covariance between location and times depends only on distance between points (distance in time and space) and not their locations or relative direction.

The covariance formula can be written as,

$$\begin{aligned} c(s,s';t,t') &= c(s'-s;t'-t) \qquad (2)\\ &= c(s'-s)\cdot c(t'-t), \qquad (3)\end{aligned}$$

where we model the covariance between location s at time t, and location s' at time t'. Equation (3) means the covariance function has been factored into a spatial and temporal component. It also means the covariance between points depends only on the distance between them, and the parameters are constant across the domain. For the models in this paper, the spatial model follows a Matérn covariance function, and the temporal covariance follows an autoregressive process of order 1, denoted as (AR1).

The independent error term is assumed to be uncorrelated Gaussian noise, where σ^2 needs to be estimated.

$$\epsilon(t,s) \sim \mathcal{N}(0,\sigma^2) \qquad (4)$$

The first extension is to model the wind speed with a mean function dependent on spatial covariates.

Instead of the mean being constant,

$$\mu_0 = \beta_0 \qquad (5)$$

we instead model it as a linear combination of covariates:

$$\mu = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \qquad (6)$$

Where x_i are covariates constant across time points but spatially varying. Initially we looked at including covariates that could be derived from open source information, such as altitude, distance from the sea and land use. However the relationship between land features and mean wind speeds is complex and non-linear, we instead used the means taken from high resolution wind maps, as discussed in section 1. Although these were produced several years ago, assuming mean wind speeds have remained approximately the same we would expect this to more accurately capture local trends.

The second extension I consider is dropping the assumption that the spatial field is stationary, and modelling the covariance as a function of covariates. This can be achieved using the stochastic partial differential equation (SPDE) formulation described in Krainski et al. (2018), which approximates the solution to the Gaussian process across a mesh (see Figures 5 and 6 for examples meshes). This allows the covariance parameters to be regressed on local covariates. The spatial covariance function we use is the Matérn, which can be parameterised in terms of the the spatial range, ρ, which denotes the distance at which the correlation goes approximately below 0.1, and σ, the standard deviation of the spatial process. See Figure 2 for an Matérn covariance function (on the left axis), and equivalent correlation function (on the right axis), for $\rho = 100$ and $\sigma = 2$.

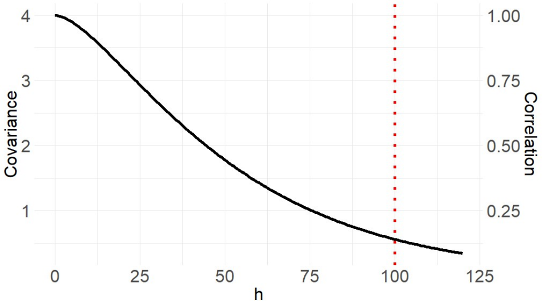


FIGURE 2. Matern covariance function with $\rho = 100$ and $\sigma^2 = 4$.

In the SPDE approach, we can allow ρ and σ to be written as a function of covariates:

- $\ln(\rho) = \theta_0^{(\rho)} + \sum_{i=1}^{n} \theta_i^{(\rho)} b(s)$
- $\ln(\sigma) = \theta_0^{(\sigma)} + \sum_{i=1}^{n} \theta_i^{(\sigma)} b(s)$,

where each θ_i is 0 for the stationary case. Examples of covariates can be similar to mean covariates for example the altitude or distance from the sea could affect how correlated two locations are. For the model in the result section, one covariate for each parameter is used, distance from the sea.

The results section contains results when tested at an hourly resolution. However as these models can be predicted at the same resolution as Met station data, it will allow for the creation of sub hourly, sub kilometre wind maps which current datasets don't contain, therefore allowing wind industry stakeholders to better estimate the potential at prospective sites.

3 Results

In order to test the various models, a leave one out approach is used, where a single met station is predicted from the remaining met stations. We apply this to a month of data, January 2024, recorded at hourly time points. All the models are fit using the INLA package in R. The accuracy of each model is compared using two metrics. The mean prediction is compared to the true observation using a mean absolute error (MAE) metric. As INLA provides a prediction distribution, we also include the percentage of true values that fall within the 95% prediction interval. The results for each error metric is shown in the following table:

TABLE 1. Results in held out Met station.

Model	**MAE**	**Contained in 95% prediction interval**
Stationary	1.92	80%
Varying Mean	1.52	65%
Varying Mean and Covariance	1.80	93%

The wind map produced by the stationary spatio-temporal model is display in Figure 3. We can observe from this that the prediction is overly smooth, as it does not incorporate local topography. In Figure 4 the model produced by a spatially varying mean is shown. This better captures local patterns.

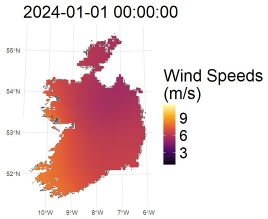


FIGURE 3. Wind Map at a single time point using stationary mean and covariance.

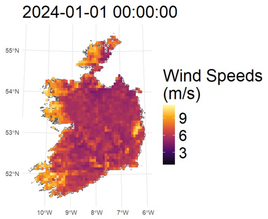


FIGURE 4. Wind Map at a single time point using spatially varying mean.

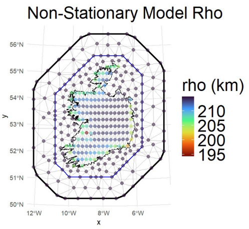


FIGURE 5. The values of ρ, the spatial range across the country.

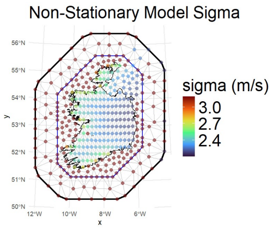


FIGURE 6. The values of σ, the spatial standard deviation across the country.

For the spatially varying covariance function, the ρ and σ parameters that vary across the mesh as a function of distance from the sea are shown in Figures 5 and 6. From this we can see coastal areas have higher variance and lower spatial range. That being said this model results in higher MAE, albeit with a larger percentage of predictions in the 95% prediction interval. This would suggest it better captures the uncertainty, especially in coastal areas, but perhaps causes overfitting, which may need to be investigated further.

References

Doddy Clarke, E., Griffin, S., McDermott, F., Correia, J.M. and Sweeney, C. (2021). Which reanalysis dataset should we use for renewable energy analysis in Ireland?. *Atmosphere*, **12**, 624.

Ingebrigtsen, R., Lindgren, F. and Steinsland, I. (2014). Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, **8**, 20 – 38.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F. and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA.* Chapman and Hall/CRC.

Lenzi, A. and Genton, M.G. (2020). Spatiotemporal probabilistic wind vector forecasting over Saudi Arabia. *The Annals of Applied Statistics*, **14**, 1359 – 1378.

Bayesian analysis of restricted mean survival time adjusted on covariates using pseudo-observations

Léa Orsini[1,2], Emmanuel Lesaffre[2], Guosheng Yin[3], Caroline Brard[4], David Dejardin[5], Gwénaël Le Teuff[1]

[1] CESP, INSERM U1018, Université Paris-Saclay, UVSQ, Villejuif, France
[2] I-Biostat, KU-Leuven, Leuven, Belgium
[3] Departement of Mathematics, Imperial College London, London, England
[4] Ipsen Innovation, Clinical Development Organisation, Les Ulis, France
[5] Product Development, Data Sciences, F. Hoffmann-La Roche AG, Basel, Switzerland.

E-mail for correspondence: `lea.orsini@gustaveroussy.fr`

Abstract: The difference in restricted mean survival time (dRMST) at a specific time point is an appropriate measure to quantify the treatment effect between two arms in randomized clinical trials (RCTs) when the proportional hazards (PH) assumption does not hold. This situation is common with immuno-oncology therapies. Several frequentist methods exist to estimate RMST adjusted on covariates based on modeling and integrating the survival function. A more natural approach is to consider a regression model on the RMST directly using pseudo-observations, which allows for a direct fit without modeling the survival function. Only two Bayesian methods exist, and both model the survival function with a nonparametric prior process. We developed a new Bayesian method based on pseudo-observations and the generalized method of moments (GMM) that offers RMST estimation adjusted on covariates without the need to model the survival function, making it attractive compared to existing Bayesian methods. A simulation study of 2-arms RCTs with different time-dependent treatment effects and covariates effects was conducted, showing that this new approach provides consistent results with existing methods, and improved precision after covariates adjustment. For illustration, the methods were applied to the Getug-AFU 15, a phase 3 trial in non-castrate metastatic prostate cancer.

Keywords: Bayesian survival analysis; Non-proportional hazards; Restricted mean survival time; Generalized method of moments; Pseudo-observations.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

1 Methods

Suppose that $\tilde{T}_i$ the time-to-event variable for the i-th subject, Z_i a p-dimensional baseline covariate vector, A_i the treatment allocation variable, and C_i a right censoring random variable, independent of $\tilde{T}_i$, Z_i and A_i. We observe $T_i = \min(\tilde{T}_i, C_i)$ and $\Delta_i = I(\tilde{T}_i \leq C_i)$ the event indicator. For a pre-specified time point of interest τ, the τ-RMST is defined as

$$\text{RMST}(\tau) = E(\min(\tilde{T}, \tau)) = \int_0^\tau S(t)dt.$$

To adjust the RMST estimation on covariates, the following regression model can be considered

$$\mu_i = E(\min(\tilde{T}_i, \tau)|A_i, Z_i) = g^{-1}(\alpha + \delta A_i + \beta_1 Z_{i1} + \cdots + \beta_p Z_{ip}),$$

where $g(\cdot)$ is a monotone differentiable link function and
$\beta = (\alpha, \delta, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ the vector of unknown parameters. With an identity link function, the regression coefficient δ, can be interpreted as the dRMST between the two arms of a RCT. In the frequentist framework, this model can be fitted using estimating equations, while censoring must be handled, for example, by using the pseudo-observations approach, see Andersen et al. (2004).
This paper extends the latter method to the Bayesian framework. Following Andersen et al. (2004), the i-th pseudo-observation is computed as

$$y_{\tau,i} = n\int_0^\tau \widehat{S}(t)dt - (n-1)\int_0^\tau \widehat{S}^{-i}(t)dt,$$

with n the sample size, $\widehat{S}(t)$ the Kaplan-Meier (KM) estimator of the survival probability, and $\widehat{S}^{-i}(t)$ the KM estimator excluding the i-th subject. Because of the unbiasedness of pseudo-observations conditional on covariates proved in Overgaard et al. (2017), we can replace the non-observed (due to censoring) $\min(\tilde{T}_i, \tau)$ by $y_{\tau,i}$ in the regression model.
The Bayesian generalized method of moments (GMM) is used to estimate the posterior distribution $p(\beta|y_\tau) \propto \tilde{L}(\beta|y_\tau)p(\beta)$ where the pseudo-likelihood $\tilde{L}(\beta|y_\tau)$ is defined following Yin (2009) as

$$\tilde{L}(\beta|y_\tau) \propto \exp\{-\frac{1}{2}U_n^{\mathrm{T}}(\beta)\Sigma_n^{-1}(\beta)U_n(\beta)\},$$

where

$$\Sigma_n(\beta) = \frac{1}{n^2}\sum_{i=1}^n u_i(\beta)u_i^{\mathrm{T}}(\beta) - \frac{1}{n}U_n(\beta)U_n^{\mathrm{T}}(\beta)$$

is a $(p+2) \times (p+2)$ matrix with $u_i(\beta) = \frac{\partial \mu_i}{\partial \beta}(y_{\tau,i} - \mu_i)$ and
$U_n(\beta) = \frac{1}{n}\sum_{i=1}^n u_i(\beta)$.

2 Simulation study

A simulation study of 2-arms RCTs was conducted to assess the performance of the Bayesian GMM with pseudo-observations, compare them with other frequentist and Bayesian RSMT estimators (Andersen et al. (2004), and Zhang and

Yin (2023)), and evaluate the impact of covariate adjustment. The event times were simulated following a Weibull distribution, with scale and shape parameters chosen to mimic different patterns of treatment effect (Figure 1: PH (scenario 1), non-PH with early effect (scenarios 2 and 4), and delayed effect (scenarios 3 and 5), with additional covariates drawn from a uniform distribution (scenario 4) or normal and binomial distributions (scenario 5). In all scenarios, 30% of censoring was considered, drawn from a uniform distribution and an administrative censoring at 8 years, the restriction time τ was set to 5 years, and 1000 replicates were generated. Noninformative priors $N(0, \sqrt{10}^2)$ were specified for all parameters of the Bayesian GMM with pseudo-observations, and a mixture of Dirichlet processes prior was applied under an exponential base measure with Gamma $\Gamma(0.01, 0.01)$ mixing distribution for the method in Zhang and Yin (2023).

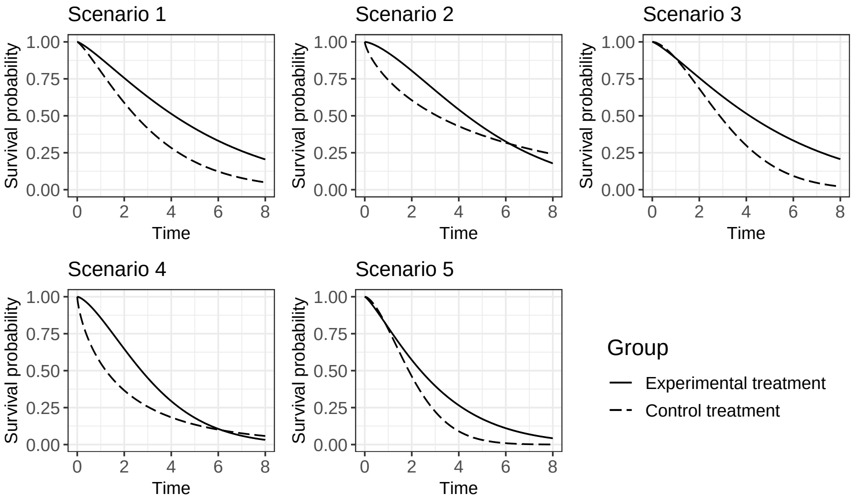


FIGURE 1. Theoretical survival curves for each simulated scenario.

In all scenarios, the Bayesian GMM with pseudo-observations gave valid unadjusted and adjusted dRMST estimations, with similar performance compared to the other methods. The main results are dRMST estimations with covariates adjustment, corresponding to Scenario 4 with $n = 500$ (Table 1). Methods allowing for adjustment on covariates produced slightly more precise estimates after adjustment. Similar results were observed with $n = 200$ with a relative gain in precision from unadjusted to adjusted estimations of 4% (scenario 4) and 8% (scenario 5) for the Bayesian GMM.

3 Real data application

For illustration, we analyzed the data from the Getug-AFU 15, a randomized phase 3 trial comparing an androgen-deprivation therapy (ADT) alone ($n = 193$) or with docetaxel ($n = 192$) in non-castrate metastatic prostate cancer. The median follow-up time was 4.2 years. We focused on the Prostate-Specific Antigen

TABLE 1. Performance of frequentist and Bayesian methods for the estimation of 5-dRMST (difference of 5-RMST between the 2 arms) for scenario 4 (n = 500) representing an early treatment effect on survival with a prognostic uniform variable.

Methods	Bias	ASE[1]	ESE[2]	RMSE[3]	Cov.[4]
Frequentist					
Kaplan-Meier estimator	0.0055	0.163	0.167	0.167	93.9
Andersen et al. (2004)	0.0055	0.163	0.167	0.167	93.9
Andersen et al. (2004)*	0.0037	0.156	0.159	0.159	94.9
Bayesian					
Zhang and Yin (2023)	0.0056	0.162	0.167	0.167	93.9
GMM	0.0044	0.163	0.166	0.166	94.2
GMM*	0.0061	0.156	0.158	0.158	94.9

[1] ASE: Average Standard Error, [2] ESE: Empirical Standard Error

[3]RMSE: Root Mean Square Error, [4] Cov.: 95% Coverage

* Model adjusted on the prognostic variable $Z_1 \sim U([0,2])$

(PSA) progression-free survival endpoint for which the PH assumption was rejected ($p = 0.00022$, Grambsch and Therneau). Without covariates adjustment, all methods yielded similar estimations of the 5-dRMST (data not shown). These results are consistent with the simulation study. After adjustment on four variables (the Gleason score, European Cooperative Oncology Group performance status, concentration of alkaline phosphatase, and presence of bone metastases) with all methods allowing for covariates adjustment, an increase in precision was observed. With the Bayesian GMM, the adjusted 5-dRMST was estimated to be 0.58 (95% CI 0.24 to 0.92) year, meaning that receiving docetaxel in addition to ADT increases the lifetime without PSA progression during the next 5 years by 0.58 year, compared to receiving ADT alone.

4 Discussion

We developed a new Bayesian approach for analyzing RMST using the GMM and pseudo-observations. This method does not require specifying the survival function to estimate RMST adjusted on covariates, making it attractive compared to the existing Bayesian methods. Caution must be taken with the potential misspecification of the model. This method will be extended to the joint analysis of RMST at multiple times.

References

Andersen, P. K., Hansen, M. G. and Klein, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis*, **10**, 335 – 350.

Overgaard, M., Parner, E. T. and Pedersen, J. (2017). Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics*, **45**, 1988 – 2015.

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis*. **4**, 191 – 208.

Zhang, C. and Yin, G. (2023). Bayesian nonparametric analysis of restricted mean survival time. *Biometrics*. **79**, 1383 – 1396.

A mathematically tractable model for information diffusion between communities

David J.P. O'Sullivan[1], Caroline Pena [1], Alina Dubovskaya[1,2]

[1] Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland

[2] Department of Psychology, University of Limerick, Limerick, Ireland

E-mail for correspondence: David.OSullivan@ul.ie

Abstract: In this paper, we build a model of information spread across networks with community structures. This allowed us to estimate new quantities of interest, such as the probability of reinforcing content to a community when it has previously stopped spreading, which show good agrement with simulation results.

Keywords: Information spread; Branching processes; Network science.

1 Introduction

Online social networks such as Facebook, Instagram, TikTok, and the platform formerly known as Twitter, serve as media for the spread of information among their users, where the users both create and share content with each other. Understanding how information spreads on social networks is of paramount importance for society (Keating et al. 2023). Given the ubiquitous use of such platforms for the dissemination of information, understanding how information spreads and is adopted is crucial. In this conference paper, we show how a Multitype Branching Process model can be used to shed light on the interplay between community structure and information spread. We model the diffusion of information using the popular independent cascade model (ICM), where node infections occur in discrete time. Nodes can be in three states: inactive, active, or removed. Active nodes attempt to activate their network neighbours once, before becoming removed themselves in the next time step. The process continues until there are no active nodes to carry on the process. In this work, we will use probability-generating functions (PGFs) to capture this stochastic process, where we focus on the offspring distributions in each community and between communities. PGFs are particularly useful in deriving expressions for the probability of extinction, hazard function, and new quantities, such as the reintroduction probability, which would be hard to analytically calculate otherwise.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

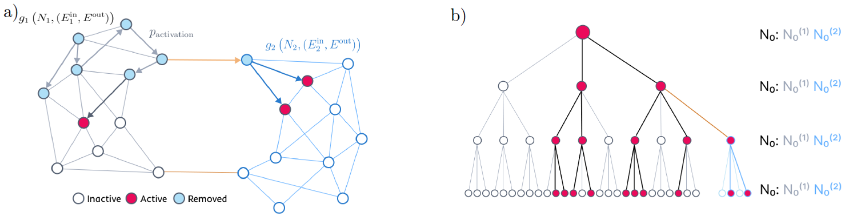

a)
$g_1\,(N_1,(E_1^{\text{in}},E^{\text{out}}))$
$p_{\text{activation}}$
$g_2\,(N_2,(E_2^{\text{in}},E^{\text{out}}))$
○ Inactive ● Active ● Removed
b)
N_0: $\text{N}_0^{(1)}$ $\text{N}_0^{(2)}$
N_0: $\text{N}_0^{(1)}$ $\text{N}_0^{(2)}$
N_0: $\text{N}_0^{(1)}$ $\text{N}_0^{(2)}$
N_0: $\text{N}_0^{(1)}$ $\text{N}_0^{(2)}$

FIGURE 1. a) Schematic illustration of the model. The network consists of two communities where we assume the degree of distribution inside and between the communities is Poisson. The process starts with a single active node in community one. b) Schematic representation of the multi-type branching process model.

2 The MTBP model for information spread

The goal is to construct a PGF for the random variable tracking the number of nodes active in each community at time t, which we denote $N(t)$. Let us consider networks that correspond to the classic *Stochastic Block Model* (SBM) for networks with communities. For simplicity, we assume that both communities have the same statistical properties. For a sufficiently large network, the degree of each node will be bivariate Poisson, where we assume the rate is λ_{in} and λ_{out} for the degree inside and between the communities, respectively, with $\lambda_{\text{in}} > \lambda_{\text{out}}$. A bivariate PGF allows us to write the probabilities as the coefficients of a power series, i.e., $G_{\boldsymbol{X}^{(1)}}(s_1, s_2) = \sum_{n,m=0}^{\infty} P[X_1^{(1)} = n, X_2^{(1)} = m]s_1^n s_2^m$, where $G_{\boldsymbol{X}^{(1)}}$ tracks the probability of having n and m activated offspring from a single node in community 1, respectively. Usefully, the Poisson distribution has a closed-form PGF. Additionally, assuming that each inactive neighbour of a node has an *i.i.d* probability of being activated, ρ, we can derive simple expressions for the offspring distribution for each community as

$$G_{\boldsymbol{X}^{(1)}}(s_1, s_2) = e^{\rho\lambda_{in}(s_1-1)}e^{\rho\lambda_{out}(s_2-1)}, \text{ and} \tag{1}$$

$$G_{\boldsymbol{X}^{(2)}}(s_1, s_2) = e^{\rho\lambda_{in}(s_2-1)}e^{\rho\lambda_{out}(s_1-1)}. \tag{2}$$

Let $N_1(t)$ be the random variable for the number of active nodes in community 1 at generation t and $N_2(t)$ be the number of active nodes in community 2. We introduce the probability generation function $G_{\boldsymbol{N}(t)}$ for $\boldsymbol{N}(t) = (N_1(t), N_2(t))$ as the iteration:

$$G_{\boldsymbol{N}(t)}(s_1, s_2) = G_{\boldsymbol{N}(t-1)}\left(G_{\boldsymbol{X}^{(1)}}(s_1, s_2), G_{\boldsymbol{X}^{(2)}}(s_1, s_2)\right). \tag{3}$$

Setting the initial condition $G_{\boldsymbol{N}(0)}(s_1, s_2) = (s_1)^1(s_2)^0$ corresponding to a single active individual in community one.

3 Estimating probabilities

Our PGF approach allows us to straightforwardly estimate common survival analysis probabilities, such as the extinction probability and hazard functions, but also, as we will see, more complex quantities, such as the probability of reintroducing an infection once it has ceased to spread in a community. This is all accomplished by simple function iteration, and setting the values of s_1 and s_2 to particular values. Let us first calculate the *extinction probability*, the probability that the process becomes extinct by generation t in a community. We can find the probability of the process going extinct by generation t only in community 1 as $q_1(t) = P[N_1(t) = 0] = G_{\boldsymbol{N}(t)}(s_1 = 0, s_2 = 1)$. To see this, note that setting $s_1 = 0$ removes all probabilities associated with having a positive number of active nodes in community 1 from the PGF, and $s_2 = 1$ marginalizes over all the probabilities associated with community 2.

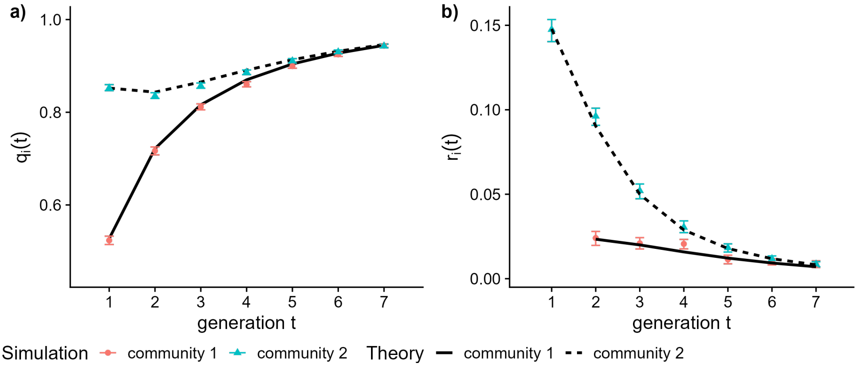


FIGURE 2. a) Extinction probability for stochastic block model network; here $\lambda_{\text{in}} = 8, ; \lambda_{\text{out}} = 2$ and $\rho = 0.06$; b) Probability of the spreading being reintroduced into a community once it has stopped spreading. 95% bootstrapped confidence intervals from the simulations are included.

We can go further and calculate what we will call *community-specific hazard functions* for community 1 and 2 as $\tilde{h}_1(t)$ and $\tilde{h}_2(t)$, respectively. They are hazard functions specified for each community, i.e., $\tilde{h}_1(t) = P[N_1(t) = 0 | N_1(t-1) > 0, N_2(t-1) \geq 0]$. The PGF iteration required is a little more complex and omitted for brevity. However, once we can calculate these community-specific hazard rates, $h_1(t)$, they allow us to estimate crucial probabilities when studying stochastic diffusion processes in the presence of community structure on networks. For example, the probability of reintroducing a pathogen to a community once it has ceased to spread in that community. We denote two new quantities, $r_i(t)$ $(c_i(t))$ the reintroduction (recurrent extinction) probability for community i, which are defined as

$$r_i(t) = 1 - P[N_i(t) = 0 | N_i(t-1) = 0] = 1 - c_i(t). \tag{4}$$

If we note its relation to the extinction probability $q_i(t) = P[N_i(t) = 0]$, where this can be written as $P[N_i(t) = 0 | N_i(t-1) = 0] P[N_i(t-1) = 0] + P[N_i(t) = 0 | N_i(t-1) > 0] P[N_i(t-1) > 0]$, whose terms we have already encountered earlier allowing us to rewrite it as $c_i(t) q_2(t-1) + \tilde{h}_i(t) \left[1 - q_i(t-1)\right]$. We can then isolate the recurrent extinction probability as

$$c_i(t) = \frac{q_i(t) - \tilde{h}_(t) \left[1 - q_i(t-1)\right]}{q_i(t-1)}.$$

Which is not only very compact but also allows us to calculate this via function iteration and the appropriate marginalization.

4 Results and conclusions

We plot the theoretical estimated probability of extinction and reintroduction against simulated values with 95% bootstrapped confidence intervals in Fig. 2. We can see that we have excellent agreement compared to a simulated branching process assuming the same network structure. We accomplished all this under the assumption of a simple SBM for the community structure; it is also worth noting that if we know the connectivity structure inside and between two communities, we can easily extend the analysis to capture much more realistic network typologies.

Acknowledgments: Science Foundation Ireland [18/CRT/6049 to C.P.]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Keating, L. A., Gleeson, J. P. and O'Sullivan, D. J.P. (2023). A generating-function approach to modelling complex contagion on clustered networks with multi-type branching processes. *Journal of Complex Networks*, **11**, cnad042.

Gene coexpression analysis with Dirichlet mixture model: accelerating model evaluation through closed-form KL divergence approximation using variational techniques

Samyajoy Pal[1], Christian Heumann[1]

[1] LMU Munich, Germany

E-mail for correspondence: `Samyajoy.Pal@stat.uni-muenchen.de`

Abstract: Gene coexpression analysis poses unique challenges, particularly in clustering normalized gene profiles where dedicated algorithms are lacking. Compositional in nature, normalized gene profiles find a fitting solution in the Dirichlet Mixture Model (DMM). This study pioneers the application of DMM for clustering normalized gene profiles, recognizing the necessity for efficient model evaluation. Central to this evaluation is the Kullback-Leibler (KL) Divergence, a critical metric for DMMs. In addressing the computational challenges associated with KL Divergence in DMMs, we introduce a novel variational approach. This method provides a closed-form solution, markedly improving computational efficiency for rapid model comparisons and robust estimation evaluations. Through validation on real and simulated data, our approach demonstrates superior efficiency and accuracy compared to traditional Monte Carlo-based methods. This innovation opens new frontiers for expeditious exploration of diverse DMM models, propelling advancements in the statistical analysis of compositional gene expression data.

Keywords: Gene coexpression; Dirichlet mixture model; Normalized gene profiles; Kullback-Leibler divergence; Variational approach.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

State-space models for clustering of compositional trajectories

Andrea Panarotto[1], Manuela Cattelan[1], Ruggero Bellio[2]

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

E-mail for correspondence: andrea.panarotto@phd.unipd.it

Abstract: Compositional data are drawing increasing interest for their ability to depict interdependent and constrained observations. While time series analysis has sometimes been employed for the study of individual compositional trajectories, little attention has been given to finding and modeling groups of trajectories. Driven by a sustainable mobility motivation, we propose a model-based approach, relying on a state space model representation and an Expectation-Maximization algorithm, for clustering compositional trajectories according to their evolution in the simplex. Trajectory covariates, not captured by the compositional representation, can be included in the component weights in a mixture of experts fashion. The method is applied to urban movement data, where people's movements are represented in the simplex by the proportions of road types in their surroundings.

Keywords: Compositional data; Expectation-Maximization algorithm; Mixture of experts model; Model-based clustering; State-space models.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Random projections for classification with high-dimensional data

Roman Parzer[1], Peter Filzmoser[1], Laura Vana-Gür[1]

[1] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Austria

E-mail for correspondence: roman.parzer@tuwien.ac.at

Abstract:
We examine the binary classification problem in a challenging high-dimensional setting with correlated predictors, where the coefficients in a logistic regression model can vary from sparse to dense. In this work, we propose the use of a data-driven random projection matrix to reduce the original feature space. The random projection combines variables considering their respective effect on the class response and such that the regression coefficient can still be recovered. In a simulation exercise, we show that the proposed random projection produces significantly better prediction results than conventional random projections, even outperforming benchmarks such as `glmnet`'s logistic regression with elastic net penalty.

Keywords: High-dimensional classification; Dimension reduction; Random projection.

1 Introduction

Over the last decades, rapid technological progress has contributed to a multitude of classification tasks arising in high-dimensional data settings.
High-dimensional data, where the number of variables p exceeds the number of observations n, i.e., where $p > n$ or even $p \gg n$, pose statistical challenges and many traditional classification algorithms become impractical or are in need to be adapted for the high-dimensional case.
In this work we address the problem of high-dimensional binary classification in a logistic regression model. To deal with the curse of dimensionality, we propose a method which relies on random projections of the data onto a lower-dimensional space. More specifically, a data-informed sparse random projection (RP) is employed to reduce the dimension of the features and a logistic regression model is fit to the resulting reduced predictors. The random projection combines variables

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

considering their respective effect on the class response and such that the coefficient in the logistic regression can still be recovered by the reduced predictors by using a quick approximate estimator of the coefficient.
The use of random projections for the high-dimensional classification problem has also been proposed in Cannings and Samworth (2017), who propose a random projection ensemble classifier for binary data, where a base classifier is applied to many "well chosen" random projections of the data. Furthermore, Xie et al. (2016) combine RP with other dimensionality reduction techniques such as PCA, LDA and feature selection for a binary classification problem in gene expression data and find that combining RP with feature selection provides superior results.
The proposed method in this paper extends the approach in Parzer et al. (2024), which deals with the linear regression setting. The methodology is flexible and can be extended to deal with other generalised linear models.

2 Method

We assume to have a binary response variable $y_i \in \{0, 1\}$, related to p-dimensional predictors $\boldsymbol{x}_i \in \mathbb{R}^p$ via a logistic regression model

$$\mathbb{P}(y_i = 1) = \frac{\exp(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta})}, \quad i \in \{1, \dots, n\} =: [n], \tag{1}$$

with the unknown coefficients $\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p$. In this section, we propose a new random projection matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m \times p}$ with $m \ll p$ tailored to generalised linear regression problems. As in Parzer et al. (2024), we let $h : [p] \to [m]$ be a random map such that for each $j \in [p]$ $h(j) = h_j$ is independently identically distributed as Unif($[m]$). Then, we let $\boldsymbol{B} \in \mathbb{R}^{m \times p}$ be a binary matrix with $\boldsymbol{B}_{h_j,j} = 1$ for all $j \in [p]$ and remaining entries 0, where we assume rank$(\boldsymbol{B}) = m$. Finally, with a diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{p \times p}$ with entries $d_j \sim \text{Unif}(\{-1, 1\}), j \in [p]$, independent of h, one can set $\boldsymbol{\Phi} = \boldsymbol{BD}$ to define a sparse random projection.
When using this random projection for the logistic regression problem (1), variables mapped to the same dimension $k \in [m]$ should not have signs conflicting their respective influence on the response, and, in general, we would wish for $\boldsymbol{\beta} \in \text{span}(\boldsymbol{\Phi}^\top)$ such that the true coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ can be recovered by the reduced predictors $z_i = \boldsymbol{\Phi}\boldsymbol{x}_i$. Both can be accomplished by setting the diagonal elements of $\boldsymbol{D}$ proportional to the coefficient $\boldsymbol{\beta}$, instead of simply picking random signs.
Parzer et al. (2024) show that in the regression setting the HOLP (High-Dimensional Ordinary Least Squares Projection, Wang and Leng 2016) estimator is a viable option to use as the diagonal elements. It is defined as the limit of the L_2-penalised least squares estimator for penalty $\lambda \to 0$ and has an explicit form.
We choose to take the same approach here and use the maximum arguments of the L_2-penalised log-likelihood for a small threshold $\lambda > 0$

$$\hat{\boldsymbol{\beta}}_{L_2} := \text{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log\Big(\frac{1}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}\Big) + \frac{\lambda}{2} \sum_{j=1}^p \boldsymbol{\beta}_j^2 \tag{2}$$

as the diagonal elements $d_j = \hat{\boldsymbol{\beta}}_{L_2,j}$. In the high-dimensional setting $p > n$, $\hat{\boldsymbol{\beta}}_{L_2}$ might diverge for $\lambda \to 0$, contrary to the regression case. Therefore, some penalisation is necessary. In practice, we calculate $\hat{\boldsymbol{\beta}}_{L_2}$ with $\lambda = 10 \cdot \max_j (\sum_{i=1}^n y_i \boldsymbol{x}_{ij})$

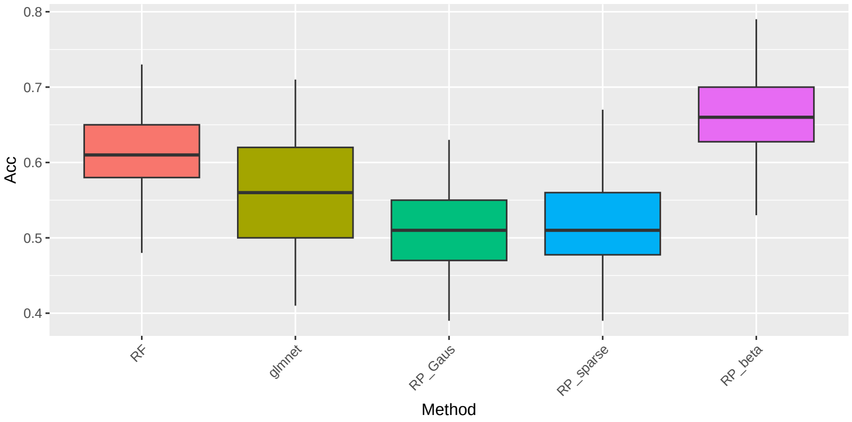


FIGURE 1. Comparison of prediction performance of random forest, `glmnet`, different conventional projections and our proposed projection, for 100 replications of the described setting with $n = 200, p = 2000$ and $\mathbf{\Sigma}_{ij} = 0.9^{|i-j|}$.

(for standardised predictors), which is also the smallest λ used by the R-function `glmnet`. Our proposed algorithm can be summarised as (i) compute $\hat{\boldsymbol{\beta}}_{L_2}$, (ii) compute $\mathbf{\Phi}$ and reduced predictors $z_i = \mathbf{\Phi}\boldsymbol{x}_i$ and (iii) estimate the model on the reduced predictors, where we employ an elastic net penalty to ensure that the model achieves a certain degree of sparsity and that the issue of separability is taken care of.

3 Preliminary simulation results

In this section, we want to demonstrate the effectiveness of this adapted random projection to improve prediction accuracy. We generate data from (1) with multivariate normal predictors $\boldsymbol{x}_i \sim N(0, \mathbf{\Sigma})$, where we choose $n = 200, p = 2000$, and $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}$ has an $AR(1)$ structure with $\rho = 0.9$.
The coefficient $\boldsymbol{\beta}$ has $a = n/2 + 2log(p)$ random entries bounded away from zero at uniformly drawn positions and is scaled such that $\text{Var}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta} = 9$, which will determine the separation between the two classes. Finally, β_0 is chosen such that $\sum_{i=1}^n \mathbb{P}(y_i = 1) = n/2$.
We compare with random forests (RF, using the R-package `randomForest`), `glmnet`, and our approach with the following three different random projection matrices $\mathbf{\Phi}$. Firstly, we use a conventional random projection $\mathbf{\Phi}$ with iid Gaussian entries (RP_Gaus), then we use the sparse version introduced above with random sign diagonal elements (RP_sparse) and, finally, our proposed version with the adapted diagonal elements (RP_beta).
Figure 1 shows the average accuracies for predicting 100 new test observations over 100 replications. We can see that the two conventional random projections RP_Gaus and RP_sparse are only slightly better than random guessing in this challenging high-dimensional setting, but our proposed adapted random projection leads to a huge improvement in accuracy, even outperforming `glmnet` and random forests.

4 Discussion

We propose the use of a data-driven random projection in a logistic regression problem. The random projection relies on a regularised estimator for $\boldsymbol{\beta}$. We will further investigate how the choice of this estimator influences the performance of the method. As next steps, similar to Parzer et al. (2024) we analyse how the performance of the proposed method changes when i) introducing a variable screening step, ii) performing the analysis for a collection of random projection and then averaging over this ensemble, and iii) introducing a threshold for sparsity and variable selection.
Finally, further possible extensions include accommodating responses from any generalised linear model in the proposed methodology.

References

Cannings, T.I. and Samworth, R.J. (2017) Random-projection ensemble classification. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **79**, 959 – 1035

Parzer, R., Vana-Gür, L. and Filzmoser, P. (2024) Sparse data-driven random projection in regression for high-dimensional data. *arXiv:*, 2312.00130.

Wang, X. and Leng, C. (2016) High-dimensional ordinary least-squares projection for screening variable. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **78**, 589 – 611

Xie, H., Zhang, Q. and Wang, Y. (2016) Comparison among dimensionality reduction techniques based on Random Projection for cancer classification. *Computational Biology and Chemistry*, **65**, 165 – 172

Shrinkage in a Bayesian panel data model with time-varying coefficients

Roman Pfeiler[1], Helga Wagner[1]

[1] [1]Johannes Kepler University, Linz, Austria

E-mail for correspondence: `roman.pfeiler@jku.at`

Abstract: We consider regression models for panel data, where regression effects and within subject dependence are allowed to vary over time. We adopt a Bayesian approach with priors that allow shrinkage to constant and zero effects as well as to simpler dependence structures. The model is evaluated in a simulation study and applied to the analysis of yearly earnings of mothers in Austria who returned to the labour market after maternity leave.

Keywords: Shrinkage prior; Factor model; Random intercept; Income dynamics.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Flexible additive models for multi-event survival analysis

Johannes Piller[1,2], Helmut Küchenhoff[1], Andreas Bender[1,2]

[1] Statistiches Beratungslabor (StaBLab), LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Munich, Germany

E-mail for correspondence: `johannes.piller@stat.uni-muenchen.de`

Abstract: The Piecewise Exponential Additive Mixed Model (PAMM) has become a popular method for complex modelling of single-event survival data. Here, we extend the framework and the relating R package `pammtools` to event-history analysis, i.e. competing risks, recurrent events and multi-state settings.

Keywords: Survival analysis; Time-to-event; Piecewise exponential; Multi-state; Competing risks.

1 Introduction

Piecewise Exponential Additive Mixed Models (PAMMs) (Bender et al., 2018) have gained popularity in various domains due to their ability to tackle a wide variety of survival problems and their flexibility to model non-linear covariate effects, including time-varying effects and cumulative effects (Bender et al., 2019). One advantage of such reduction techniques is that they do not require any specialised software for the estimation of the model parameters. Thus, in the case of the PAMM, they can be conveniently estimated using generalized additive mixed modeling methodology or, for example, respective boosting or deep learning based approaches (Bender et al., 2022). Nevertheless, their use in practice requires pre-processing, which differs depending on the survival task at hand (e.g. left-truncation, competing risks, etc.) and post-processing (e.g. transforming estimated parameters to useful quantities like survival or transition probabilities). The R package `pammtools` facilitates the entire modeling process, so far, however, only for single-event data. Here we extend the framework and package capabilities to handle general multi-state models.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

2 Methods

We consider the general multistate setting with observed data

$$(y_{i,k,e}^{entry}, y_{i,k,e}^{exit}, \delta_{i,k,e}, \mathbf{x}_{i,k,e}),$$

where $y_{i,k,e}^{entry}$ is the entry time of subject $i = 1, \ldots, n$ into the risk set for transition $k = 1, \ldots, q$ in episode $e = 1, \ldots, m$ and $y_{i,k,e}^{exit}$ the respective exit time, either due to the transition occurring or censoring. Here, $\delta_{i,k,e}$ is the corresponding status indicator and $\mathbf{x}_{i,k,e}^{\top} = (x_{i,k,e,1} \cdots x_{i,k,e,p})$ the covariate row-vector, which, besides subject specific information, can contain information about the past (e.g. number and timing of past transitions and number and timing of previous episodes). Note that we need both indices, k and e, as, in the general case, some transitions could occur more often, for example in case of recurrent events or back-transitions. The transitions are modeled via log-hazard rates

$$\log(h_{k,e}(t|\mathbf{x}_{i,k,e}, \ell_i)) = \beta_{0,k,e} + f_{0,k,e}(t) + \sum_{p=1}^{P} f_p(x_{i,k,e,p}, t) + b_{\ell_i}. \qquad (1)$$

In Eq. (1) the log-baseline is given by $\log(h_{0,k,e}(t)) = \beta_{0,k,e} + f_{0,k,e}(t)$, $f_p(x_{i,k,e,p}, t)$ are potentially non-linear, potentially time-varying effects of covariates and b_{ℓ_i} potential random effects for cluster ℓ to which subject i belongs to. Eq. (1) could be further extended to include stratification of the baseline hazard according to subsets of subjects, interactions between different covariates (e.g. via tensor products), cumulative effects of time-dependent covariates, and more complex random-effect structures, but this is omitted here for simplicity.

Using PAMMs, we estimate Eq. (1) by splitting the follow-up into J intervals, with intervals $(\kappa_{j-1}, \kappa_j]$, $j = 1, \ldots, J$, transforming the raw-data accodringly and estimating the interval-specific hazard rates $h_{k,e}(t|\mathbf{x}_{i,k,e}) = h_{k,e,j}(\mathbf{x}_{i,k,e})$, for all $t \in (\kappa_{j-1}, \kappa_j]$. Non-linear functions are parameterized as $f_{p,k,e}(x_{i,k,e,p}) = \sum_{g=1}^{G} \gamma_{p,g} B_{p,g}(x_{i,k,e,p})$ with basis coefficients $\gamma_{p,g}$ and suitable basis functions $B_{p,g}$ (e.g. B-splines). The respective parameters are estimated by maximizing the corresponding penalized Poisson likelihood (Wood, 2020) or other suitable estimation techniques. Note that a separate smooth baseline hazard is estimated for each transition and episode, but could also be reduced to transition specific baseline hazard rates. Once the transition specific hazard rates in Eq. (1) are estimated, the transition probabilities are obtained via the empirical transition probability matrix (Beyersmann et al., 2011). To calculate the empirical probability matrix, first, we discretize the time component $t \in \mathcal{T}$, with $\mathcal{T} = \{T \in \mathbb{N} : t_1 < \ldots < t_T\}$, i.e. t stems from a grid of ordered time points. Second, we integrate the specific hazard rates and obtain the cumulative transition hazards, i.e. $H_{k,e}(t|\mathbf{x}) = \int_0^t h_{k,e}(u|\mathbf{x})du$. Next, we define transition matrices, which contain the transition probabilities. Let Q be the set of all states for each transition k and $Q_l \subseteq Q, l \in Q$ the set of all possible states after transitioning from l, then there exists a tuple $(l, o) \in Q \times Q_l$, describing the k-th transition from state l to state o, which, in the following, is denoted by $l \to o$. The transition matrix is then given by a finite matrix product over all event times t and matrices $\mathbf{I} + d\hat{\mathbf{H}}(t|\mathbf{x})$, with entries $d\hat{H}_{l\to o,e}(t|\mathbf{x}) = \hat{H}_{l\to o,e}(t|\mathbf{x}) - \hat{H}_{l\to o,e}(t - |\mathbf{x})$, i.e.

$$\prod_{t\in\mathcal{T}} \left(\mathbf{I} + d\hat{\mathbf{H}}(t|\mathbf{x})\right) \qquad (2)$$

Since the rows of probability matrices must sum up to one, the diagonal elements of $\hat{\mathbf{H}}(t|\mathbf{x})$, are defined to be $\hat{H}_{l\to l,e}(t|\mathbf{x}) = -\sum_{o\in Q(l)} \hat{H}_{l\to o,e}(t|\mathbf{x})$.
Using PAMMs, we estimate Eq. (2) by applying the same discretization and data transformation for the interval- and transition-specific hazard rates Eq. (1) and calculate

$$H_{k,e}(t|\mathbf{x}) = \sum_{l=1}^{j(t)-1} h_{k,e,j}(\mathbf{x})(\kappa_l - \kappa_{l-1}) \cdot h_{k,e,j(t)}(\mathbf{x})(t - \kappa_{j(t)-1}) \qquad (3)$$

where $j(t)$ the index of the interval for which $t \in (\kappa_{j(t)-1}, \kappa_{j(t)}]$.

3 Simulation

In this section, we illustrate the capabilities of the model via simulation studies. While we focus on simple settings (e.g. to illustrate equivalence to Aalen-Johanson estimator), the model class can be used in more complex settings.
In the first simulation study, we sample an illness-death multi-state setting with 200 subjects and constant transition-specific log-hazards $h_{0\to1} = 0.3, h_{0\to2} = 0.6$, and $h_{1\to2} = 0.5$. Using the functionalities of the `mvga` package, see (Allignol et al., 2008), we calculate the Aalen-Johanson estimator for the simulated multi-state model. Using PAMMs, we estimate the log-hazards and calculate the transition probabilites. Figure 1 shows equivalence of the two approaches.

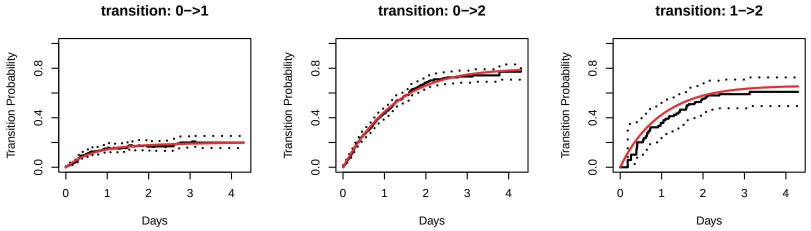


FIGURE 1. Shown are the transition probabilities based on the Aalen-Johanson estimator (solid, black) with confidence bands (dotted, black), and the estimation using `pammtools` (solid, red).

In the second simulation study, we sample an illness-death multi-state setting with non-linear covariate effects and time dependent transition-specific log-hazards given by $h_{0\to1}(t|(x)) = 0.5+0.25x^3$, $h_{0\to2}(t|(x)) = \frac{1}{2^8\cdot\Gamma(8)}t^7e^{-t/2}$, and $h_{1\to2}(t|(x)) = 0.4 - x^2$. Note that the hazards are by construction highly non-linear and also time-dependent. The simulation is built on a data set with 1500 subjects, a covariate x, sampled uniformly from $[-3, 3]$, and 100 repetitions. Using PAMMs, for each repetition, we fitted a model with non-linear effects in `x` and `t`. Finally, we calculated the average and compared the true curve with the average fit. Figure 2 shows that the average fit of the non-linear log-hazards (red) is close to the pre-defined transition-specific log-hazards of the simulation setup.

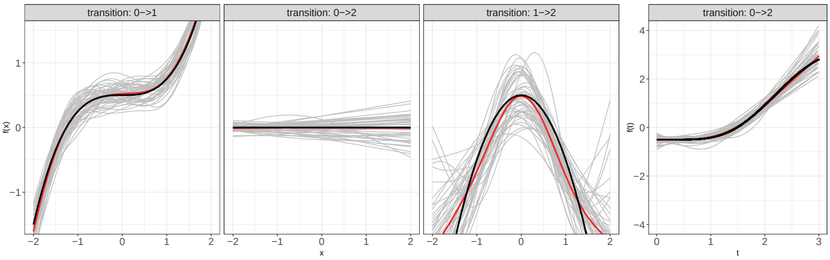


FIGURE 2. The first three facets depict the non-linear effect of `x` for each transition. The last facet depicts the non-linear effect of `t` for $0 \to 2$. Results from each iteration are grey, the average is red, and the true function is black.

4 Application

The presented PAMMs approach enables smooth effects to be used to estimate cumulative hazards and transition probabilities. To illustrate the flexibility, we used the `pammtools` package to analyze the `mgus2` data set in the `survival` package, see (Kyle et al., 2002). The data set contains a classical illness-death setup with possible transitions $0 \to 1$, i.e. progression to a plasma cell malignancy (pcm), $0 \to 2$, i.e. death, or $1 \to 2$, i.e. progression from pcm to death. We estimated the hazards with a linear and a non-linear effect of hemoglobin (`hgb`). Figure 3 visualizes the non-linear effect of `hgb` and the time `tend` on the logarithmic hazard rates for the transition $0 \to 2$ and the transition probabilities for the linear and non-linear model. As can be seen in the left facet of the figure, the change of the amount of `hgb` within $[11, 14]$ in $[g/dl]$ has a stronger decreasing effect on the hazard rates than for amounts within $[9, 11)$ c.p. The overall fit and the fact that the effective degrees of freedom are $4.65 > 1$ undermine the need of the more flexible smooth functions for modeling the effect of `hgb`. Figure 3

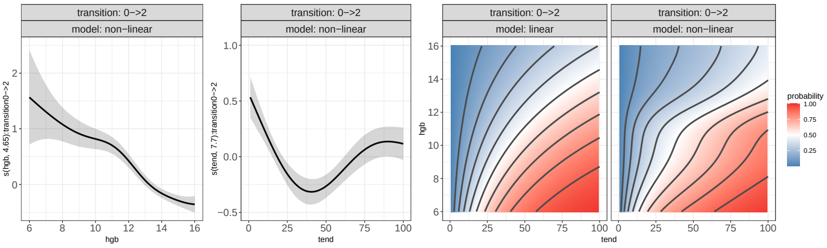


FIGURE 3. For transitions into the death stage, the first facet depicts the non–linear effect of `hgb`, the second depicts the non-linear effect of time, the third and forth facets depict the transition probabilities depending on the time and `hgb`, modelled with a linear effect (third facet) or a non-linear effect (fourth facet).

compares further the influence of the linearly (third facet) and the non-linearly (right facet) modeled hazard rates on the transition probabilities. In the linear case, the transition probabilities change equally over time and different levels of hgb. In the non-linear case, the transition probabilities change according to the structure of the smooth functions in the first and second facet of Figure 3.

5 Discussion

This article illustrates a workflow for reducing complex multi-state tasks to more standard regression tasks using the R package pammtools. All transition hazards and probabilities, including their dependencies on covariates and time, can be estimated in one single model. The additive predictor used to define the transition hazards can be very flexible, including non-linear terms, spatial effects, random effects and their interactions. While this approach was illustrated mainly on non-linearity with a focus on extending single-event survival models to multi-state settings in Sections 3 and 4, the approach will be particularly useful when the Markov assumption may be violated, i.e., when the hazard at time t depends on the past. In this approach, one could model such dependencies by introducing time-varying covariates like number of past transitions or time spent in previous states into the restructured data set, which natively supports dependencies on multiple timescales (Iacobelli et al., 2013). One disadvantage of the data transformation, in particular in the multi-state setting, is the increase of the data set. However, since the estimation problem is reduced to a Poisson regression task, available techniques for efficient estimation in the big data context (Wood et. al., 2017, Reulen et. al., 2015, Sennhenn-Reulen et. al., 2016) are applicable, such that even models with millions of rows and complex predictors are estimable within reasonable time and memory requirements. Nevertheless, further reduction of the transformed data set size is desirable and could be addressed in future iterations of the implementation.

References

Allignol, A., Beyersman, J. and Schumacher, M. (2008). mvna: An R Package for the Nelson–Aalen estimator in multistate models. *R News*, **8**, 48 – 50.

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

Bender, A, Rügamer, D., Scheipl, F. and Bischl B. (2020). A General Machine Learning Framework for Survival Analysis In: *Hutter F, Kersting K, Lijffijt J, Valera I, editors, Machine learning and knowledge discovery in databases*, Lecture Notes in Computer Sciences, Springer International Publishing, 158 – 173.

Bender, A, Scheipl, F., Hartl, W., Day, A.G. and Küchenhoff H. (2019). Penalized estimation of complex, non-linear exposure-lag-response associations. *Biostatistics*, **20(2)**, 315 – 331.

Beyersmann, J., Allignol, A. and Schumacher M. (2011). *Competing Risks and Multistate Models with R.* New York: Springer.

Iacobelli, S. and Carstensen, B. (2013). Multiple time scales in multi-state models. *Statistics in Medicine*, **32**, 5315 – 5327.

Kyle, R., Therneau, T., Rajkumar, V., Offord, J., Larson, D., Plevak, M. and Melton, L. J. (2002). A long-terms study of prognosis in monoclonal gammopathy of undertermined significance. *New England Journal of Medicine*, **346**, 564 – 569.

Reulen, H. and Kneib, T. (2015). Boosting multi-state models. *Lifetime Data Analysis*, **22**, 241 – 262.

Sennhenn-Reulen, H. and Kneib, T. (2016). Structured fusion lasso penalized multi-state models. *Statistics in Medicine*, **35**, 4637 – 4659.

Wood, S.N. (2020). Inference and computation with generalized additive models and their extensions. *TEST*, **29**, 307 – 339.

Wood, S.N., Li, Z., Shaddick, G. and Augustin, N.H. (2017). Generalized additive models for gigadata: Modelling the U.K. black smoke network daily data. *Journal of the American Statistical Association*, **112**, 1199 – 1210.

Latent Dirichlet allocation and hidden Markov models to identify public perception of sustainability in social media data

Luigi Cao Pinna[1], Claire Miller[1], Marian Scott[1]

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

E-mail for correspondence: `luigi.caopinna@glasgow.ac.uk`

Abstract: To help guide a just transition to a sustainable society and onboard the local communities, researchers can identify events of public interest through access to data from community engagement activities and social media content. However, novel analytic methods are required to process and analyse data in unstructured formats (e.g. transcripts, text and images) and to extract useful information for decision-making. This paper proposes an analytics pipeline combining latent Dirichlet allocation and hidden Markov models for automatically detecting multiple latent changepoints in topics over time, without prior knowledge of their occurrence. Analysing social media content (i.e., tweets) related to Glasgow, we identified events that captured social media users' public interest, demonstrating the potential of our method to inform timely and relevant policy making.

Keywords: Sustainable society; Social media content; Latent Dirichlet allocation; Multiple latent changepoints.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

Model selection procedure in multilevel cross-classified latent class models

Nicola Piras[1], Silvia Columbu[1], Jeroen K. Vermunt[2]

[1] Department of Mathematics and Computer Science, University of Cagliari, Italy
[2] Department of Methodology and Statistic, Tilburg University, The Netherlands

E-mail for correspondence: nicola.piras97@unica.it

Abstract: The availability of proper selection criteria is of fundamental importance in the definition of latent class analysis models. When the data structure is multilevel the selection procedure must be applied to each level of the model. In the case of the multilevel cross-classified extension, we propose to apply a three step procedure that takes into account the mutually dependence between the two levels of the structure in the selection. The performances of the method are investigated through simulation studies in which different information criteria are considered. The definition of these criteria are based on approximations of the log-likelihood, which is intractable in such a cross-classified structure.

Keywords: Information Criteria; Multilevel Cross-Classified; Latent Class.

1 Introduction

When performing a latent class (LC) analysis it is crucial to apply model selection criteria for determining the number of latent classes. The decision on that number is affected by different factors (like the sample size and the level of separation between the classes) that must be taken into account in the definition of selection measures. In case of multilevel data, the task of selection gets even more complex as it concerns latent classes at all levels of the data structure. Indeed, the extension of latent class models to multilevel data can be defined taking a set of membership variables for each level. The problem of selection in the hierarchical model has been exhaustively discussed in (Lukociene, Varriale, Vermunt 2010). In this work we focus instead on the cross-classified case, introduced in (Columbu, Piras, Vermunt 2023), in which units are simultaneously nested within two higher level groups (for instance, children belonging to both schools and neighborhoods). Therefore, the selection applies to the number of level-1 latent classes (L) and the number of two separate sets of level-2 latent classes (H and R).

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

We propose to consider a three step approach similar to what done in the hierarchical frame. That is, in a first step we determine the number of lower level classes as in a standard LC, ignoring the multilevel structure. Then we determine the number of higher level classes, taking fixed the number of level-1 latent classes selected at the previous step. Finally, we determine again the number of lower level classes fixing the number of higher level classes at the values we obtained from the second step.

The implementation of a three step procedure enables to take under consideration the mutually dependence between the levels of the data structure. The selection of latent classes at level-1 influences that of level-2, the reverse is also true, however in the second direction the dependency is typically weaker.

In the LC framework, and more in general in mixture modeling, the problem of choosing the right number of classes corresponds to a model selection problem. In this context, information criteria (IC) based on some form of log-likelihood penalization are usually considered. Such statistics are built as a weighted combination between the fit of the model and its complexity, measured in terms of number of free parameters involved. The missing data structure typical of LC, with the presence of unknown membership variables to be estimated, requires the introduction of a complete version of the likelihood. Therefore, the goodness of fit can be measured taking either the log-likelihood (logL) value or the complete log-likelihood (CL). In the extension to the multilevel cross-classified version the definition of IC gets more complex as both levels of the structure must be considered and no finite expression of the data likelihood is available. Indeed, the double missing data structure at the higher level makes the likelihood untractable, implying that approximated versions of it can be computed. We will consider multiple IC statistics by letting vary the penalization terms that are computed considering the number of latent classes and/or the sample size (in a BIC fashion). In addition, for each of them we take two versions, one based on the logL and a second based on the CL. Given the presence of two levels in the data, in the level-2 of the structure the sample size can be intended as the total number of individuals (n) or the number of combinations of level-2 cross classified units (K and Q) present in the data.

2 Three-step procedure

The procedure we consider consists of three steps, and is based on three criteria for level-1 and their correspondent definition for the cross-classified level-2. For each IC we consider also a version with CL, generally denoted as IC_e, and for what concerns the BIC we consider alternatives penalized by the number of level-2 groups within which the observations are nested (BIC_g and BIC_{eg}). The steps taken are:

1) Determine the number of lower level classes L fitting standard LC, without taking into account the multilevel structure. Level-1 information criteria used are:

$$BIC = -2 \cdot logL + ((L-1) + npar) \cdot log(n)$$

$$AIC = -2 \cdot logL + 2 \cdot ((L-1) + npar)$$

$$AIC3 = -2 \cdot logL + 3 \cdot ((L-1) + npar)$$

$npar = \sum_{i=1}^{I} L \cdot (C_i - 1)$ is the number of free distribution parameters with I categorical indicators of C_i categories. The lower the value of an IC, the better the model.

2) Determine the number of higher level cross-classified classes H and R fixing the number of lower level classes to that selected in step 1. Level-2 information criteria used are defined as:

$$BIC = -2 \cdot logL + ((R-1) + (H-1) + (H \cdot R \cdot (L-1)) + npar) \cdot log(n)$$

$$AIC = -2 \cdot logL + 2 \cdot ((R-1) + (H-1) + (H \cdot R \cdot (L-1)) + npar)$$

$$AIC3 = -2 \cdot logL + 3 \cdot ((R-1) + (H-1) + (H \cdot R \cdot (L-1)) + npar)$$

the analogous definition taking CL instead of logL and substituting n with the number of group combinations available in the data can be also considered.

3) Determine again the number of level-1 classes L, setting this time the number of level-2 cross-classified classes to the value selected in step 2. This third step allows to evaluate if the suitable number of lower level classes can change after taking into account the multilevel structure. The same criteria as in step 2 are applied.

TABLE 1. Number of replicates in which the investigated criteria, at step 2 and 3, estimated various combinations of level-1 and level-2 cross-classified latent classes. The correct combination is $(4, 3, 2)$.

		Step 2				Step 3		
IC	n_{kq}	4 2 2	**4 3 2**	4 2 3	4 3 3	3 3 2	**4 3 2**	5 3 2
BIC	6	5	33	50	12	0	100	0
	8	0	20	55	24	0	100	0
BIC_g	6	1	32	50	17	0	100	0
	8	0	14	56	30	0	100	0
AIC	6	0	23	47	30	0	100	0
	8	0	4	48	48	0	100	0
$AIC3$	6	26	47	27	1	0	100	0
	8	0	4	48	48	0	100	0
BIC_e	6	37	60	1	2	86	14	0
	8	16	84	0	0	100	0	0
BIC_{eg}	6	29	66	1	4	78	22	0
	8	8	92	0	0	100	0	0
AIC_e	6	13	72	1	14	78	22	0
	8	0	88	0	12	100	0	0
$AIC3_e$	6	14	73	1	12	78	22	0
	8	0	92	0	8	100	0	0

2.1 Simulation results

To evaluate the performances of the criteria proposed and the three step procedure, we carried out simulations taking dataset with two different separation

conditions, obtained setting different values of level-1 units within each k, q group ($n_{kq} = \{6, 8\}$). For each scenario we generated 100 datasets with six binary indicators, and set $K = 30$, $Q = 12$, $L = 4$, $H = 3$ and $R = 2$.
The results of simulations are summarized in Table 1. We observe that at step 2 the versions of the IC based on the CL are those selecting in higher proportion the right number of level-2 classes, keeping fixed the level-1 classes to the correct one $L = 4$. For step 3, once fixed the number of cross-classified classes to the real one ($H = 3$ and $R = 2$) the measures based on logL give always the correct selection. In particular we highlight that at step 2 $AIC3$ performs better with lower separation, while increasing the separation BIC_{eg} provides also good results. These preliminary results are in line with what found in multilevel hierarchical latent class models.

References

Columbu, S., Piras, N. and Vermunt, J. K. (2023). Multilevel Cross-Classified Latent Class Models. In: *CLADAG 2023 Book of Abstracts and Short Papers, ISBN: 9788891935632*, 390 – 393.

Lukociene, O., Varriale, R. and Vermunt, J., K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, **40**, 247 – 283.

Semi-Markov multistate model with interval-censored transition times

Xavier Piulachs[1], Klaus Langohr[2], Guadalupe Gómez[2]

[1] Polytechnic University of Catalonia, Campus Terrassa, Terrassa, Spain,
[2] Polytechnic University of Catalonia, Campus Nord, Barcelona, Spain

E-mail for correspondence: xavier.piulachs@upc.edu

Abstract: A Cox-based multistate model is proposed for analyzing a multi-cohort event history process with interval-censored transition times. The cohort is included as a stratum variable when modeling each transition hazard, while testing the compliance with the Markov property conditional on the prognostic covariates. Whenever the Markovian assumption does not hold for a given transition, the time of entry into the current state is incorporated in the modeling procedure, yielding a semi-Markov process. To deal with interval censoring, an easy-to-implement procedure is based on performing a multiple imputation of the unknown transition times within the specified intervals. The corresponding artificially completed datasets are separately fitted using the proposed multistate model, each providing inference on the target population quantity. Finally, the overall collected information is properly combined. The described methodology is applied to a three-wave dataset of COVID-19-hospitalized adults.

Keywords: Semi-Markov multistate model; Interval-censored data; Multiple imputation; COVID-19.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Joint models for longitudinal and time-to-event data in social science research

Sophie Potts[1], Anja Rappl[2], Karin Kurz[1], Elisabeth Bergherr[1]

[1] University of Goettingen, Goettingen, Germany
[2] Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany

E-mail for correspondence: sophie.potts@uni-goettingen.de

Abstract: A joint model for longitudinal and time-to-event data is a well-established estimation method in biostatistics because it exploits data richness, handles endogenous covariates in survival models, and controls for non-random dropout in longitudinal studies. However, this method is not yet part of the standard toolkit of social scientists even though the growing mass of longitudinal and survival data sets requires appropriate analysis tools. Therefore, this contribution provides a gentle introduction to the method of joint models and highlights its advantages for social science research questions. We demonstrate its usage and usefulness using an application on marriage dissolution and marital satisfaction and compare the results with classical approaches. In addition to demonstrating the method, our results suggest that shared household work in a marriage has no direct effect on the risk of marital dissolution for women but for men. Further, there is a strong indirect effect of shared household work on the risk of marriage dissolution via marital satisfaction for both genders.

Keywords: Joint models; Longitudinal data; Survival data; Marriage dissolution; Relationship satisfaction.

1 Introduction

Research questions pointing to the timing of an event as well as respective data sets are frequently found in social science research. Using time-varying covariates (TVC) allows to model the impact of changing covariates over time in survival models. Many social science research fields use some version of these time-to-event data analysis such as family formation, educational attainment, recidivism or reemployment.

However, the concept of a time-varying covariate assumes that it does not change between the observation times and is exogenous. Especially for frequently changing, self-reported and person-related variables as they are common in social sciences these assumptions are questionable.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

Both problems, endogeneity and missing data between observation times, are tackled by the statistical concept of joint models for longitudinal and time-to-event data (Wulfsohn & Tsiatis, 1997).
Besides their frequent usage in the field of biostatistics, Cremers et al. (2021) marked that joint models are underused in social sciences. In order to increase the usage of these models and to highlight the advantages and possible applications a low-threshold introduction on how to use these models is needed. Since a joint model represents a useful tool to analyse complex social phenomena data, we illustrate the application of it in the field of social sciences. Therefore this contribution aims to provide an understandable introduction on how to use joint models.
We apply the model to marriage data in order to analyse the relationship of the trajectory of marital satisfaction (longitudinal) and the timing of marriage dissolution (survival). The example is executed using the Software `R`.

2 Joint models for longitudinal and time-to-event data

Joint Models are a class of statistical models that combine a longitudinal outcome and a time-to-event outcome. Hereby, the former TVC is modelled via a linear mixed model (LMM), allowing for intra-individual variance along the time axis (t) captured by random intercepts (b_{0i}) and possibly random slopes (b_{1i}) for each individual i. Joint models consist of two submodels: By incorporating the predictions of a longitudinal model in a survival model, the two models are linked and estimation for both submodels is performed simultaneously. The model can be written as

$$h(t|M_i(t), \boldsymbol{x_i}) = h_0(t)\exp[\boldsymbol{\gamma}^T \boldsymbol{x_{i,\mathrm{surv}}} + \alpha m_i(t)]$$

with $m_i(t) = \boldsymbol{\beta}^T \boldsymbol{x_{i,\mathrm{long}}} + \boldsymbol{b}_i^T \boldsymbol{z_i}$. Besides the classical baseline covariates $\boldsymbol{x_{i,\mathrm{surv}}}$, the estimated value of the TVC $m_i(t)$ enters the model and is equipped by a coefficient α which is called the *association parameter*. Including a covariate in both submodels its direct and indirect effect on the time to event can be separated. By estimating a unique $\hat{\beta}_k$ coefficient as well as a $\hat{\gamma}_k$ coefficient and the association parameter $\hat{\alpha}$ we decompose the total effect via: $\hat{\alpha}\hat{\beta}_k + \hat{\gamma}_k$. Estimation of the coefficients can be done using different estimation strategies, which are presented and compared by Rappl et al. (2021).

3 Data set on marriage satisfaction and time to marriage dissolution

In order to demonstrate the use of joint models in sociology, the relationship between satisfaction with the marriage and the time to marriage dissolution in first marriages is investigated. To the best of our knowledge, no one used a joint model for longitudinal and time-to-event analysis in order to exploit the whole richness of data, i.e. the longitudinal character of the data as well as the information of timing of an event for this use case. As a data base the German pairfam ("Panel Analysis of Intimate Relationships and Family Dynamics") data set (Huinik et al., 2011) is selected. All available waves (2008/09 – 2021/22) are used and the final sample consists of $N = 3,559$ first marriages with at least three time-points during the study. A share of approx. 7% stated an end of the relationship during the observation period (number of events). We did not take the actual month of divorce as event time but the stated end of relationship. The endogeneity of relationship satisfaction when analysing the time of marriage dissolution is obvious, as relationship satisfaction is highly influenced by the occurrence of the event. Figure 1 is an indicator for state dependence (Kalbfleisch and Prentice, 2002) of marital satisfaction. It shows the smoothed average trajectory of persons still in the relationship (left) and persons that ended their relationship to their married partner (right). Since these trajectories (both, for men and women) differ considerably between the two groups, it is necessary to employ a modelling technique that addresses the issue of endogeneity appropriately. Additionally, it is reasonable to assume that satisfaction with the marriage does not only change at the time points of the interview but throughout the whole observation period. The *longitudinal model* on satisfaction with the marriage will be modelled by an LMM including a random intercept and a random slope term for the duration of marriage (t). The *time-to event model* on time to marriage dissolution will be modelled jointly with the longitudinal model, taking the estimated values of satisfaction as a main covariate for the hazard of dissolution. Specifically, the time-to-event model is chosen to be a Cox-proportional-hazards model with a spline baseline approximation.

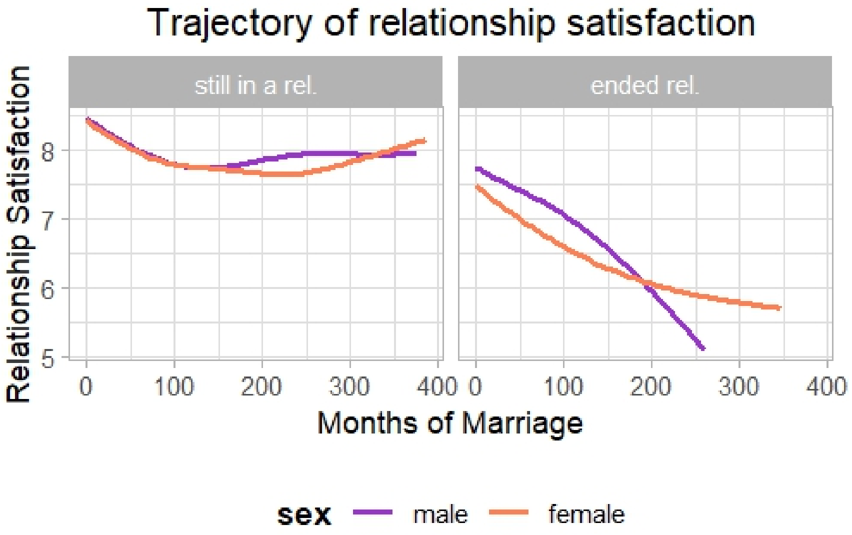


FIGURE 1. Estimated average trajectory of relationship satisfaction of persons by event status and gender. Non-linear smoother by gender (dark: male, light: female).

4 Results

With our separate joint models by gender using the JM package (Rizopoulos, 2010) we find the expected negative current value association for marital satisfaction and the risk of marriage dissolution (see Table 1). This effect is highly statistically significant for women ($\hat{\alpha} = -0.55$) and men ($\hat{\alpha} = -0.47$).

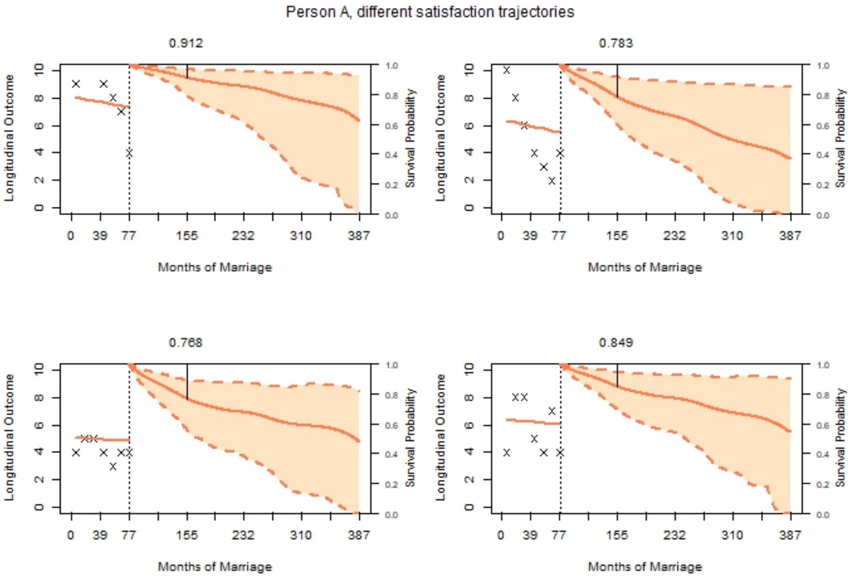


FIGURE 2. Predicted survival probability for a fictional person varying only the marital satisfaction trajectory. Covariate values: female, part-time working, 2 children, at least one preschool child, premarital cohabitation, median values (for females) for the other covariates.

Individual and dynamic survival predictions such as the one in Figure 2 can be made with joint models. The illustration shows the advantage of the modelling approach compared to a classical TVC approach, as all satisfaction trajectories have the same last value but still result in different predicted survival curves for the individual due to the fitting procedure.

Using a TVC model, the effect of marital satisfaction on the risk of marriage dissolution is highly underestimated (women: $\hat{\gamma} = -0.31$, men: $\hat{\gamma} = -0.28$) and a direct effect of the amount of shared household work on the risk of marriage dissolution would be assigned. Applying the joint model we are able to decompose the effect of shared household work into an insignificant *direct* effect and a highly significant *indirect* effect via marital satisfaction for women. The decomposition for men results in both effects being significant, i.e. that higher share of household work for men is associated with a higher risk of marriage dissolution through both pathways – directly and indirectly via marital satisfaction. The models control for the standard socio-economic variables, premarital cohabitation, children as well as for gender-role attitudes.

		longitudinal submodel (marital satisfaction)			time-to-event submodel (risk of marriage dissolution)		
	variable	estimate	Std. Err.	p-value	estimate	Std. Err.	p-value
women	household work	-0.1369	0.0191	0.0000	0.1136	0.0794	0.1527
	mar. satisfaction				-0.5469	0.0552	0.0000
men	household work	-0.0603	0.0267	0.0237	0.2400	0.1312	0.0673
	mar. satisfaction				-0.4693	0.0696	0.0000

TABLE 1. Joint model estimation results for women and men. Only selected coefficients presented.

This application highlights the advantage of using a joint model in terms of additional knowledge production via decomposing the covariate effects.

5 Summary and conclusion

This work aimed to introduce the method of a joint model for longitudinal and time-to-event data in the field of social science research. We demonstrated its usage and usefulness using pairfam data on marriage dissolution and marital satisfaction. The results indicate that higher amount of shared household work in a marriage done by the respondent has no direct effect on the risk of marital dissolution for women but yet a strong indirect effect via marital satisfaction. There is a vast amount of extensions of the rather basic application regarding e.g. the number of longitudinal outcomes, association structures, distributions of the longitudinal outcome and estimation methods.

Acknowledgments: This paper uses data from the German Family Panel pairfam, coordinated by Josef Brüderl, Sonja Drobnič, Karsten Hank, Johannes Huinink, Bernhard Nauck, Franz J. Neyer, and Sabine Walper. The study was funded from 2004 to 2022 as a priority program and long-term project by the German Research Foundation (DFG). The work on this paper was supported by the DFG (Number 426493614) and the Volkswagen Foundation (Freigeist Fellowship).

References

Cremers, J., Mortensen, L. H. and Ekstrøm, C. T. (2021). A joint model for longitudinal and time-to-event data in social and life course research: employment status and time to retirement. *Sociological Methods & Research*, **53**, 1 – 36.

Huinink, J., Brüderl, J., Nauck, B., Walper, S., Castiglioni, L. and Feldhaus, M. (2011). Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. *Zeitschrift Für Familienforschung*, **23**, 77 -– 101.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons.

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. *The International Journal of Biostatistics*, **18**, 127 -– 149.

Rizopoulos, D. (2010). JM: An R Package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, **35**, 1 -– 33.

Wulfsohn, M.S. and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330 – 339.

Non-parametric smoothing for the diffusion of armed conflict

Daniel Racek[1], Paul Thurner[2], Göran Kauermann[1]

[1] Institute of Statistics, Ludwig-Maximilians-University Munich, Germany
[2] Institute of Political Science, Ludwig-Maximilians-University Munich, Germany

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

Abstract: The field of conflict research has moved towards predicting and understanding conflict at subnational levels. However, studies still oversimplify the complex spatio-temporal dynamics of conflict. To address this, this paper introduces a statistical model that captures both the spatial as well as the temporal dimension of conflict diffusion. Using fine-grained conflict data on Africa, we demonstrate that our fully-interpretable diffusion model outperforms models typically employed in the field.

Keywords: Conflict Research; Statistical Modelling; Generalized Additive Model.

1 Introduction

Predicting conflict and understanding its determinants has been the key focus of conflict research for decades (Hegre et al., 2017). In recent years the field has moved towards analyzing conflict in more fine-grained subnational areas, as novel conflict event databases have become available. Paired with the emergence of new data sources such as social media and remote sensing data, numerous new studies have been published that analyze and/or try to forecast conflict on a local level (Bazzi et al., 2022). With this, more advanced statistical models as well as machine learning techniques are finding their way into the forefront of the field (Vesco et al., 2022).

However, many of the utilized models still oversimplify the complex dynamics of conflict and do not adequately account for its dependence over both time and space. As a result, predictive models suffer performance losses, while models studying the determinants of conflict may over- or underestimate the impact of the predictors of interest (Cook et al., 2023).

Many subnational studies are conducted across Africa utilizing 0.5×0.5 decimal degree monthly grid cell observations. But they do not (fully) account for the diffusion of armed conflict across this lattice grid. Schutte and Weidmann (2011) have shown that armed conflict indeed exhibits patterns of spatial and temporal

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

diffusion, i.e., future conflict is influenced by past conflict within a grid cell but also by past conflict in its (further-away) neighbours. However, including these patterns into classical regression models poses a challenge. Hence, most studies simply treat "this dependence [...] as a nuisance" (Schutte and Weidmann, 2011, p.152).

In this work we propose a regression model that is able to flexibly incorporate both the spatial as well as temporal dimension of conflict diffusion, while all of its effects remain fully and easily interpretable. More specifically, we design a generalized additive model with a flexible non-parametric spatio-temporal smoothing component over past conflict, to predict monthly conflict fatalities across grid cells. We show that our proposed model captures the complex dynamics of armed conflict across this lattice grid much better than models typically employed in the field.

2 Data

We draw on conflict data from the widely known UCDP GED (Sundberg and Melander, 2013), which reports events of organized violence. Each reported event is assigned an approximate date, location, type of violence and an estimated number of fatalities. We match these events to the commonly employed PRIO grid cells of size 0.5 × 0.5 decimal degrees ($\sim$ 55 × 55km at the equator). To assess our approach, we investigate Africa, where many research studies have been conducted in. We analyze conflict on a monthly basis, as a more fine-grained temporal resolution becomes problematic due to imprecision in reported event dates.

We focus on the monthly amount of battle-related fatalities in each cell from 2000 to 2020. This implies we have a total of 10,640 grid cells and 252 months, resulting in 2,681,280 observations. Note, that armed conflicts are in most parts of the world, including Africa, extremely rare events. Only 0.42% of all observations exhibit one or more conflict fatalities.

Population has been shown to be one of the key predictors of armed conflict (Raleigh and Hegre, 2009). Thus, we also draw on population data from the WorldPop project (Tatem, 2017). Using satellite imagery, census data and various other geospatial datasets, it estimates yearly population numbers across the world. We employ the 1km resolution dataset and derive the total amount of population for each cell for each year.

3 Method

Let $Y_{t,s}$ denote the number of conflict fatalities occurring in month t in cell location $\boldsymbol{s}$. We define $\boldsymbol{s} = (r, c)^{\mathrm{T}}$ as a bivariate location vector, where r refers to the row, and c to the column of the respective location in the grid. As the number of fatalities are count data we assume that $Y_{t,s} \sim \text{Poisson}(\lambda_{t,s})$. We define the intensity as

$$\lambda_{t,s} = exp(\boldsymbol{x}_{t,s}^{\mathrm{T}}\boldsymbol{\beta}_x + g(\boldsymbol{s}) + \gamma(H_{t,s})), \tag{1}$$

where $\boldsymbol{x}_{t,s}$ is a feature vector of intercept, cell size and the time-varying lagged population of each cell. The component $g(\boldsymbol{s})$ represents a smooth location effect, for which we use thin plate regression splines (Wood, 2003). We denote $\gamma(H_{t,s})$

as our smooth diffusion effect of conflict. We use the notation $H_{t,s}$ to express that we are utilizing the history and neighbouring history H of a cell location $\boldsymbol{s}$ at time point t in our spline representation. We define $\gamma(H_{t,s}) = b(H_{t,s})^{\mathrm{T}}\boldsymbol{u}$, which we model as follows. Let $\tau > 0$ be the maximum time lag and $\delta \geq 0$ the maximum distance considered. We define our basis as

$$\boldsymbol{b}(H_{t,s}) = \sum_{\tilde{t}=t-\tau}^{t-1} \sum_{\tilde{s}\in N_\delta(s)} \boldsymbol{a}(\tilde{t},t) \otimes \boldsymbol{o}(\tilde{\boldsymbol{s}},\boldsymbol{s})\ log(Y_{\tilde{t},\tilde{s}}+1) \tag{2}$$

where $\otimes$ is the Kronecker product of all basis vectors in time ($\boldsymbol{a}$) and space ($\boldsymbol{o}$). We define the neighborhood of a location as $N_\delta(\boldsymbol{s}) = \{\tilde{\boldsymbol{s}} : ||\tilde{\boldsymbol{s}} - \boldsymbol{s}|| \leq \delta\}$. In practice, this means we sum up our time-space basis vector (that results from the Kronecker product) over all past time points for which $t - \tilde{t} \leq \tau$ and all neighbouring cells for which $||\tilde{\boldsymbol{s}} - \boldsymbol{s}|| \leq \delta$.

The individual basis functions in both time and space are exponential decay functions, with $f(x) = exp(-w\ x)$, where $w > 0$ is a pre-defined decay rate and $x \geq 0$. We scale them, such that $f(x_{min}) = 1\ \forall\ w$ and $f(x_{max}) = 0\ \forall\ w$, where x_{min} and x_{max} are the minimum and maximum values allowed for x. We define the decay rates as $w_k = \{w_1, ..., w_K\}$ in time and $v_g = \{v_1, ..., v_G\}$ in space. Hence, $\boldsymbol{b}(H_{t,s}) \in \mathbb{R}^{(KG)\times 1}$. To guarantee smoothness of $\gamma(H_{t,s})$, we employ a ridge penalty, i.e. we estimate the penalized log likelihood

$$l_{pen}(\boldsymbol{\theta}, \boldsymbol{u}, \rho) = l(\boldsymbol{\theta}, \boldsymbol{u}) - \frac{1}{2}\rho\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}, \tag{3}$$

where ℓ is the log-likelihood, ρ the penalty and $\boldsymbol{\theta}$ the remaining parameter vector. Following a Bayesian view, we can incorporate the estimation of this penalty as a random effect (Kauermann, 2005), i.e. we assume $u_i \overset{\text{i.i.d.}}{\sim} N(0, \rho^{-1})$. This allows us to integrate the smoothing into mixed model estimation routines such as the well-known R package *mgcv* (Wood, 2017).

For our use case, we use a set of ten basis functions each. We set $\tau = 24$ to consider the past 24 lags in time, and set $\delta = 10$ to consider all neighbouring cells up to a distance of roughly 550 km from the source cell (horizontally this includes cells up to the 10th-order neighbour), based on upper bounds identified in the literature (Zhukov, 2012; Mueller et al., 2022). This gives us in total a combination of $10 \times 10 = 100$ basis functions that differ across time and/or space. As highest decay rate we choose $w = 5$, resulting in a basis function that only captures the first temporal lag respectively the cell itself (i.e. no spatial lag). As lowest decay rate we choose $w = 0.05$, resulting in a basis function which is (almost) linear. The remaining w are chosen such that there is a constant multiplicative increase in their rate. To be specific, we choose $w_k = v_g = \{0.05, 0.0834, 0.1391, ..., 5\}$. We visualize the set of spatial basis functions $f(s)$ in Figure 1.

We fit our model on data from 2000-2018 and evaluate out-of-sample performance on all observations from 2019 to 2020. Out-of-sample evaluation is necessary, as predicting conflict has been shown to be a particularly difficult task, due to the large amount of noise and no-conflict observations (Racek et al., 2024). To allow for a model comparison both in- as well as out-of-sample, rooted in statistical theory (Dunn et al., 2018), and to understand how well the models are performing relatively compared to a null (intercept-only) model, we compute the explained deviance defined as

$$\text{Explained Deviance} = D_{\text{expl.}} = 1 - \frac{D}{D_0} = 1 - \frac{-2\ell(\hat{\boldsymbol{\theta}})}{-2\ell(\hat{\boldsymbol{\theta}}_0)}, \tag{4}$$

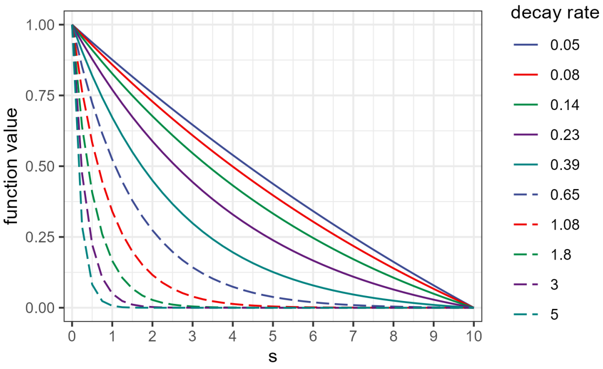


FIGURE 1. Spatial basis functions.

where D denotes the deviance of the respective model and the 0-subscript the null model. We fit a separate null-model on our out-of-sample observations to better understand generalization behaviour. Additionally, we also look at in-sample performance by comparing our models using the AIC.

4 Results

Table 1 provides a summary of our main results. For simplicity, we will refer to baseline models employed in the literature as M0, and to our proposed diffusion model as M1. Baseline M0-1 replaces the diffusion term $\gamma(H_{t,s})$ by temporal lags only (no smoothing), M0-2 extends this to also include the lags of first-order neighbours. Our proposed model (M1) outperforms the baselines both in- as well as out-of-sample (higher AIC, higher explained deviance). Particularly out-of-sample its performance excels, with an increase in 3.4 percentage points (+11.3%) in explained deviance. Hence, we can conclude that our proposed diffusion model has the superior generalization behaviour, and better captures the underlying patterns of conflict diffusion.

TABLE 1. Comparison of proposed diffusion model (M1) with baselines (M0). Lags refers to the number of temporal lags included in the respective baseline. Best performance metrics are denoted in bold.

Model	AIC	$D_{\text{expl.}}$ In-sample	$D_{\text{expl.}}$ Out-of-sample
M0-1, 12 Lags	1614475	0.3397	0.2655
M0-1, 24 Lags	1611291	0.3411	0.2630
M0-2, 12 Lags	1580450	0.3540	0.2984
M0-2, 24 Lags	1573515	0.3569	0.2957
M1	**1553154**	**0.3654**	**0.3322**

In Figure 2 we visualize the diffusion coefficients of $\gamma(H_{t,s})$ for the first four temporal lags t on a spatial grid map with rows (r) and columns (c). Hence, these coefficients capture the effect of one-unit increases in past logged fatalities on $\lambda_{t,s}$.

The visualized diffusion effects are symmetrical, i.e. we can understand them as the increase conflict in cell $(0, 0)$ has on all surrounding cells, and, as the effect conflict in all surrounding cells has on cell $(0, 0)$. For illustration, the coefficient for the first temporal lag at $(0, 0)$ is 0.5483, hence 1 logged fatality in the past month within the same grid cell increases the predicted fatalities by 73.03%. Overall, naturally, the effect is largest in the origin (temporal diffusion only) and decreases as the spatial distance increases (spatio-temporal diffusion). This pattern holds for all monthly lags. Surprisingly, we notice two ring patterns that still require further investigation. More generally, we observe an exponentially decreasing effect over both time and space and the desired smooth effect.

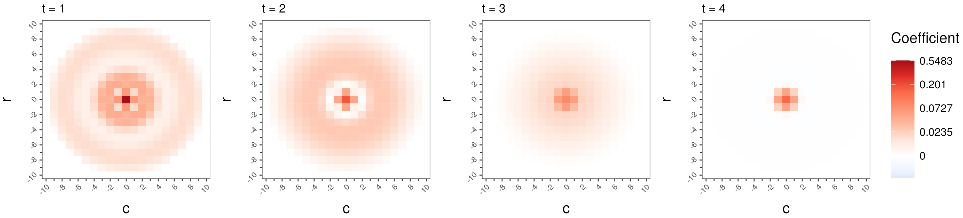


FIGURE 2. Diffusion coefficients on spatial grid map (log colour scale). r refers to the row, c to the column in the spatial grid. t refers to the respective time lag.

References

Bazzi, S., Blair, R.A., Blattman, C., Dube, O., Gudgeon, M. and Peck, R. (2022). The promise and pitfalls of conflict prediction: evidence from Colombia and Indonesia. *Review of Economics and Statistics*, **104**, 764 – 779.

Cook, S.J., Hays, J.C. and Franzese, R.J. (2023). Stadl up! the spatiotemporal autoregressive distributed lag model for tscs data analysis. *American Political Science Review*, **117**, 59 – 79.

Dunn, P.K. and Smyth, G.K. (2018). *Generalized Linear Models with Examples in R*. Volume 53. Springer.

Hegre, H., Metternich, N.W., Nygård, H.M. and Wucherpfennig, J. (2017). Introduction: Forecasting in peace research. *Journal of Peace Research*, **54**, 113 – 124.

Kauermann, G. (2005). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning and Inference*, **127**, 53 – 69.

Racek, D., Thurner, P.W., Davidson, B.I., Zhu, X.X. and Kauermann, G. (2024). Conflict forecasting using remote sensing data: An application to the Syrian civil war. *International Journal of Forecasting*, **40**, 373 – 391.

Raleigh, C. and Hegre, H. (2009). Population size, concentration, and civil war. A geographically disaggregated analysis. *Political geography*, **28**, 224 – 238.

Schutte, S. and Weidmann, N.B. (2011). Diffusion patterns of violence in civil wars. *Political Geography*, **30**, 143 – 152.

Sundberg, R. and Melander, E. (2013). Introducing the ucdp georeferenced event dataset. *Journal of peace research*, **50**, 523 – 523.

Tatem, A.J. (2017). Worldpop, open data for spatial demography. *Scientific data*, **4**, 1 – 4.

Vesco, P., Hegre, H., Colaresi, M., Jansen, R.B., Lo, A., Reisch, G. and Weidmann, N.B. (2022). United they stand: Findings from an escalation prediction competition. *International Interactions*, **48**, 860 – 896.

Wood, S.N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **65**, 95 – 114.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*. CRC press.

Zhukov, Y.M. (2012). Roads and the diffusion of insurgent violence: The logistics of conflict in Russia's north Caucasus. *Political Geography*, **31**, 144 – 156.

Residual analysis in information retrieval models

Sumal Randeni[1], Kenan M. Matawie[1], Laurence A. F. Park[1]

[1] School of Computing, Data and Mathematical Sciences, Western Sydney University, Australia

E-mail for correspondence: `S.Randeni@westernsydney.edu.au`

Abstract: Optimising Information Retrieval (IR) models is vital for improved performance. While various strategies, including parameter estimation and tuning, can contribute to these improvements, our focus lies in utilising a statistical modelling approach, particularly through residual analysis. This paper investigates the application of the Logistic Regression approach within Generalised Linear Models (GLMs) to enhance BM25 IR models. Notably, the utilisation of residual analysis and GLMs for IR model enhancement remains largely unexplored in the domain of Information Retrieval research.

Keywords: Residuals; Information retrieval; Logistic regression; Generalised linear models.

1 Introduction

The function of an information retrieval (IR) system is to provide a set of candidate documents that are predicted to be relevant to a provided query. In this process, retrieval systems use an embedded retrieval model to provide a score to each document reflecting its relevance to the query, where the greater the score, the more likely the document is relevant. For this purpose, search engines in IR systems have been designed to produce a ranked list of documents in a decreasing order of relevance score.

Assessing the retrieval effectiveness is the method of evaluating the performance known as precision of a retrieval model and it is computed by means of an evaluation function that is available in IR. Prior to use these evaluation functions, each document in the ranked list should be marked as relevant or not relevant by means of a manual relevant judgement process . These relevant judgments are readily available for standard document collections that are published by Text Retrieval Conference (TREC) [Donna Harman (1993)] in which they use a binary system as 1 for relevant and 0 for non-relevant document.

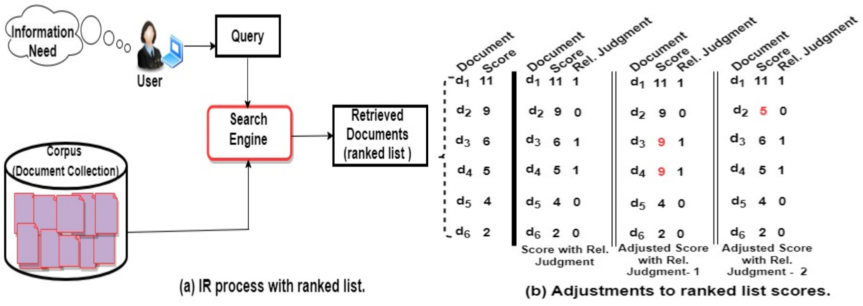
For example, Figure 1(a) shows the IR process with basic components and a ranked list which contains 6 documents $(d_1, d_2, ..., d_6)$ with scores 11, 9, 6, 5, 4 and 2 respectively. The documents in this list are ranked by score, and we can see there that in position 2 (document d_2), the ranked score (9) is higher than the score 6 in position 3 (document d_3), but the document d_2 in position 2 is not relevant (first block in Figure 1(b)). This means that ranking and relevance judgments are not in order and we say that it is not a perfect ranking. The model's performance improves when a greater number of relevant documents appear at the top of the ranked list.

Let us say we equate the score of d_2 to d_3. The error in this case is 3. Similarly, we add 4 to d_4, which result in an error of 4 for d_4. After making these adjustments, the ranked scores become 11, 9, 6+3, 5+4, 4, and 2. This is shown in the second block in Figure 1(b).

Instead, we can subtract from the scores as 11, 9-4, 6, 5, 4, and 2 to obtain the perfect ranking where the error in this case is 4. This is shown in the third block in Figure 1(b).

If we are trying to minimise the change in a score, the second method is better. The adjustments we have discussed here represent the residuals in the ranking of IR for achieving the perfect ranking.

This study aims to improve precision in the retrieval model by prioritising relevant documents at the top of the ranked list. We investigate integrating Logistic Regression [Hosmer and Lemesshow, (2000)] within Generalised Linear Models [Nelder and McCullagh, (1989)] with the IR retrieval model BM25 to create a mixed model. This integrated model is utilised for generating and optimising residuals to enhance IR model performance. The performance of IR models, measured by accuracy or precision, is evaluated using Mean Average Precision (MAP).


Information Need
User
Query
Search Engine
Retrieved Documents (ranked list)
Corpus (Document Collection)

Document	Score
d1	11
d2	9
d3	6
d4	5
d5	4
d6	2

Document	Score	Rel. Judgment
d1	11	1
d2	9	0
d3	6	1
d4	5	1
d5	4	0
d6	2	0

Score with Rel. Judgment

Document	Score	Rel. Judgment
d1	11	1
d2	9	0
d3	9	1
d4	9	1
d5	4	0
d6	2	0

Adjusted Score with Rel. Judgment- 1

Document	Score	Rel. Judgment
d1	11	1
d2	5	0
d3	6	1
d4	5	1
d5	4	0
d6	2	0

Adjusted Score with Rel. Judgment - 2

(a) IR process with ranked list.

(b) Adjustments to ranked list scores.

FIGURE 1. High Level IR process with ranked list. Vertical bar in black separates figures (a) and (b). In (a), a ranked list is shown with 6 documents $(d_1, d_2, ..., d_6)$ with scores. There are 3 blocks (left to right) in (b) and each shows three columns representing documents, scores and relevant judgment assigned to each document. First block shows how relevant judgment (Rel. Judgement) is assigned initially to each document. In the second and third blocks, proposed adjustments to scores are shown in red to push the relevant documents to the top of the list.

2 Residuals, BM25 and MAP

Similar to residual analysis in regression modeling, we introduce predicted and observed scores in Information Retrieval (IR). Optimising the difference between these scores, termed residuals, improves model performance. While achieving a perfect ranking is impossible, our goal is to elevate and identify the position of most relevant documents in the ranked list.
The formula of the BM25 model is given below [Spärk Jones et al., (2000)]:

$$\mathcal{R}(d, Q; k_1, b) = \sum_{t \in Q} \left(\frac{f_{t,d}(k_1 + 1)}{f_{t,d} + k_1 \left(1 - b + b \frac{l_d}{l_{\text{avg}}}\right)} \right) \log \left(\frac{N - df_t + 0.5}{df_t + 0.5} \right)$$

where d is the document, Q is the set of query terms, $f_{t,d}$ is the count of term t in document d, l_d is the length (number of terms) in document d, l_{avg} is the average document length for the document collection, k_1 and b are model parameters, N is the number of documents in the collection, and df_t is the number of documents that contain term t in the collection.

If the set of relevant documents for a query $q_j \in Q$ is $\{d_1,..d_{mj}\}$, and r_{jk} is the set of ranked retrieval results from the top of the ranked list until we get to document d_k, then the Mean Average Precision (MAP)[Manning et al., (2008)] is given as:

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(r_{jk})$$

In this equation, $Precision$ signifies the proportion of retrieved documents that are relevant, while MAP ranges between 0 and 1, with higher scores indicating increased model accuracy.

3 Methodology and results

3.1 Mapping IR residuals to logistic regression

To formalise our problem and define residuals in IR models, let's introduce the information available to us from a dataset using the following notation:

1. Q is the $q \times t$ matrix of query vectors, where Q represents a set of queries in which q is a query and t is a term. Please note that "×" is the multiplication symbol.
2. d is a document vector.
3. r is a vector of relevance judgement for document d across all queries in Q, where r is binary (0,1).

Our goal is to model $r = f(d, Q)$. We want to discover the function f, ideally, $r{=}Q{\times}d$. This is possible if the function (f) is linear, but it is not since r contains 0 (irrelevant) and 1 (relevant), and $Q{\times}d$ is real due to the discrete nature of relevance variable r. To address this problem, let us define a new vector p as the probability of relevance for document d over all queries in Q (p is like r but contains probabilities). We can write the generalise equation:

$$logit(p) = log(p/(1 - p)) = Q \times d$$

This has transformed into a logistic regression problem. We aim to maximise the likelihood of p (to be the same as r) by setting the document weights d. The problem of the logit equation is that the document weight d may overfit the given query and not generalise to new queries. We address this problem by adding a bias term to the logit equation shown above so that it becomes generalise to any query. Thus, we rewrite our logit equation as:

$$logit(p) = log(p/(1-p)) = Q \times (dBM25 + e)$$

where $dBM25$ are the document weights computed using BM25 and e is residuals. We simplify it further and get:

$$logit(p) = Q \times dBM25 + Q \times e = o + Q \times e$$

where o = Q × dBM25 is the score. Using this formula, we can fit e that will generalise to new queries. Also, we could use regularisation, so the optimisation becomes:

$$maximise \;\; \ell(d) - \alpha \times ||dBM25 + e||$$

where $\ell(d)$ is the likelihood function of the logistic regression and α will control how good the fit is. Instead of maximising, we minimise the negative of Log Likelihood. Therefore, instead of $\ell(d)$, we introduce $\ell\ell(d)$ which represents the Log Likelihood. So, our loss function becomes:

$$minimise \;\; -\ell\ell(d) + \alpha \times ||dBM25 + e||$$

While estimating parameters, our models may overfit due to high variance in some data, leading to an increase in sample error. Regularised regression, also known as penalised models, is introduced to mitigate this overfitting.

So, the regularisation penalises high coefficients by adding the regularisation term $G(\beta)$ multiplied by the parameter $\lambda (\in \mathbb{R})$ to the objective function. We write the negative *Log Likelihood* function ($-\ell\ell$), with the regularisation term (with control parameter λ) and relevant judgments (r_j), and the score s_j, for a given query j, as follows:

$$\hat{\beta} = \min_{\beta} -\ell\ell(\beta; r_j, s_j) + \lambda G(\beta)$$

The λ parameter controls how much emphasis is given to the penalty term. Initially, note that s_j is $Q_j \times (dBM25)$, and it is then updated with adjusted score represented by $Q_j \times (dBM25 + e_j)$. The residual for query j, e_j is computed by means of the binomial model of GLM. Now, we need to execute this integrated function for each query against all the documents in our dataset, as describe next.

3.2 Implementation and results

To implement our integrated function for optimisation, we use *glmnet* package in *R* on our data source known as "Cranfield dataset". This data set contains 1398 documents with 4499 terms, 225 queries and the corresponding TREC relevance judgments. Unlike in other large TREC collections [(Donna Harman (1993)], in this data set, an exhaustive relevance judgments of all (query, document) pairs have been conducted. From this collection, we have selected documents that have more than one relevance judgement for each query because GLM application fails to respond for documents that have one or zero relevant judgments. After excluding documents with one or zero relevant judgments for queries, our document set has been reduced to 474. Additionally, we have limited our experiments to the top 100 queries. So, in the end, our filtered data set contains 474 documents,

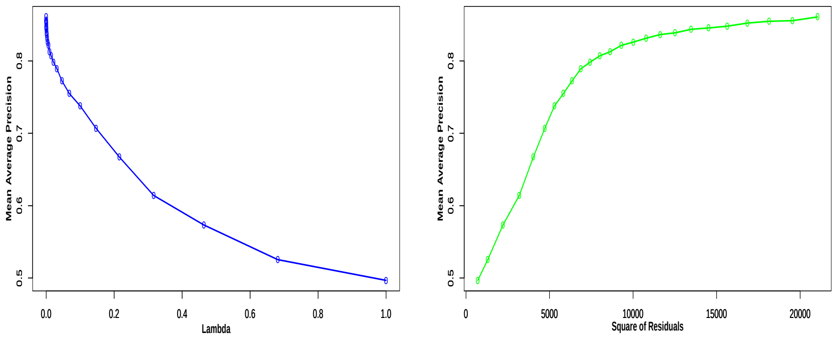


FIGURE 2. MAP with Lambda values (left) and MAP with Square of Residuals (right). Both used ridge model.

4370 terms and 97 queries.

We conducted preliminary experiments to determine a suitable range of λ values, settling on 25 values ranging from 1 to 0.0001. In this paper, we present a segment of our experiments, focusing on the results obtained for 25 λ values in the Ridge Penalty model. We computed Mean Average Precision (MAP) for each λ, and for residual sum of squares (RSS), and the results are shown in Figure 2. As expected, for this specific dataset, MAP values approach 1 as λ decreases, reflecting the reduced penalisation of residuals (see left graph). This suggests that a significant number of relevant documents are consistently placed at the top of the ranked list. Similarly, the graph in the right shows that MAP is approaching to 1 and saturating with the increase of RSS. Additional analysis is currently ongoing.

4 Conclusion

In this study, we explored the application of residuals in Information Retrieval through logistic regression and investigated the optimisation of IR models using Generalized Linear Models. This is a largely unexplored area in the domain of IR research. In this work, the BM25 model exhibited increased precision, particularly for lower λ values. Our approach suggests a potential framework for evaluating the precision of IR models via residuals, which may be validated and explored further in future studies. The experiment and results are ongoing research, and further analysis and modelling, are planned for potential enhancements.

References

Harman, D. (1993). The first text retrieval conference (TREC-1), special publication (NIST sp, U.S. Department of Commerce), 500 – 207.

Hosmer, D. W. and Lemesshow, S. (2000). *Applied Logistic Regression.* John Wiley & Sons, New York.

Manning, C.D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Nelder, J. A. and McCullagh, P. (1989). *Generalised Linear Models.* Chapman & Hall/CRC.

Sparck Jones, K., Walker, S. and Robertson, S. E. (2000). A probabilistic model of information retrieval. *Information Processing and Management*, **36**, 809 – 840.

Bridging the data gap - Estimate the exit rates, entries, and exits using only data on occupancy

Martje Rave[1], Göran Kauermann[1]

[1] Department of Statistics, Faculty of Mathematics, Informatics and Statistics, Ludwig-Maximilians-Universität München, Germany

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

Abstract: In this abstract, we focus on illustrating the method which enables us to estimate the length of stay, along with the number of entries and exits observing only the occupancy and covariates. The application of this method is particularly beneficial in data situations where we face missing response data. Though this method can be used in a plethora of situations, we will showcase two data sets in the conference poster. These are; first the number of incoming and outgoing COVID-19 patients in the ICU, and second the number of bikes rented and returned at bike stations in Vienna. For the purpose of this abstract, we will present the method and its application to simulated data only.

Keywords: Stochastic EM algorithm; Skellam distribution; Imputation.

1 Method

We start with the definition outlined in Equation 1.

$$\Delta_{(i)} \equiv O_{(i)} - O_{(i-1)} = I_{(i)} - R_{(i)} \tag{1}$$

Here, $\Delta_{(i)}$, is defined to be the difference in the observed occupancy, $O_{(i)}$, at a given observation and the observed occupancy, $O_{(i-1)}$, immediately preceding (i). This is equal to the number of entries during this period, $I_{(i)}$, minus the number of exits during this period, $R_{(i)}$. Both $I_{(i)}$ and $R_{(i)}$ are unobserved. Since the number of entries and exits are counting processes, it is reasonable to assume that they are Poisson distributed. The difference between two Poisson-distributed random variables follows a Skellam distribution, Skellam (1948).

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

$$
\begin{aligned}
I_{(i)} &\sim Poisson(\lambda^I_{(i)}) \\
R_{(i)} &\sim Poisson(\lambda^R_{(i)}) \\
\Delta_{(i)} \equiv I_{(i)} - R_{(i)} &\sim Skellam(\lambda^I_{(i)}, \lambda^R_{(i)}),
\end{aligned} \tag{2}
$$

where $\lambda^I_{(i)}$ and $\lambda^R_{(i)}$ are both the intensity parameters of the given Poisson distributions but also the parameters of the Skellam distribution.

We can use the stochastic-expectation-maximization (stEM) to iteratively simulate from a Skellam distribution to obtain the number of entries and exits, use this to estimate the intensity parameters, using two GAMs, which are then used in the Skellam distribution in the simulation step, until convergence is reached. Here, the exit rates were taken to be fixed and explained in detail in our first paper, Rave et al. (2023).

The extension to the initial approach allows us to estimate the exit rates, as well as the entries and exits, where the intensity parameters for the entries are estimated using a generalized additive model. The functions for the intensity parameters are shown in Equation 3.

$$
\begin{aligned}
\lambda^I_{(i)} &= \exp\left(\eta^I_{(i)}\right), \\
\lambda^R_{(i)} &= \exp\left(\eta^R_{(i)}\right)\left(\sum_{j=1}^{\tau} \omega_j \hat{I}_{(i-j)}\right),
\end{aligned} \tag{3}
$$

with $\eta^I_{(i)}$ and $\eta^R_{(i)}$ being the linear combinations of the covariates. ω_j refers to the exit rate at the j^{th} lag. The j^{th} lag is the j^{th} time unit prior to the time of observation, with maximum lag being τ, and the minimum lag being 1.

The intensity parameters for the exits, $\lambda^R_{(i)}$, are now estimated using a seesaw algorithm as they are subject to constraints. The convergence of the methodology is elaborated on by Spall (2012). Namely, the constraints are that the exit rates, $\{\omega_1, \ldots, \omega_\tau\}$, must sum up to 1 and each must be larger or equal to 0. To achieve this, we use an approximation to the log-likelihood given by Equation 4, shown by, Lindstrom et al. (1990).

$$
\begin{aligned}
l^R_P(\boldsymbol{\omega}) &\approx l^R_P\left(\hat{\boldsymbol{\omega}}^{(q)}\right) + s^T\left(\hat{\boldsymbol{\omega}}^{(q)}\right)\left(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}^{(q)}\right) - \frac{1}{2}\left(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}^{(q)}\right)^T J\left(\hat{\boldsymbol{\omega}}^{(q)}\right)\left(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}^{(q)}\right) \\
&\approx [s^T\left(\hat{\boldsymbol{\omega}}^{(q)}\right) + \left(\hat{\boldsymbol{\omega}}^{(q)}\right)^T J\left(\hat{\boldsymbol{\omega}}^{(q)}\right)]\boldsymbol{\omega} - \frac{1}{2}\left(\boldsymbol{\omega}^T J\left(\hat{\boldsymbol{\omega}}^{(q)}\right)\boldsymbol{\omega}\right) + K,
\end{aligned} \tag{4}
$$

where $s(\boldsymbol{\omega})$ and $J(\boldsymbol{\omega})$ are the score function, shown in Equation 6, and information matrix, shown in Equation 7, respectively. Both are derived from the log-likelihood function shown in Equation 5 and $\hat{\boldsymbol{\omega}}^{(q)}$ are the estimated exit rates at the q^{th} iteration of the seesaw algorithm.

$$
l^R_P(\boldsymbol{\omega}) = \sum_{i=1}^{N} R_{(i)} log(\sum_{j=1}^{\tau} \omega_j \hat{I}_{(i-j,d)}) - exp(\beta^R_0) \sum_{j=1}^{\tau} \omega_j \hat{I}_{(i-j,d)}. \tag{5}
$$

$$
s(\omega_i) = \frac{\partial l^R_P(\boldsymbol{\omega})}{\partial \omega_s} = \sum_{i=1}^{N} R_{(i)} \Big(\frac{\hat{I}_{(i-s)} - \hat{I}_{(i-\tau)}}{\sum_{j=1}^{\tau} \omega_j \hat{I}_{(i-j)}}\Big) - exp(\beta^R_0)(\hat{I}_{(i-s)} - \hat{I}_{(i-\tau)}). \tag{6}
$$

$$
J_{[s,k]}(\boldsymbol{\omega}) = -\frac{\partial l^R_P(\boldsymbol{\omega})^2}{\partial \omega_s \partial \omega_k} = \sum_{i=1}^{N} R_{(i)} \frac{(\hat{I}_{(i-s)} - \hat{I}_{(i-\tau)})(\hat{I}_{(i-k)} - \hat{I}_{(i-\tau)})}{(\sum_{j=1}^{\tau} \omega_j \hat{I}_{(i-j)})^2}, \tag{7}
$$

with $[s, k]$ being the s^{th} row entry and k^{th} column entry of the information matrix.

2 Simulation study

Simulated Data We simulate ten simple data sets in which the entries are generated from a Poisson distribution with intensity parameter, $\lambda_I^{sim} = 10$, as shown in Equation 8.

$$\begin{aligned} I_{(i,j)}^{sim} &\sim Pois(\lambda_I^{sim}), \\ &\forall i \in \{1, \dots, 300\} \text{ and } j \in \{1, \dots, 200\}. \end{aligned} \tag{8}$$

Equation 9 shows how the number of exits are generated. The exit rates, $P(\text{lag} = t) = \omega_t$ are randomly chosen to be $(\omega_1 = 0.5, \omega_2 = 0.2, \omega_3 = 0.2, \omega_4 = 0.1)$, thus 50% of entries exit after one unit of time following the date of observation, 20% of entries exist two units of time following the day of entry, and so forth. The maximum lag, $T_{max} = 4$, is also randomly chosen. From this probability mass function, we simulate the length of stay for each entry; $(1, \dots, I_{(i,j)})$.

$$\begin{aligned} I_{(i,j)}^{sim}(\text{lag} = t) &= \sum_{l=1}^{I_{(i,j)}^{sim}} \mathbb{I}(l = t), \\ R_{(i,j)}^{sim} &= \sum_{z=1}^{T_{max}} I_{(i,j)}^{sim}(\text{lag} = z), \end{aligned} \tag{9}$$

with $\mathbb{I}(l = t)$ being an indicator function, which takes the value of 1, if the simulated length of stay, l, is equal to the t^{th} lag, at time point, i, for a given group, j, and 0 otherwise.

Results The results of the very simplified version of these applications are illustrated for one of the simulated data sets. We observe that indeed the simulated data are estimated by the method reasonably well, as seen in Figure 1. Further results are going to be presented in the final poster.

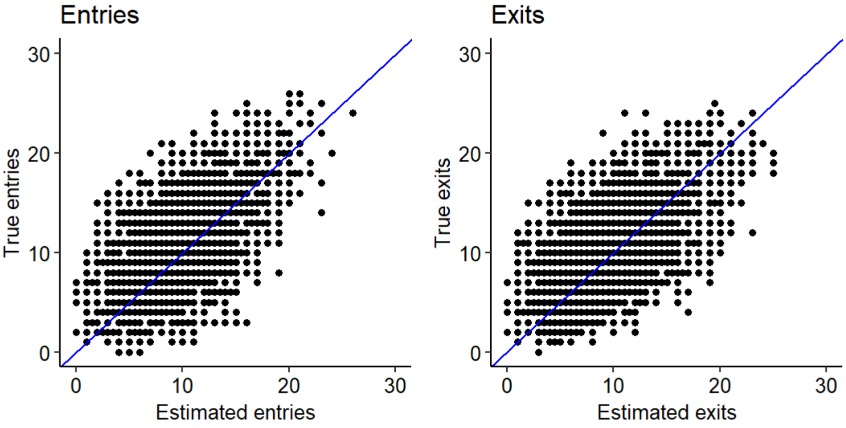


FIGURE 1. Simulated vs estimated entries and exits.

References

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673 – 687.

Rave, M. and Kauermann, G. (2023). The Skellam distribution revisited: Estimating the unobserved incoming and outgoing ICU Covid-19 patients on a regional level in Germany. *Statistical Modelling*, doi:10.1177/1471082X241235024.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B (Methodological)*, **10**, 257 — 261.

Spall, J. C. (2012). Cyclic seesaw process for optimization and identification. *Journal of Optimization Theory and Applications*, **154**, 187 — 208.

Optimizing variable selection in multi-omics datasets: a focus on exclusive lasso

Dayasri Ravi[1], Andreas Groll[1]

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

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

Abstract: Multi-omics datasets pose significant challenges due to their structured nature, where highly correlated variables are grouped within a complex, high-dimensional framework. Traditional Lasso methods encounter limitations in handling correlated features within these groups effectively. To address this issue, we propose using Exclusive Lasso, focusing on inducing sparsity at the intra-group level. Additionally, we introduce an efficient algorithm for solving the related optimization problem. By prioritizing feature selection robustness within correlated group structures, our proposed methodology offers a promising solution to the challenges inherent in analyzing biological datasets. This advancement enhances our ability to extract meaningful insights from multi-omics data, thus facilitating deeper understanding and exploration of complex biological systems.

Keywords: Variable selection; Composite penalty; Exclusive lasso.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

A comparison of extreme gradient and Gaussian process boosting for a spatial logistic regression on satellite data

Michael Renfrew[1], Bruce J. Worton[1]

[1] School of Mathematics, The University of Edinburgh, United Kingdom

E-mail for correspondence: `M.A.Renfrew@sms.ed.ac.uk`

Abstract: A popular and successful method of obtaining regression models using decision tree learners is XGBoost. However, the method implicitly assumes conditional independence of the predictions given the data and is not statistically efficient for autocorrelated data, as arises in spatial statistics. GPBoost incorporates a Gaussian process in a mixed effects model, and is demonstrated for our remote sensing model to reduce the generalisation error dramatically.

Keywords: Gaussian processes; Gradient boosting; Logistic regression; Machine learning; Spatial regression.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Design optimality for a general alcohol model for the human body

Juan M. Rodríguez-Díaz[1], M. Teresa Santos-Martín[1], Irene Mariñas-Collado[2]

[1] Department of Statistics, University of Salamanca, Spain
[2] Department of Statistics and Operations Research and Mathematics Didactics, University of Oviedo, Spain

E-mail for correspondence: `juanmrod@usal.es`

Abstract: The usual equation employed to estimate a person's blood alcohol concentration after consuming alcoholic drinks assumes zero-order kinetics in the ethanol elimination phase. This implies that the elimination process occurs in the body at a uniform rate as a function of the ethyl-oxidation constant. The model, formulated by Widmark, does not consider the phase of increase in concentration, and approximates the phase of elimination in a linear way, which may be insufficient if the tests are carried out in the first phases of alcohol intake. A new model addressing alcohol absorbtion and elimination phases in the human body is proposed, which has several advantages over other models existing in literature. Optimal designs are computed for different types of drinking subjects. Furthermore, the case of several alcohol incorporations, not very much treated in literature, is analyzed and a convenient model is proposed as well. Finally, aA proposal for constructing design that are quasi-optimal with little computational effort is presented, which could be used to create tables of optimal designs from a very easy way.

Keywords: Absorbtion phase; Alcohol model; Elimination phase; Optimal designs; Widmark equation.

1 Introduction

Alcohol is probably the most extended legal drug. In the human body, most of the ingested alcohol (90-98%) is processed in the liver, with the remaining 2-10% eliminated unchanged by breath, perspiration, and urine.

There are several equations that can be used to model the pharmacokinetics of ethanol and thus the Blood Alcohol Concentration (BAC) in the body, but it has traditionally been modelled making use of the Widmark equation, which was

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

first developed in the 1930s (Watson, 1981). In these equation, it is assumed that the BAC decreases at a constant rate per a unit time, i.e. zero-order elimination rate. Although BAC is the most reliable indicator of alcoholic drunkenness, police frequently use a Breath Alcohol Concentrarion (BrAC) estimate obtained with a breathalyzer, which is a less intrusive and more practical tool and, therefore, the Widmark equation has been adapted for these type of tests. In order to match the estimates of blood and breath concentration temporal patterns, a blood/breath alcohol ratio of 2,300:1 is commonly advised.

Since the main metabolizing enzyme is saturated at low blood alcohol concentrations, ethanol is a good example of a drug that usually displays dose-dependent or saturation kinetics and, for questions arising in forensic science and legal medicine (BAC of 50-500 mg), zero-order kinetics is a reasonable assumption for characterizing blood ethanol elimination. However, below a BAC of 5–10 mg% the metabolizing enzymes are no longer saturated with substrate and first-order kinetics apply (Jones, 2019). The linear model relates to the one-compartment model with zero-order elimination kinetic, leaving the absorption kinetics out of the study. Thus, the traditional linear model may be ineffective at forecasting BAC at various time points, as well as estimating the time when the maximum is reached.

The Widmark model is the most used in forensic medicine and by traffic officers, due to its simplicity and because it adjusts quite well the alcohol elimination phase. However, in some situations the interest is not in estimating the level of alcohol at the present time, but in past temporal points (e.g. forensic science trying to estimate the level of alcohol of the driver at the exact time of a car accident that happened some time ago: minutes, hours...). In this scenario, it would not be reasonable to estimate the level of alcohol in the past by the decreasing linear trend. The linear regression model works well in a local environment of the lab test, but it does not when going backwards. As can be seen in the Figure 1, the green point would represent the real alcohol concentration and the red point the estimate made with the Widmark line. A non-linear trend model is needed, more specifically, a hill-shaped function that could describe not only the clearance but also the absorption phase of alcohol intake.

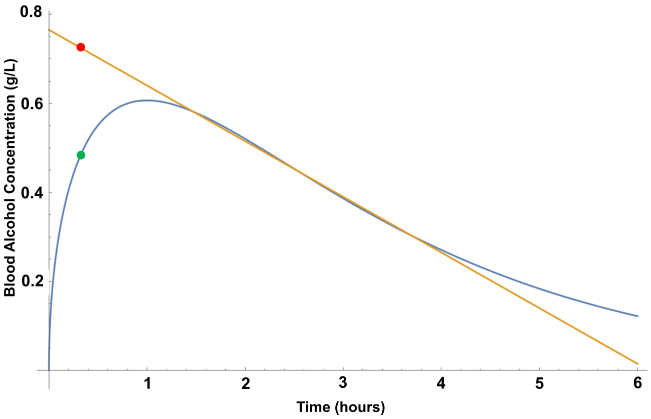


FIGURE 1. Widmark line versus blood alcohol concentration

Regardless of the model used, usually the aim is to obtain good estimation of the parameters, and for this reason it is important to take observations at the moments that give more information, that is to use Optimal Design of Experiments methodology. An experiment with a good design and a proper model not only yields more information than an experiment with a worse design, but it also makes it possible to provide the best conditions for the experiment. D-optimality is the most used criterion to measure the goodness of a design, providing the time points $\{t_1, t_2, ...\}$ at which to take samples in order to get the best estimators of the parameters of model, that is, the estimators with minimum variance. Nonlinear models arise in scientific experiments in a variety of areas, such as pharmacology, biology and agriculture. Determination of optimal designs for these models is more tricky and it usually involves linearization.

As an alternative to different hill-shaped models found in literature, the simplified-Gamma model is presented Mariñas-Collado et al (2023), which can capture all the different phases of the BAC while remaining user-friendly. It can be expressed as:

$$C_{SG}(t) = st^a e^{-bt}, \tag{1}$$

with a, $b > 0$. This model has always a hill shape, and the maximum value is attained at $t^* = a/b$. The simplified model, just like the Gamma function, enables the fitting of a wide range of hill-shape models. However, since the most important fact is indeed this hilly shape, there is no need for the density-function constraints; these can be removed to obtain a model that is easier to handle and work with.

2 Optimal and quasi-optimal designs for single/several alcohol intakes

Most of existing works in literature assume a single and instantaneous alcohol intake, both of the assumptions not very realistic actually. However, sometimes they may approximate the real situation. Mariñas-Collado et al (2023) use the simplified-Gamma model to study the best designs were observations should be taken in order to get a better estimation of the model, for different subject characteristics (gender, body weight, quantity of alcohol ingested), proposing as well equally-spaced designs with several points. This approach is quite popular among practitioners, that usually prefer to take observations 'filling' the design space instead of taking the (usually few) samples asked by the optimal design.

However, in some situations the single intake is no longer valid, and models with several intakes should be considered. When assuming a second intake at time m, the intuitive approach is to fit a piecewise-defined model:

$$C_{SG_2*}(t) = \begin{cases} st^a e^{-bt}, & t \leq m, \\ s\left(t^a e^{-bt} + (t-m)^a e^{-b(t-m)}\right), & t > m. \end{cases}$$

However, the main problems of this model are that it is not differentiable at m, and has abrupt intake effects. An approximate model that avoids the first problem and soften the second one will be proposed, and optimal designs will be computed for different values of m. Moreover, quasi-optimal designs will be proposed. To get these designs for a specific set of subject characteristics (gender, body weight,...)

and quantity of alcohol consumed, optimal designs should be computed just for the two extreme values of m, and they can be used to get high-efficient designs for the rest of the values of the second alcohol incorporation.

References

Jones, A.W. (2019). Alcohol, its absorption, distribution, metabolism, and excretion in the body and pharmacokinetic calcula- tions. *Wiley Interdisciplinary Reviews: Forensic Science*, **1**, e1340.

Mariñas-Collado, I., Rodríguez-Díaz, J. M. and Santos-Martín, M.T. (2023). Optimal designs for a non-linear model for the pharmacokinetics of ethanol elimination in the human body. *Chemometrics and Intelligent Laboratory Systems*, **238**, 104839.

Watson, P.E., Watson, I.D. and Batt, R.D. (1981). Prediction of blood alcohol concentrations in human subjects. updating the widmark equation. *Journal of studies on alcohol*, **42**, 547 – 556.

A suggestion for a test if a calibrated quantitative adverse outcome pathway is chemical agnostic based on between chemical heterogeneity

Ullrika Sahlin[1], Zheng Zhou[1]

[1] Lund University, Sweden

E-mail for correspondence: `ullrika.sahlin@cec.lu.se`

Abstract: A quantitative Adverse Outcome Pathway (AOP) is a chemical agnostic model predicting an adverse outcome given molecular initiating event. A calibrated qAOP can be used to derive a point of departure when assessing the health effects of chemicals under next generation risk assessment. We propose a way to test if a calibrated quantitative AOP is chemical agnostic, which consider the trade-off between between chemical heterogeneity and predictive performance. To do this, we formulate a statistical model for in vitro and in vivo dose-response data on multiple key events in an AOP collected from several chemicals. The performance of calibration and the proposed test is evaluated by a simulation experiment.

Keywords: qAOP; Evidence synthesis; Calibration; heterogeneity.

1 Introduction

1.1 qAOPs

Next Generation Risk Assessment relies on new approach methodologies involving non-human non-animal data. An adverse outcome pathway (AOP) is according to Villeneuve et al. (2014), a conceptual framework that organizes existing knowledge concerning biologically plausible, and empirically supported, links between molecular-level perturbation of a biological system (key events, KEs) and an adverse outcome (AO) at a level of biological organization of regulatory relevance (Figure 1).
An AOP is supposed to be chemical agnostic, i.e. given a level of the MIE, it should not matter which chemical that is triggering the MIE.

1.2 Hazard Assessment

A quantitative AOP is a model that predict the AO given available information of events in the AOP, at least the Molecular Initiating Event (MIE). The probability

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

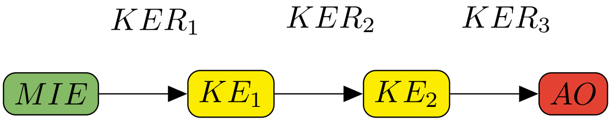


FIGURE 1. An AOP contains relationships between them (KERs), leading from a molecular initiating event (MIE) triggered by a stressor to an adverse health effect on the organism or population level (AO).

Pr(AO—MIE) can be understood as how certain we are that the AO will occur given the MIE (Perkins et al. 2019). It has been suggested to use a calibrated quantitative AOP together with a chemical specific dose-response model, as a new method to derive a Point of Departure (PoD) in Human Health Hazard Assessment (Conolly et al. 2017). The qAOP is calibrated for in vitro data or for non-human data, and therefore, the reference point for the PoD, might have to be informed by a extrapolation from in vitro to in vivo (qINVIVE), or the internal in vitro exposure is linked to external exposure by a Physiologically based pharmacokinetic modelling (PBPK) modelling.

$$\underbrace{\overbrace{P(AO_{in\ vivo}|AO_{in\ vitro})}^{qINVIVE} \cdot \overbrace{P(AO_{in\ vitro}|MIE)}^{qAOP}}_{chemical\ agnostic} \cdot \underbrace{\overbrace{P(MIE|d_{int\ expo})}^{dose-response} \cdot \overbrace{P(d_{int\ expo}|d_{ext\ expo})}^{PBPK}}_{chemical\ specific}$$

1.3 Data to calibrate a qAOP

Quantitative AOPs are statistical models of the KERs calibrated by dose-response data (Figure 2). The data for each KERs consists of in vitro or in vivo dose-response data for different chemicals.

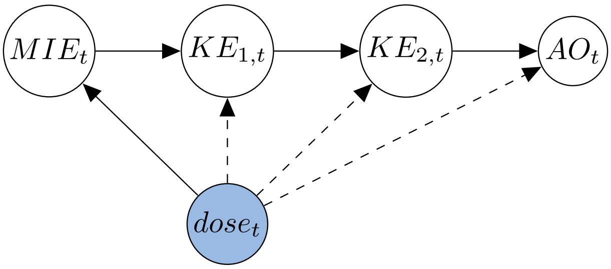


FIGURE 2. In vitro or in vivo studies generate dose-response data measured over time that can be used to calibrate a qAOP.

2 Aim

The aim is to

1. formulate a statistical model for calibration of qAOP based on dose-response data from multiple chemicals, and
2. propose a performance measure based on between chemical heterogeneity to test if the calibrated qAOP is chemical agnostic

3 Method

3.1 Dose-response model

The functional form of the dose-response relationship for a response measured as a proportion and thereby bounded in the unit interval was chosen to be:

$$H(t) = \nu + (1-\nu)\frac{t^{\beta}}{\alpha^{\beta} + t^{\beta}}$$

where $0 < \nu < 1$, $\alpha > 0$, $\beta > 0$.
In our example, we consider in vitro data on the Key Event for chemical k as provided for doses at different levels i, $d_{i,k}$, and the corresponding response measured as the proportion of events $x_{i,k}$ with a standard error $se_{i,k}$).
The statistical model proposed for a dose-relationship for chemical k is:

$$x_{i,k} \sim N(z_{i,k}, se_{i,k})$$
$$z_{i,k} = \nu_k + (1-\nu_k)\frac{d_{i,k}^{\beta_k}}{\alpha_k^{\beta_k} + d_{i,k}^{\beta_k}}$$

The background response is assumed to be zero $\nu_k = 0$. We use suitable priors for the chemical specific parameters α_k and β_k for all k that ensure increasing dose-response relationships within reasonable limits.

3.2 Response-response model

The functional form of the response-response model is specified by a modification of the distribution function of the new power function distribution (NPFD) defined by Iqbal et al. (2021)

$$G(t) = 1 - \left(\frac{1-t}{(\delta-1)t+1}\right)^{\eta}$$

where $0 < t < 1$, $\eta > 0$ and $\delta > 0$.
Here, in vivo data on the Adverse Outcome is assumed to consist of doses at different levels i for chemical k, $d_{i,k}$ and the corresponding response measured as the number of tumours $y_{i,k}$ and total number of individuals (often rats) $n_{i,k}$).
The statistical model proposed for the response-response-relationship for is:

$$y_{i,k} \sim Bin(n_{i,k}, \pi_{i,k})$$
$$\pi_{i,k} = G(z_{i,k})$$
$$log(\delta_k) \sim N(\mu_{\delta}, \sigma_{\delta})$$
$$log(\eta_k) \sim N(\mu_{\eta}, \sigma_{\eta})$$

Suitable priors for the parameters μ_{δ}, σ_{δ}, μ_{η}, σ_{η} are chosen to ensure convergence when combining dose-response and response-response models for a set of chemicals $k = 1, \ldots, K$.

3.3 Performance measure

The performance of a calibrated qAOP is evaluated with respect to predictive performance for all chemicals (jointly and separately), and the chemical heterogeneity in the model for the KER based on the parameters σ_δ and σ_η. The qAOP is chemical agnostic if σ_δ and σ_η are small. We test if a qAOP is chemical agnostic by evaluating the trade-off in predictive performance for a response-response model with no chemical specific parameters.

3.4 Simulation study

We use a simulation study to

1. demonstrate how well the calibrated model is able to estimate the true KERs, and
2. how well the proposed test can discriminate between qAOPs that are more or less chemical agnostic

References

Villeneuve, D.L., et al. (2014). Adverse outcome pathway (AOP) development I: strategies and principles. *Toxicological Sciences*, **142**, 312 — 320.

Conolly, R.B., et al. (2017). Quantitative adverse outcome pathways and their application to predictive toxicology. *Environmental Science & Technology*, **51**, 4661 – 4672.

Iqbal, M.Z., Özel, G. and Balogun, O.S. (2021). A better approach to discuss medical science and engineering data with a modified Lehmann type-II model. *F1000Research*, **10**, 823.

Perkins, E.J., et al. (2019). Building and applying quantitative adverse outcome pathway models for chemical hazard and risk assessment. *Environmental Toxicology and Chemistry*, **38**, 1850 – 1865.

Employing random effects in variance components modelling

Kristína Sakmárová[1], Arnošt Komárek[1], Martin Otava[2]

[1] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[2] Janssen-Cilag s.r.o, Czech Republic

E-mail for correspondence: `sakmarova@karlin.mff.cuni.cz`

Abstract: Typical models used for the analysis of the concentration of the active pharmaceutical ingredient include fixed run effects model and random run effects model, while both models assume homogenous random location and residual effects. In this article we choose random run effects model, as it has more benefits than fixed run effects model. Location and residual errors however do not have to be necessarily i.i.d., which further requires population effect and random effect parametrization. For parameter estimation Bayesian methods were used, using brms package. Assuming hierarchical random effects model we provide results for population residual and location-to-location variability among batches.

Keywords: Residual error variation; Random effects standard deviation variation; Bayesian methods.

1 Introduction

The uniformity of dosage units of oral solid dosage (i.e. tablets) refers to the degree of variability of the content of the active pharmaceutical ingredient (API) in the tablet. It is the critical quality attribute, as the amount of the API needs to be tightly controlled in order to deliver to patient the required dose. Both underdosing and overdosing can have severe implications for the patient. Further, stratified content uniformity is often assessed during the process development for a deeper understanding of the variability within the batch, separating location-to-location variability and residual variability. The residual variability, i.e. variance among tablets manufactured at nearly same timepoint, originates from micromixing. This phenomenon refers to local variability within the blend, resulting in a certain degree of variability in the content uniformity of the tablet cores sampled at single sampling times. Variability between locations is rarely interesting in batch processes, where lot of premixing is done at various stages, but becomes a factor for continuous manufacturing (CM). CM of oral solid dosage continues

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

to demonstrate has advantages in pharmaceutical manufacturing. In contrast to traditional batch processes, CM provides reduced footprint, easier scale-up of production and much faster product delivery and reaction to the market. The fast production processes were enabled by the introduction of reliable fast measuring techniques that allow real time assessment of the product quality during production. The waiting time for the lab analysis has been main drawback of traditional reference methods. Besides increased speed, process analytical technology (PAT), often based on near-infrared spectroscopy (NIR), is much faster and cheaper to implement individual measurements, resulting into considerably larger sample size obtainable from the batch which further improves product quality evaluation. The regular sampling is important part of CM, as variability on the feeders of various components of tablets together with absence of extensive premixing may lead to large variation in time, i.e. among the sampling locations. Additionally, correct assessment of the within-location variability is also important, especially during early development, from the perspective of early product quality evaluation and determination of expected errors around the residence time distribution (RTD). RTD estimates the average blend properties based on feeder performance, but cannot reflect directly the variability at the tablets level, so it is important to take residual error into account when establishing control limits. An assumption of common variance among batches for both within- and between-locations often holds during the late stages of development (such as process validation), where the process is tightly controlled and few batches are explored. However, such an assumption may not be valid during early development evaluation, when process parameters settings and material properties may vary among batches, having possibly considerable impact on the blending performance both in long-term as well as directly impacting micro-mixing. In such cases, heteroscedasticity should be considered, e.g. with separate residual error per batch and possibly as well separate variance of between-location means. However, the actual value of variance for a given batch would rarely be of interest. Instead, we would like to know what we can say about the population of batches regarding their residual and location-to-location variability.

2 Methods

2.1 Data description

For confidentiality reasons we provide only simulated data based on the real-life values. The data consist of 600 observations of concentration of API, generated in 3 samples. In each sample, we measure the concentration in 10 runs and within each run in 20 locations. The simulated data are plotted in Figure 1. We can see, that for each of the 10 runs we have concentrations among 20 locations and 3 values of the outcome for each location. The data show different variability within location and even between locations. We thus incorporate random effects in variance of residuals and location effects in statistical models in the next section to capture this phenomenon. Based on Figure 2 we can assume homogeneity of variance within runs.

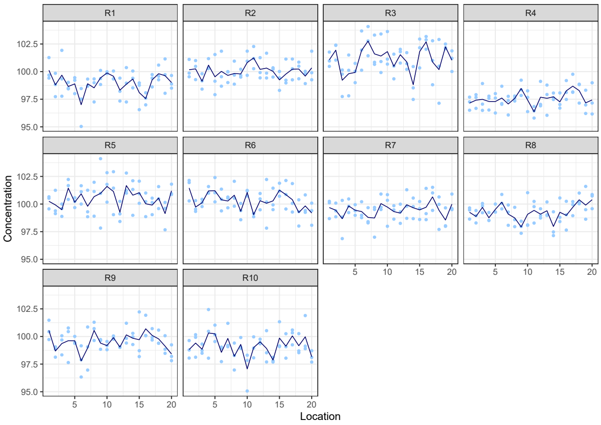


FIGURE 1. Artificial data resembling real CM line data.

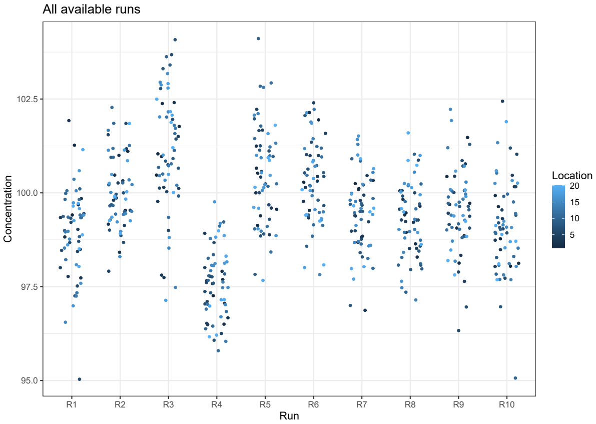


FIGURE 2. Within-run variability displayed.

2.2 Model structure

The simplest model used to analyze the concentration of API (Y_{ijk}, where $i = 1, \ldots, 10$ represents the index of run, $j = 1, \ldots, 20$ is the location index and $k = 1, 2, 3$ is the sample index) is the fixed run effects (R_i) model with normally distributed random effect of location (l_{ij}) and normally distributed random errors (ε_{ijk}):

$$Y_{ijk} = R_i + l_{ij} + \epsilon_{ijk}, \qquad l_{ij} \sim N(0, \sigma^2_{loc}), \qquad \epsilon_{ijk} \sim N(0, \sigma^2_{res}). \quad (1)$$

To acknowledge the fact, that the effect of the run is allowed to vary across all its levels, we need to include variability among runs as well. This leads us to use of random run effects (r_i) model:

$$Y_{ijk} = \mu + r_i + l_{ij} + \epsilon_{ijk}, \qquad\qquad r_i \sim N(0, \sigma^2_{run}), \tag{2}$$

where μ represents global process mean and distribution of location effects and random errors remains the same as in the first model. However, location effects and residuals are not necessarily i.i.d. We thus take into account the two following scenarios:

1. different variances among residuals w.r.t. to the run,
2. different variances among residuals and locations w.r.t. to the run.

The first scenario leads to use of random effect of the run on residuals in the model (2). We use logarithmic transformation of standard deviations to re-scale the values on the interval $(-\infty, \infty)$:

$$\epsilon_{ijk} \sim N(0, \sigma^2_{res,i}), \qquad log\sqrt{\sigma^2_{res,i}} = \alpha_{res} + s_i, \qquad s_i \sim N(0, \sigma^2_{runSD}). \tag{3}$$

In the second scenario we simply analogically add random effect of the run on location to the model (3). This time we don't use the logarithmic transformation to preserve the simplicity of implementation of the model in the brms package.

$$l_{ij} \sim N(0, \sigma^2_{loc,i}), \qquad \sigma_{loc,i} = \alpha_{loc} + t_i, \qquad t_i \sim N(0, \sigma^2_{runLocSD}). \tag{4}$$

We use Bayesian statistics to estimate the parameters and implement the models (3) and (4) using brms package created by Bürkner (2017,2018) in R by R Development Core Team (2023). All results in the next section were verified by direct implementation in Stan by Stan Development Team (2024). Application of random variance models is pretty straightforward and have been already used e.g. by Wright and Simon (2003) or Williams, Rodriguez and Bürkner (2021).

3 Results

Estimated parameters from model (3) can be found in Table 1 and from model (4) in Table 2. All estimates were calculated as posterior means from 3 chains with 4 000 iterations that followed 1 000 burn-in iterations. Differences between es-

TABLE 1. Model (3) with random residual error across batches.

Parameter	Interpretation	Estimate	95% Cred. Int.
μ	Global process mean	99.58	(98.87, 100.32)
σ_{run}	Run SD	1.11	(0.68, 1.88)
σ_{loc}	Location SD	0.32	(0.08, 0.50)
α_{res}	Log Average/Population Residual SD	0.08	$(-0.05, 0.21)$
$exp(\alpha_{res})$	Average/Population Residual SD	1.09	(0.96, 1.24)
σ_{runSD}	Run SD on Log Residual SD	0.16	(0.06, 0.31)

timates of common parameters of the models (3) and (4) are negligible. Interest arouses narrower credible interval for the population location s.d. (α_{loc}) in comparison to the location s.d. σ_{loc} from model (3) (Table 1). This result together with rather narrow credible interval for run s.d. on population location s.d. ($\sigma_{runLocSD}$) indicates advisable use of random effects on location effects as well as on residuals. This is confirmed by relatively high $\sigma_{runLocSD}$ in comparison to σ_{loc}.

TABLE 2. Model (4) with random residual and location error across batches.

Parameter	Interpretation	Estimate	95% Cred. Int.
μ	Global process mean	99.59	(98.88, 100.30)
σ_{run}	Run SD	1.10	(0.67, 1.85)
α_{loc}	Average/Population Location SD	0.41	(0.28, 0.55)
$\sigma_{runLocSD}$	Run SD on Population Location SD	0.29	(0.17, 0.44)
α_{res}	Log Average/Population Residual SD	0.07	$(-0.04, 0.18)$
$exp(\alpha_{res})$	Average/Population Residual SD	1.07	(0.96, 1.19)
σ_{runSD}	Run SD on Log Residual SD	0.12	(0.02, 0.27)

4 Discussion

Use and interpretation of random effects on variance are uncomplicated. Both of considered variations can indeed be very different in a developmental run and averaging them into a single value (as it is the case when homoscedasticity is assumed) could lead to overoptimistic assessment of the process performance. Attention should be given to heteroscedasticity within the random effects and not only to heteroscedasticity in the residual errors as it is often the case. Residual error variation represents different degrees of micromixing whereas location s.d. variation introduces different stability of the run. Finally, representation of the random effect variability via log transformation adds efficiency into the estimation, but requires more complex implementation within the existing software.

Acknowledgments: Special Thanks belongs to Martin Modrák , Department of Bioinformatics, 2nd Faculty of Medicine, Charles University, Prague, Czech Republic for providing technical details and backround on brms and Stan interface.

References

Bürkner, P.C. (2017). brms: An R Package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, **80**, 1 – 28.

Bürkner, P.C. (2018). Advanced Bayesian multilevel modeling with the R Package brms. *The R Journal*, **10**, 395 – 411.

R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Stan Development Team (2024). *Stan Modeling Language Users Guide and Reference Manual, Version 2.34*.

Williams, D.R., Rodriguez, J.E. and Bürkner, P.C. (2021). Putting variation into variance: modeling between-study heterogeneity in meta-analysis. *PsyArXiv*.

Wright, G. W. and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448 – 2455.

Using linear mixed models to compare a self-assessed frailty score with clinician assessed scores in patients approaching major surgery

Mohammad Sayari[1], James Durrand[2], Christopher Taylor[3], Jochen Einbeck[1], Ehsan Kharatikoopaei[4], Joshua Craig[5], Nathan Griffiths[2]

[1] Durham University, UK
[2] The Newcastle upon Tyne NHS Foundation Trust, UK
[3] The South Tees NHS Foundation Trust, UK
[4] Manchester Metropolitan University, UK
[5] Northumbria Healthcare NHS Foundation Trust

E-mail for correspondence: mohammad.sayari@durham.ac.uk

Abstract: Frailty is a syndrome of reduced physiological and cognitive reserve resulting in vulnerability to physiological insult and delayed recovery. It is a recognised predictor of poor perioperative outcomes. The Rockwood clinical frailty score (CFS) is a validated frailty screening tool based on the appearance of the patient in clinics. A study sponsored by South Tees Hospitals NHS Foundation Trust investigated whether patients may be able to self-assess their frailty utilizing a modified Rockwood CFS, by benchmarking the self-assessed scores with a clinician- and a researcher-assessed CFS score. A linear mixed-effects model, involving covariates such as age and ASA scores, was used to compare the CFS frailty scores and to identify any differences in their agreement. Linear mixed-effect model trees were also used for a better understanding of interactions of covariates and scorer effects.

Keywords: Frailty; Peri-operative care; Clinical frailty scale; Cohen's kappa statistic; Linear mixed-effects model.

1 Introduction

Frailty is a recognised predictor of poor perioperative outcomes (Lin et al., 2016). Preoperative assessment of frailty is key to allow planning of perioperative care,

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

and discussions with patients to manage risk, expectations, and facilitate shared decision-making and informed consent.

The Rockwood clinical frailty score (CFS) is a validated scoring system-based global clinical impression of frailty based on the appearance of the patient in clinic. It is in routine use in patients over 64 in the perioperative and wider clinical settings. The CFS groups patients into 9 classes ranging from very fit to severe frailty, each allocated a numerical value of 1-9, increasing with rising frailty (Rockwood et al., 2005). Typically, a person allocated a score of 1-3 is labelled as 'non-frail'. A person scoring 4 is labelled 'pre-frail', a score of 5-8 is 'frail' and 9 is 'terminally ill'.

Recently, drivers toward a digitalised NHS, along with the COVID-19 pandemic, have encouraged remote clinical working and telemedicine to deliver patient care. This limits the applicability of the CFS without a face-to-face patient contact, removing a key component of comprehensive preoperative assessment.

A surrogate marker for frailty is required. We propose that patients may be able to self-assess their frailty utilizing a modified Rockwood CFS. If patient self-assessment is feasible and agreement with clinician assessed CFS is acceptable, this would be a stepping stone to wider validation and utilisation as a remotely delivered preoperative frailty assessment tool.

2 Methods

Initially, agreement between CFS frailty scores was examined using the quadratic weighted Cohen's Kappa. Values for levels of agreement using the Kappa coefficient are interpreted as follows: < 0 = no agreement; 0.00-0.20 = slight agreement; 0.21-0.40 = fair agreement; 0.41-0.60 = moderate agreement; 0.61-0.80 = substantial agreement; 0.81-1.00 = almost perfect agreement (Landis et al., 1977).

However, such an analysis does not allow for the investigation of covariate effects such as age or ASA score, on the strength of agreement between scores. Hence, a linear mixed-effects model was set up to compare the CFS frailty scores. The linear mixed-effects model allows assessing covariate impacts and interactions when comparing CFS frailty scores, and accounts for intra-patient correlation using a patient-level random effect, hence enabling the computation of robust standard errors to minimise the likelihood of false conclusions. We consider the scores produced by the patient, clinician, and researcher, as pertaining to assessment groups $j = 0, 1$ and 2, respectively. We denote by y_{ij} the measured score for patient i on group j, and by the vector x_i any covariates of interest for patient i. Then a linear mixed-effects model can be formulated as:

$$y_{ij} = \sum_{j=1}^{2} \gamma_j 1_{\{\text{group}=j\}} + x_i^T \beta + u_i + \varepsilon_{ij},$$

where the terms involving the γ_j and β are fixed effects, and u_i is a patient-level random intercept. In the summation term, the patient self-assessment ($j = 0$) serves as the reference category. The fixed effect parameters γ_j capture the agreement differences of interest. Additionally, and not displayed here notationally, we considered models using interaction terms between the grouping variables and the covariates age and ASA score. The linear mixed-effects models were fitted using function `lmer` in R package **lme4**.

In addition, for a more comprehensive understanding of the interaction between covariates and scorers, we used linear mixed-effects model trees. The GLMM tree algorithm is an extension of the model-based recursive partitioning (MOB) method. The MOB method uses a parameter instability test to select partitioning variables. However, MOB is not suitable for multilevel data. To address this limitation, the GLMM tree algorithm was developed to incorporate random effects into the analysis (Fokkema et al., 2018). While random effects are estimated globally using all observations, the fixed effects are estimated locally. The dataset is partitioned based on additional covariates or partitioning variables, and fixed effects are estimated for each partition cell. The GLMM tree model was estimated using the function `lmertree` from the R package **glmertree**.

3 Results

All patients aged 65 or over who were listed for major surgery were included in the study ($n = 80$). Table 1 presents the inter-rater reliability of the CFS frailty scores using Cohen's Kappa. The results demonstrate a moderate agreement between patient-allocated self-score and pre-assessment score on the 9-level scale ($\kappa = 0.43$). There was also a moderate agreement between the pre-assessment score and the research team score on the 9-level scale ($\kappa = 0.59$). There was a substantial agreement on the 9-level scale CFS between the patient-allocated self-score and the research team score ($\kappa = 0.62$). On the 3-level scale, the results indicate a fair agreement between the patient-allocated self-score and pre-assessment score ($\kappa = 0.32$). There was a substantial agreement on the 3-level scale CFS between the patient-allocated self-score and the research team score ($\kappa = 0.68$). There was a moderate agreement between the pre-assessment score and the research team score on the 3-level scale ($\kappa = 0.55$).

TABLE 1. Inter-rater reliability on 9-point and 3-level clinical frailty score.

Agreement measured (Kappa statistics)	9-point scale	3-level scale
Patient allocated self-score vs. pre-assessment score	0.433	0.319
Patient allocated self-score vs. research team score	0.622	0.683
Pre-assessment score vs. research team score	0.591	0.554

All P-values < 0.01

Table 2 represents the results of the linear mixed-effect model. The results show that the patient-allocated self-scores were higher than pre-assessment scores (model 1, $p = 0.015$). There were no significant differences between the patient-allocated self-score and the research team score (model 1, $p = 0.588$). In model 2, patient-assessed scores tend to be higher than the other ones, but older patients (age $> 74y$) behave differently than younger patients in the sense that older patients do not assess themselves frailer than the other scores would indicate. The results for interaction between ASA (American society of anesthesiology) and groups (model 3) indicate that there were no significant differences between patients with $ASA \geqslant 3$ and $ASA < 3$ when comparing the pre-assessment and research team with patient-allocated self-scores.

TABLE 2. Results of linear mixed-effects models.

Model	Formula	Fixed effect	Effect Estimate (β coefficient)	Standard error	P-value
1	CFS score $\sim$ group + $(1\|ID)$	(Intercept)	3.539	0.126	< 0.001
		group (pre-assessment)	-0.289	0.117	0.015
		group (research team)	-0.064	0.117	0.588
2	CFS score $\sim$ group + age$>$74 + group:age$>$74 + $(1\|ID)$	(Intercept)	3.498	0.164	< 0.001
		group (pre-assessment)	-0.498	0.154	0.001
		group (research team)	-0.253	0.154	0.101
		age$>$74	0.084	0.25	0.737
		group (pre-assessment):age$>$74	0.487	0.235	0.039
		group (research team):age$>$74	0.443	0.235	0.06
3	CFS score $\sim$ group + ASA $\geqslant$ 3 + group:ASA $\geqslant$ 3 + $(1\|ID)$	(Intercept)	3.171	0.176	< 0.001
		group (pre-assessment)	-0.229	0.171	0.184
		group (research team)	-0.200	0.171	0.245
		ASA $\geqslant$ 3	0.660	0.24	0.006
		group (pre-assessment):ASA $\geqslant$ 3	-0.114	0.235	0.629
		group (research team):ASA $\geqslant$ 3	0.236	0.235	0.316

Reference category: patient self-assessment score

CFS clinical frailty score, ASA American society of anesthesiology

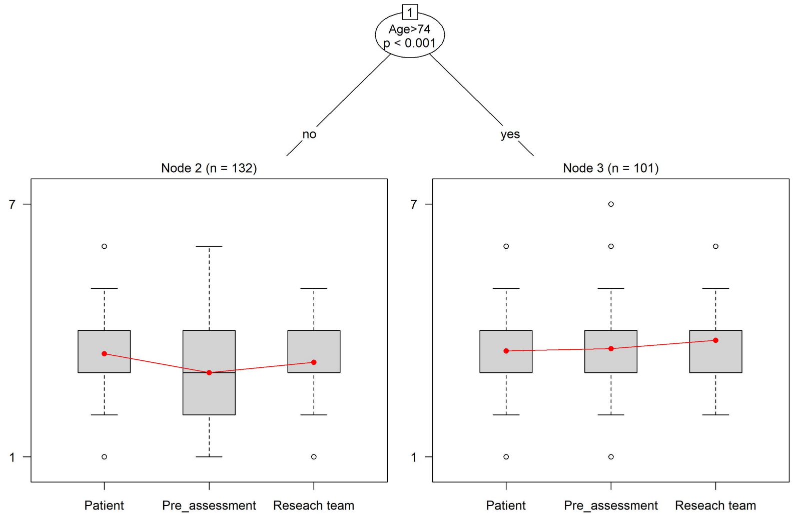


FIGURE 1. Fitted linear mixed-effects model tree for model 2

To gain a deeper understanding of how covariates and scorers interact in models 2 and 3, we used the GLMM tree model. The GLMM trees for models 2 and 3 are shown in Figures 1 and 2, respectively. In each inner node of the plotted trees, the splitting variable and corresponding p-value from the parameter stability test are reported. The diagram in each figure shows two terminal nodes for CFS. In Figure 1, Node 2 shows that patients 74 years of age or younger had higher self-assessment scores compared to pre-assessment and research team scores. In node 3 (patients over 74), there were no substantial differences between scores. In Figure 2, patients with ASA scores under 3 had slightly higher self-assessment scores than pre-assessment and research team scores, and in node 4, patients

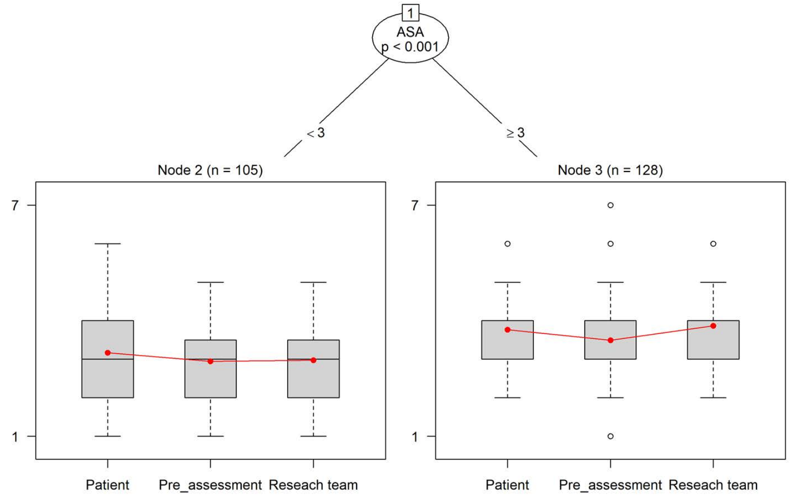


FIGURE 2. Fitted linear mixed-effects model tree for model 3

with ASA scores of 3 or higher had slightly lower pre-assessment compared to the other scores. Please note that the p-values displayed in the top node indicate the significance of the split; that is the existence of subgroups with differing behavior. They make no statement on significant differences of scorer types within those subgroups.

4 Conclusion

In this study, we evaluated the use of patient self-assessment as a surrogate marker for clinician-assessed frailty. Our findings suggest that patients can evaluate their frailty by using a modified Rockwood CFS. In an additional analysis, we also assessed the agreement between CFS frailty scores using the intraclass correlation coefficient and Bland–Altman plots. All of the results confirmed that there was an acceptable agreement between the self-scores allocated by the patients and the research team scores, with some tendency for relatively younger patients to assign themselves larger frailty scores.

References

Fokkema, M., Smits, N., Zeileis, A., et al. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior research methods*, **50**, 2016–2034.

Landis, J.R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Lin, H.S., Watts, J.N., Peel, N.M. and Hubbard, R.E. (2016). Frailty and post-operative outcomes in older surgical patients: a systematic review. *BMC Geriatrics*, **16**, 1–12.

Rockwood, K., Song, X., MacKnight, C., Bergman, H., et al. (2005). A global clinical measure of fitness and frailty in elderly people. *Canadian Medical Association Journal*, **173**, 489 – 495.

Function-on-scalar regression via first-order gradient-based optimization

Quentin Edward Seifert[1], Elisabeth Bergherr[1], Tobias Hepp[1,2]

[1] University of Göttingen, Germany
[2] Friedrich-Alexander-Universität Erlangen-Nuremburg, Germany

E-mail for correspondence: quentinedward.seifert@uni-goettingen.de

Abstract: Functional regression models can quickly become computationally expensive with either the response or some covariates being functional. We suggest to use gradient descent based optimization algorithms to estimate such models as an easily scalable alternative to established approaches. Preliminary simulation results show our approach to perform reliably. We apply our approach to supermarket parking data recorded during the first months of the Covid-19 pandemic in Germany.

Keywords: Functional regression; Gradient descent; Covid-19 restrictions.

1 Introduction

Functional regression models are models that are able to include functional observations. This framework encompasses datasets with functional covariates and scalar responses, functional responses with scalar covariates or datasets in which both, response and covariates, are functional (Ramsay and Silverman, 2005; Reiss et al., 2010). Due to the nature of the data, the models can already become computationally expensive with a comparably low number of observations. To accommodate this increased complexity, we introduce gradient descent based functional regression. The idea is to fit functional regression models using gradient descent based optimization algorithms and estimate the model parameters as one would estimate the parameters of neural networks.

In the following sections we introduce the general framework, demonstrate the applicability of our approach on simulated data and finally apply our model to supermarket parking data to analyze the effect of Covid-19 restrictions on supermarket visits during the spring of 2020 in Germany.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

2 Methods

Function-on-scalar regression models with functional response $y(t)$ and scalar covariates can be represented as

$$\boldsymbol{y}(t) = \boldsymbol{Z}\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}(t)$$

where $\boldsymbol{y}(t)$ is an N- dimensional vector containing the functional responses, $\boldsymbol{Z}$ is a $N \times q$ design matrix and $\boldsymbol{\beta}(t)$ is a vector of coefficient functions defined as $\boldsymbol{\beta}(t) = (\beta_1(t), ..., \beta_q(t))^T$ that are represented by linear combinations of basis functions $\boldsymbol{\theta}(t) = (\theta_1(t), ..., \theta_K(t))$ such that

$$\beta_j(t) = \boldsymbol{b}_j^T \boldsymbol{\theta}(t) \; .$$

In matrix notation, the model becomes

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{\Theta}^T + \boldsymbol{E} \; ,$$

with $\boldsymbol{Y}$ containing the functional responses collected at $t_1, ..., t_n$, $\boldsymbol{B}$ containing the coefficients to be estimated and $\boldsymbol{\Theta}$ being matrix of evaluated basis functions. Applying the vec-operator and expressing the model in terms of Kronecker products, this model can be expressed in the form of a standard regression model

$$\text{vec}(Y^T) = \text{vec}\left[(\boldsymbol{Z}\boldsymbol{B}\boldsymbol{\Theta}^T)^T\right] + \boldsymbol{E}$$

$$\text{vec}(Y^T) = (\boldsymbol{Z} \otimes \boldsymbol{\Theta})\text{vec}(\boldsymbol{B}^T)$$

with $\boldsymbol{Z} \otimes \boldsymbol{\Theta}$ as the design matrix (Reiss et al., 2010). Usually, the coefficient functions $\beta_k(t)$ are estimated using penalized splines. The framework is not limited to simple scalar covariates as depicted here and can furthermore be augmented to include multi-dimensional splines to model interactions between t and covariates. Since a single observation actually consists of n observations, these models can become computationally expensive very easily. For large datasets, we therefore propose gradient descent based functional regression. In our approach, the models are set up as usual, the model parameters are then trained as if they were the weights and biases of a neural network without hidden layers. Rather than relying on the conventional fitting methods which require the whole dataset every step of the way, the use of optimizers using mini-batching allows the models to scale really well. Their reliance on only the gradient of the loss function furthermore makes the models very easy to setup and flexibly adjustable.

3 Simulations

We demonstrate the general functionality of our model based on a simulation setup adapted from the R-package `refund` (Goldsmith et al., 2023) with a functional response $y(t)$ with $N = 50$ and $n = 40$, a functional intercept $\beta_0(t)$ and a single scalar covariate x with the corresponding coefficient function $\beta_1(t)$ such that

$$y(t) = \beta_0(t) + x\beta_1(t) + \epsilon(t)$$

with $\epsilon \sim \text{N}(0, 1)$. The model is set up with cubic regression splines with $k = 10$ and fitted by minimizing the negative log-likelihood of a Gaussian distribution using the Adam-optimizer (Kingma and Ba, 2014). The left panel of Figure 1 shows a few exemplary observations. Each curve in the plot depicts a single observation. The other two panels show the results of 100 simulation runs using our proposed fitting methods and confirm the general ability of our approach to appropriately deal with functional data. Throughout, the model is able to reliably approximate the true coefficient functions.

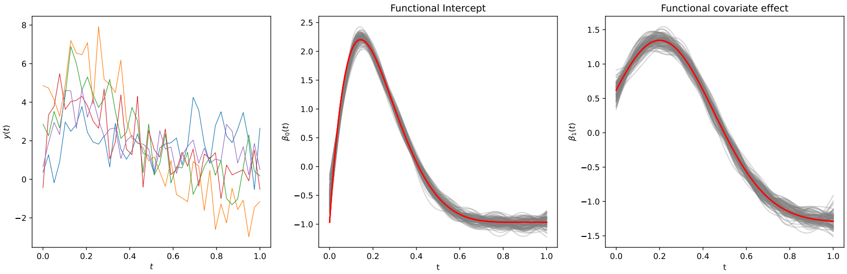


FIGURE 1. The left panel shows 5 exemplary observations, each depicted by one line, the center and right panel show the functional intercept and the functional covariate effect.

4 Covid parking application

We apply our method to data provided by Smart City System Parking Solutions, a company that produces smart parking sensors which record the occupancy status of parking spots. The dataset consists of 89 supermarket parking lots from all over Germany. The data spans 120 days, starting on February 1, around the time Covid-19 was already known to be on the rise but before the introduction of the extensive contact restrictions, until May 31. Most but not all lots were observed every day. We exclude the obervations from 0:00 to 6:00 and aggregate the used parking spots on the respective lots every 15 minutes and use their utilization levels as the response. With $N = 10,431$ and $n = 64$, the dataset effectively consists of $10431 * 64 = 667{,}584$ observations. We fit a model with a functional intercept which is specified to be a cubic regression spline and a tensor-product spline to estimate the interaction of time of day and time within the year and assume the response to follow a Beta-distribution. The predictor for the mean is estimated as

$$\eta(t) = \alpha + \beta_0(t) + f(day, t) \ .$$

where α is a regular intercept. The results are depicted in Figure 2. The functional intercept shows the expected pattern of high utilization during the day and low utilization during the early morning and late evening. The contour plot shows some interesting patterns, such as shifts towards overall fewer supermarket visits, especially in the afternoon, following the introduction of the restrictions in March 2020.

While efficiently programmed libraries such as `refund` are still able to estimate the above model despite the already large amount of data points totaling more 600,000, extending the time frame to several years will drastically increase the volume and ultimately make the application of stochastic gradient descent methods indispensable.

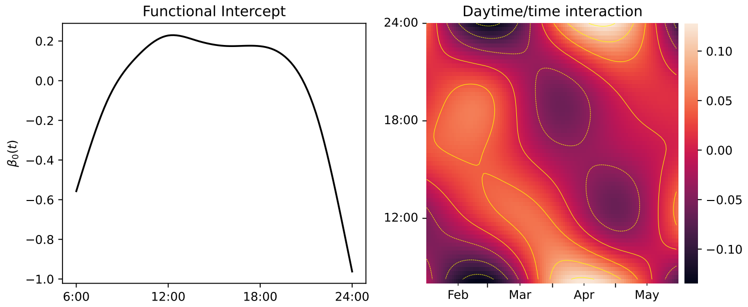


FIGURE 2. Functional intercept and interaction of day and daytime as a smooth surface.

References

Goldsmith J., Scheipl F., Huang L., Wrobel J., Di C., Gellar J., Harezlak J., McLean M.W., Swihart B., Xiao L., Crainiceanu C. and Reiss P.T. (2023). refund: Regression with Functional Data.
R package version 0.1-34, *https://CRAN.R-project.org/package=refund.*

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, **1**, 297 – 310. (2014).

Kingma, D.P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint: arXiv:1412.6980.*

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer NY, New York.

Reiss, P.T., Huang, L. and Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *The International Journal of Biostatistics* **6**, 28.

Confidence intervals for tree-structured varying coefficients based on parametric bootstrap

Nikolai Spuck[1], Matthias Schmid[1], Moritz Berger[1]

[1] Institute of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Germany

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

Abstract: The tree-structured varying-coefficient model (TSVC) is a flexible regression approach that allows the effects of covariates to vary with the values of the effect modifiers. Relevant effect modifiers are identified inherently using recursive partitioning. To quantify uncertainty in TSVC models, we propose a procedure to construct confidence intervals. This task constitutes a selective inference problem as the coefficients of a TSVC model result from the data-driven model building. To account for this issue, we introduce a parametric bootstrap approach, which is tailored to the complex structure of TSVC. Coverage proportions of the proposed confidence intervals were evaluated in a simulation study. For illustration, we considered an application to data from patients suffering from Covid-19.

Keywords: Varying coefficents; Tree-based modeling; Selective inference; Parametric bootstrap.

1 Tree-structured varying coefficients

The classical varying coefficient model proposed by Hastie and Tibshirani (1993) allows the modeling of an outcome variable Y in a generalized regression framework with predictor function

$$\eta(\boldsymbol{X}, \boldsymbol{Z}) = \beta_0 + \beta_1(Z_1)X_1 + \ldots + \beta_p(Z_p)X_p\,, \tag{1}$$

where $Z_1, \ldots, Z_p$ denote random variables that serve as *effect modifiers* and change the linear effects of $X_1, \ldots, X_p$ through an unspecified functional form. The model in (1) requires the effect modifiers to be specified beforehand, and each varying coefficient may only depend on one effect modifier. To address these limitations, Berger et al. (2019) proposed the tree-structured varying coefficient (TSVC) model, which applies a recursive partitioning technique to inherently

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

detect relevant effect modifiers. The predictor function of a TSVC model $\mathcal{M}$ is given by

$$\eta^{\mathcal{M}}(\boldsymbol{X}) = \beta_0^{\mathcal{M}} + \beta_1^{\mathcal{M}}(\boldsymbol{X}_{[-1]})X_1 + \ldots + \beta_p^{\mathcal{M}}(\boldsymbol{X}_{[-p]})X_p\,, \tag{2}$$

where $\boldsymbol{X}_{[-j]}$ denotes the set of covariates $X_1, \ldots, X_p$ excluding X_j. Accordingly, the effect of each covariate can be modified by each other covariate except itself. The functions $\beta_j^{\mathcal{M}}(\cdot)$ are each determined by a tree structure. This means that each function $\beta_j^{\mathcal{M}}(\cdot)$ sequentially partitions the observations into disjoint subsets N_{jm}, $m = 1, ..., M_j$, based on the values of the selected effect modifiers and assigns a different regression coefficient for X_j to each partition N_{jm}. These functions can be written as

$$\beta_j^{\mathcal{M}}(\boldsymbol{X}_{[-j]}) = \sum_{m=1}^{M_j} \beta_{jm}^{\mathcal{M}} I(\boldsymbol{X}_{[-j]} \in N_{jm})\,. \tag{3}$$

where $I(\cdot)$ denotes the indicator function. Hence, the structure of TSVC model $\mathcal{M}$ is characterized by the set of partitions $\mathcal{M} = \{\{N_{jm}\,, m = 1, \ldots, M_j\}\,, j = 1, \ldots, p\}$. Each coefficient is derived from binary splits partitioning the observations of one parental node into two child nodes iteratively.

In each step of the TSVC fitting algorithm, the best splitting rule among all covariates X_j, respective candidate splitting variables $X_k, k \neq j$, and possible split points is selected, starting from a linear predictor without varying coefficients. For this, all candidate models with one additional split are evaluated and the best performing one is selected based on the minimal deviance. From the resulting sequence of hierarchical models, an optimal model can finally be selected using a likelihood-based measure, e.g. the Bayesian information criterion (BIC).

2 Selective confidence intervals

Our objective is to construct confidence intervals (CIs) for the best approximating varying linear effects $\boldsymbol{\beta}^{\mathcal{M}}$ of a TSVC model $\mathcal{M}$ fitted to the data $\mathcal{D} = \{(y_i, \boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})^T), i = 1, \ldots, n\}$.

In order to do so, we need to account for the fact that the coefficients of interest $\beta_{jm}^{\mathcal{M}}$ arise out of model structure $\mathcal{M}$ being selected by the TSVC fitting procedure, that is, that the model selection event $\widehat{\mathcal{M}} = \mathcal{M}$ occurred. Therefore, a $100(1-\alpha)\%$ CI of $\beta_{jm}^{\mathcal{M}}$ is supposed to satisfy

$$\mathbb{P}\left(\beta_{jm}^{\mathcal{M}} \in CI_{jm}^{\mathcal{M}} \mid \widehat{\mathcal{M}} = \mathcal{M}\right) \geq 1 - \alpha\,, \tag{4}$$

which constitutes a so-called *selective inference* or *post selection inference* problem (Berk et al., 2013, Fithian et al, 2014). In linear regression models with LASSO penalization Lee et al. (2016) found that if the selection event $\widehat{\mathcal{M}} = \mathcal{M}$ can be characterized by a set of inequalities that fulfill specific condtions, $\widehat{\mathcal{M}} = \mathcal{M}$ constitutes a *linear selection event* and exact statistical inference of the coefficients conditional on the selection event can be performed. Specification of the selection event for TSVC models, however, would require a vast number of inequalities. The main reason is that the TSVC algorithm involves the fitting of several trees, which is considerably more complex than fitting of a single tree or a predictor function with interactions of predefined order (scenarios investigated by Neufeld et al., 2022, and Suzumura et al., 2017). In the first iteration of the

TSVC algorithm, the event of selecting one specific splitting rule is characterized by $p(p-1)n$ inequalities, assuming p continuous covariates with n possible split points each. Overall, $\mathcal{O}(np^2S)$ inequalities are required to describe the selection of one specific sequence of nested TSVC models $\mathcal{M}^{[s]}, s = 1, \ldots, S$, and an optimal model $\mathcal{M}$ out of it. Since there are cases where the same model structure $\mathcal{M}$ can be described by a number of different sequences of nested models (e.g. when the same splits are performed in a different order), the conditioning set that characterizes the selection event $\widehat{\mathcal{M}} = \mathcal{M}$ is a union of sets. Moreover, after the selected number of splits is reached, any combination of additional splitting rules can result in the same model structure as long as the respective models' BIC is larger than the BIC of the selected model.

To tackle this complex mechanism and construct a CI for the effect $\beta_{jm}^{\mathcal{M}}$ satisfying Equation (4), we propose a parametric bootstrap approach tailored to the selective inference problem at hand. Specifically, we compute estimates for $\beta_{jm}^{\mathcal{M}}$ from a set of bootstrap samples $\mathcal{D}_b, b = 1, \ldots, B$. Simply enforcing the structure of the original model $\mathcal{M}$ on each sample neglects the uncertainty induced by the data-driven model building, resulting in CIs that tend to be too short. To account for this uncertainty, we first apply the TSVC fitting procedure to the samples $\mathcal{D}_b$, resulting in B different models $\mathcal{M}_b$ with predictor functions

$$\eta^{\mathcal{M}_b}(\boldsymbol{X}) = \hat{\beta}_0^{\mathcal{M}_b} + \hat{\beta}_1^{\mathcal{M}_b}(\boldsymbol{X}_{[-1]})X_1 + \ldots + \hat{\beta}_p^{\mathcal{M}_b}(\boldsymbol{X}_{[-p]})X_p. \tag{5}$$

Secondly, we determine an estimate of the effect $\beta_{jm}^{\mathcal{M}}$ from the original model based on bootstrap sample $\mathcal{D}_b$ by averaging the node-specific effect estimates $\hat{\beta}_{jm}^{\mathcal{M}_b}$ with regard to partition N_{jm} from the original model yielding

$$\bar{\beta}_{jm}^{(b)} = \frac{1}{|N_{jm}|} \sum_{i:\boldsymbol{x}_i \in N_{jm}} \hat{\beta}_j^{\mathcal{M}_b}(\boldsymbol{x}_{i[-j]})\,. \tag{6}$$

By definition of (6), for covariate X_j each observation is assigned to one of the subsets N_{jm} that was identified by the original model $\mathcal{M}$, and subsequently the average value of the function $\hat{\beta}_j^{\mathcal{M}_b}(\cdot)$ from model $\mathcal{M}_b$ across the observations in N_{jm} is calculated. Finally, a $100(1-\alpha)\%$ CI for $\beta_{jm}^{\mathcal{M}}$ is obtained by computing the $\alpha/2$ and $(1-\alpha/2)$ percentiles from the bootstrap estimates $\bar{\beta}_{jm}^{(1)}, \ldots, \bar{\beta}_{jm}^{(B)}$.

Equation (6) allows to determine bootstrap estimates of the coefficients of interest $\beta_{jm}^{\mathcal{M}}$. Yet, calculating these estimates does in itself not condition on the model selection event. In order to mimic the conditioning on $\widehat{\mathcal{M}} = \mathcal{M}$, we apply a parametric bootstrap scheme to generate bootstrap samples $\mathcal{D}_b$, where the bootstrap values of the outcome variable $y_i^{(b)}$ are drawn from the conditional distribution of $Y|\, \boldsymbol{X} = \boldsymbol{x}_i$ indicated by the fitted TSVC model $\mathcal{M}$. That is, the new outcome values $y_i^{(b)}$ are generated from a distribution with expectation

$$\mathbb{E}(Y|\, \boldsymbol{X} = \boldsymbol{x}_i) = g^{-1}\left(\eta^{\mathcal{M}}(\boldsymbol{x}_i)\right)\,, \tag{7}$$

where $\eta^{\mathcal{M}}(\cdot)$ is the predictor function of the original TSVC model $\mathcal{M}$ fitted to data $\mathcal{D}$.

3 Empirical evaluations

Coverage proportions of the proposed CIs were assessed in a simulation study and compared to those of classical asymptotic normal distribution-based Wald

TABLE 1. Coverage proportions of 95% CIs for TSVCs averaged across all effects.

DGP	CI method	n	200		500		1000	
		σ	1	2	1	2	1	2
Linear	Wald		.833	.833	.875	.875	.900	.900
	Parametric percentile		.951	.951	.949	.949	949	949
TSVC	Wald		.795	.795	.843	.852	.865	.867
	Parametric percentile		.968	.971	.966	.970	.963	.965

TABLE 2. Restults of fitting the TSVC model to the Covid-19 patient data.

Variable	Partition	$\hat{\beta}$	$\exp(\hat{\beta})$	95% CI of $\exp(\hat{\beta})$
Age	—	0.008	1.008	[1.006; 1.051]
Treatment antibodies	Age$\leq$ 60	-2.924	0.054	[0.000; 0.252]
	Age$>$ 60	-0.986	0.373	[0.001; 8.629]

CIs. We considered a non-varying linear data generating process (DGP) with one informative variable X_1 and one noise variable X_2

$$y_i = 0.25\, x_{i1} + \varepsilon_i,\ i = 1, \ldots, n\,, \tag{8}$$

where $x_{i1}, x_{i2} \sim N(0, 1)$. We also considered a TSVC DGP

$$y_i = 0.5\, I(x_{2i} > 0.5)x_{i1} - I(x_{2i} \leq 0.5 \wedge x_{i3} = 0)x_{i1} + \varepsilon_i,\ i = 1, \ldots, n\,, \tag{9}$$

where $x_{i3} \sim \text{Bin}(1, 0.5)$. For both DGPs, we set $\varepsilon_i \sim N(0, \sigma^2)$ with standard deviations $\sigma \in \{1, 2\}$ and sample sizes $n \in \{200, 500, 1000\}$. The proposed parametric percentile CIs were based on $B = 1000$ bootstrap samples. The average coverage proportions were calculated based on $R = 5000$ replications. The average coverage proportion across all covariates was calculated as

$$C_{\text{av}} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{p}\sum_{j=1}^{p}\frac{1}{M_j^r}\sum_{m=1}^{M_j^r} I\left(\beta_{jm}^{\mathcal{M}_r} \in CI(\beta_{jm}^{\mathcal{M}_r})\right)\,, \tag{10}$$

where $\mathcal{M}_r$ denotes the TSVC model fitted in the r-th replication and M_j^r denotes the number of coefficients of X_j in model $\mathcal{M}_r$.

The results in Table 1 for the linear DGP show that the proposed CIs yield coverage proportions close to the nominal level whereas coverage proportions of the Wald CIs are far too low but increased with larger sample size. Of note, even with this simple underlying linear DGP without varying effects, neglecting the fact that constructing CIs for TSVCs is a selective inference problem (e.g. by applying a naive Wald type CI) may yield highly anti-conservative results with low coverage. For the TSVC DGP, the proposed CIs tended to be rather conservative but approached the nominal level for larger sample size and lower noise. The naive Wald type approach resulted in insufficient coverage proportions across all settings.

In addition, we applied a logistic TSVC model to data from Covid-19 patients (Huebner et al., 2023). The main objective was to investigate the effect of age and of treatment with antibodies on the need for oxygen support and to detect possible interactions between treatment effect and age. The results in Table 2 indicate a linear non-varying effect of age and a treatment effect modified by age. Based on the proposed CIs, a significant effect of age and of treatment with antibodies for patients aged 60 years or younger at significance level $\alpha = 0.05$ was shown but no significant treatment effect for patients older than 60 years.

4 Summary

TSVC models are flexible tools for generalized regression that allow the linear effects of the covariates to vary with the effect modifiers. They can be fitted using the eponymous R add-on package (Berger, 2021). The TSVC fitting procedure is able to inherently detect relevant effect modifiers and relaxes the prerequisite that effect modifiers need to be specified before model fitting. Constructing CIs for TSVCs is a selective inference problem as statistical inference is performed after model selection. In this vein, we proposed parametric bootstrap-based method tailored to the complex selection mechanism of TSVC as an approximate solution. The applications to real-world data from COVID-19 patients showed that the proposed CIs may differ strongly from naive Wald type CIs and lead to different conclusions when assessing statistical significance of the coefficients. The effect of CVIV in the group of elderly patients is highly clinically meaningful. This highlights that accounting for the selective inference problem is essential when statistical inference on the parameters of a TSVC model is of interest. In the simulation study, our approach yielded coverage proportions close to the nominal level for the linear DGP whereas the simple Wald type CI showed insufficient coverage. In the more complex scenario with varying coefficnets, the proposed approach showed slightly conservative results, while Wald type CIs yielded coverage proportions that were far too low.

References

Berger, M. (2021). TSVC: Tree-Structured Modeling of Varying coefficnets. *https://CRAN.R-project.org/package=TSVC*, R package version 1.5.3.

Berger, M., Tutz, G. and Schmid, M. (2019). Tree-structured modeling of varying coefficients. *Statistics & Computing*, **29**, 217 – 229.

Berk R., Brown, L., Buja, A., Zhang, K. and Thao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, **41**, 802 – 837.

Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. *arXiv 1410.2597*

Hastie, T. and Tibshirani, R. (1993). Varying coefficient models. *Journal of the Royal Statistical Society, Series B*, **55** 757 – 779.

Huebner, Y.R., Spuck, N., Berger, M. et al. (2023). Antiviral treatment of covid-19: which role can clinical parameters play in therapy evaluation? *Infection*, **51**, 1855 – 1861.

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

Neufeld, A.C., Gao, L.L. and Witten, D. M. (2022). Tree-values: Selective inference for regression trees. *Journal of Machine Learning and Research*, **23**, 1 – 43.

Suzumura, S., Nakagawa, K., Umezu, Y., Tsuda, K. and Takeguchi, I. (2017). Selective inference for sparse high-order interaction models. In *International Conference on Machine Learning*, 3338 – 3347.

A model-based boosting approach to deal with dependent censoring

Annika Strömer[1], Nadja Klein[2], Andreas Mayr[1]

[1] Department of Medical Biometrics, Informatics and Epidemiology, University of Bonn, Bonn, Germany

[2] Chair of Uncertainty Quantification and Statistical Learning, Research Center Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technische Universität Dortmund), Dortmund, Germany

E-mail for correspondence: `stroemer@imbie.uni-bonn.de`

Abstract: In survival analysis, censoring is an inherent observation that is usually assumed to be unrelated to the event of interest. When this assumption is not fulfilled, traditional methods like the Cox model may yield skewed or biased results. For example, if a patient's health deteriorates and the patient chooses to withdraw from the trial due to a poor prognosis, the time of censoring depends on the patient's health status. To deal with dependent censoring, in this work we propose to utilize distributional copula regression via model-based boosting. This approach allows to model the joint distribution of survival and censoring times by linking appropriately marginal distributions for T and C through a parametric copula. Rather than assuming the marginals are known, all distribution parameters (including the copula parameter) are estimated simultaneously as functions of (potentially different) covariates. A key merit of boosting is that estimation is even feasible for high-dimensional data with $p > n$, when classical estimation frameworks easily meet their limits. In addition, the boosting algorithm includes data-driven variable selection. To investigate the performance of our approach under controlled conditions, we first conduct a simulation study. Furthermore, we illustrate its practical application analysing the survival of colon cancer patients from an observational study.

Keywords: Copula; Dependent censoring; Model-based boosting.

1 Introduction

In time-to-event analysis, it is inherent that we encounter censored observations, whereby some patients in a study are not observed until the occurrence of the event of interest. Most widely used approaches, such as the Kaplan-Meier estimator or the Cox proportional hazard model, can handle time-to-event data with,

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

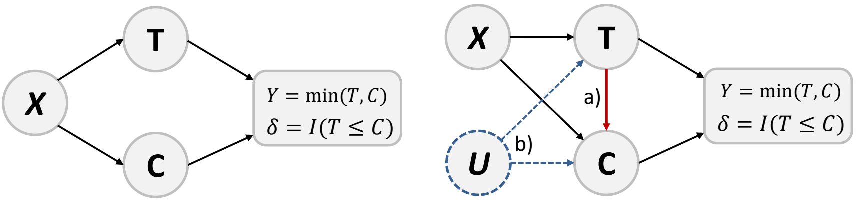


FIGURE 1. Graphical representation of two survival scenarios, showing the dependency between covariates $\boldsymbol{X}$ and the event time T and the censoring time C. Here, Y is the last observed time point and δ is the event indicator. The left-hand graph shows the case of independent censoring and the right-hand graph shows two cases of dependent censoring: a) direct dependence between T and C and b) through an unmeasured covariate U.

for example, right-censored observations. In these approaches, it is assumed that the survival time T and censoring time C are statistically independent for given covariates, see Figure 1 (left) for a graphical illustration.

However, this assumption may not be fulfilled in many situations, e.g. if patients. withdraw from a study due to their poor state of health. In these cases classical approaches that assume independent censoring could lead to biased results. This is due to a direct link between the survival time and the censoring time shown with a) in Figure 1 (right). Thus, if we assume that sicker patients drop out of the study due to poor health, then censored patients are more likely to be sicker than the non-censored patients, and the survival time of the patients may be overestimated. A further dependence can exist if it is caused by unobserved confounding variables, as shown in b) in the right graph of Figure 1, where the unobserved variable U influences survival and censoring time. Therefore, it is of great importance to take the dependency in such cases into account. In practice, however, one observes either the event or censoring, which makes it hard to verify the dependency.

Copula models linking two random variables by specifying their dependence structure (Midtfjord et al., 2022; Czado and Van Keilegom, 2023), gained increasing interest in recent years. Most of the proposed models in the literature for dependent censoring rely on a completely known copula. However, in practice, the association parameter is often unknown and can have a major influence on the resulting estimators of the marginal distributions (Huang and Zhang, 2008).

2 Methodology

In the following, we propose a boosting approach based on copulas to deal with dependent censoring without an assumed copula. Let T be the survival time and C the censoring time and based on the assumption of random right censoring, we have $Y = \min(T, C)$ and $\Delta = I(T \leq C)$, where $\mathcal{I}(A) = 1$ if A is true and the event was observed. According to Sklar's theorem, we can express the joint conditional cumulative distribution function (CDF) of T and C given covariate information $\boldsymbol{x}$ as

$$F_{T,C}(t, c|\boldsymbol{\alpha}) = C\{F_T(t|\boldsymbol{\theta_T}), F_C(c|\boldsymbol{\theta_C})|\theta\},$$

where $\boldsymbol{\alpha} = (\boldsymbol{\theta_T}, \boldsymbol{\theta_C}, \theta)^T$ is the vector of model parameters and $F_T(t|\boldsymbol{\theta_T})$ and $F_C(c|\boldsymbol{\theta_C})$ are the marginal CDFs, which are non-negative, continuous and parametric. The copula $C(.,\theta)$ is uniquely defined and parametric with a copula parameter θ that describes the dependency between T and C. Due to identifiability of the complete model, we consider the log-normal and Weibull distribution as marginal distributions and the Clayton, Gauss and Gumbel copula (Czado and Van Keilegom, 2023).

For estimation, we assume that we have an independent and identical distributed sample $\mathcal{D} = \{(y_i, \delta_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$. Then the joint log-likelihood is

$$\begin{aligned}\ell(\alpha; \mathcal{D}) = &\sum_{\delta_i=1} \log(f_{T,\theta_T}(y_i)\left[1 - h_{C|T,\theta}\{F_{C,\theta_C}(y_i)|F_{T,\theta_T}(y_i)\}\right]) \\ &+ \sum_{\delta_i=0} \log(f_{C,\theta_C}(y_i)\left[1 - h_{T|C,\theta}\{F_{T,\theta_T}(y_i)|F_{C,\theta_C}(y_i)\}\right]),\end{aligned}$$

where $h_{C|T,\theta}\{F_{C,\theta_C}(y_i)|F_{T,\theta_T}(y_i)\}$ and $h_{T|C,\theta}\{F_{T,\theta_T}(y_i)|F_{C,\theta_C}(y_i)\}$ express the conditional distribution function in terms of their associated copula.

Model-based boosting is used for the estimation of the proposed copula model, which allows modeling all distribution parameters simultaneously while being able to process high-dimensional data problems with $p > n$ (Hans et al., 2023). Furthermore, it leads to a data-driven variable selection which is controlled by the number of boosting iterations (Mayr et al., 2012).

3 Simulation

To study the performance of the proposed approach, we conducted a detailed simulation study. We considered not only low- and high-dimensional settings but also different censoring rates (20%, 50% and 80%) and an independent setting modeled via copula regression. The results of 100 simulation runs for the low-dimensional setting indicate that the model was able to identify the informative variables and accurately estimate the true effects for all distribution parameters, while also demonstrating stability across varying censoring rates. The high-dimensional setting performed similarly but it was more challenging to estimate the association parameter. Also the independent setting was estimated quite well for the informative variables regarding the marginal distributions, however, the model included many noise variables. We also compared our model with a classical Cox model and accelerated failure time models (AFT) models. We evaluated the estimated coefficients as well as the prediction performance using various metrics, i.e., C-index, Brier score, integrated Brier score, integrated absolute error and integrated squared error. Overall, the Cox model always performed considerably worse concerning the predictive performance than the AFT and copula models. Note, that the estimated coefficients of the Cox model are not directly comparable to the copula approach, because the coefficients are interpreted as hazard ratios.

For the AFT, the estimated coefficients of the informative variables were slightly overestimated and the model included some noise variables, particularly those that were informative for the censoring time. In terms of predictive performance, the copula model always performed best for the Brier score and integrated Brier Score; for the other metrics, the AFT model was similar to the copula model

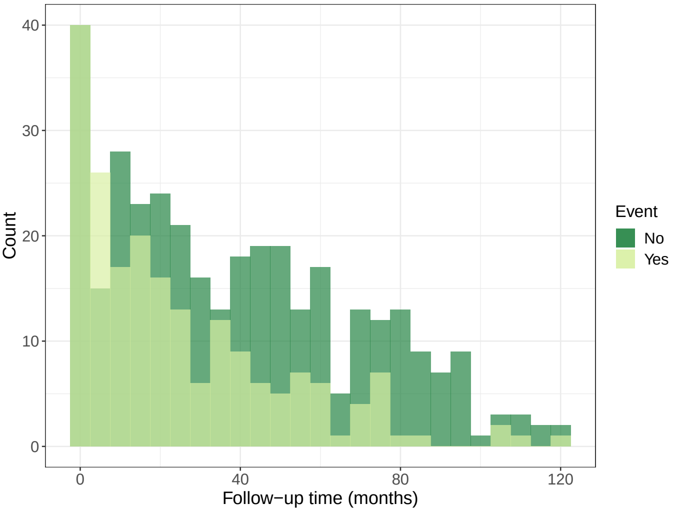


FIGURE 2. Histogram of the follow-up time in months for patients with event (light green) and without event (dark green).

from time to time. No major differences were found for the C-index between all three models.

4 Modelling the survival of colon cancer patients

We illustrate the new boosting approach for dependent censoring on a dataset on colon cancer. The dataset contains $n = 546$ patients, listed in a registry of a local German acute care hospital where all patients underwent the surgical resection of the affected part of the intestine with radical regional excision of adjacent lymph node stations, following the corresponding guidelines. We focus on the outcome of overall survival since surgery, where the event was observed in 201 patients, while 345 patients were right-censored, with an illustration of the follow-up time in Figure 2. The following covariates were included in the model: chemotherapy (yes/no), ASA score (general health status, mild/severe), UICC cancer stage (I-IV), age, LNE (number of pathologically examined lymph nodes), LNR (lymph node ratio, cancerous lymph nodes / examined lymph nodes), sex, R status (residual tumor, yes/no) and preexisting cancer (yes/no). The outcome was modeled by Weibull distributed margins and the Clayton copula chosen by the comparison of the empirical risk. All variables were included with simple linear models as base-learners.

Table 1 shows the selected and estimated coefficients for each parameter of the boosted copula model. Negative estimated coefficients for the mean survival time, for example, increase the chance of experiencing the event. Positive coefficients would indicate a longer survival or more precisely, chemotherapy shows a coefficient of 0.54 for survival time, which means that patients who have undergone chemotherapy are more likely to have a longer survival time. Chemotherapy was also selected for the correlation parameter with a coefficient of -1.170, which

TABLE 1. Resulting estimated coefficients for the dataset on colon cancer.

	μ_T	σ_T	μ_C	σ_C	ρ
Intercept	6.070	-0.270	3.061	-0.428	0.546
Age	-0.016	-0.001	0.007	0.002	-
Sex (female/male)	-0.033	0.071	0.006	-0.081	-
Chemotherapy (yes/no)	0.540	0.592	0.117	0.186	-1.170
ASA score (mild/severe)	-0.693	-0.195	-	-0.143	-
UICC cancer stage II	-0.092	-	-	0.029	-
UICC cancer stage III	-0.393	0.052	0.111	0.231	-
UICC cancer stage IV	-1.081	-0.147	-	-	-0.250
LNE	0.009	0.008	0.007	0.011	-
LNR	-1.616	0.434	0.560	-	-
R status	-0.153	-0.014	0.220	0.519	-
Preexisting cancer	-0.398	-0.020	-0.060	-0.014	-

indicates a negative impact on the association between survival and censoring time.

5 Conclusion

We have introduced a copula-based boosting approach to deal with dependent censoring. The simulation study indicates a promising performance of our approach, achieving stable results even with low and high censoring rates, also in comparison to classical approaches. Nevertheless, the challenge persists in evaluating the model adequately to facilitate comparison with traditional methods like the Cox model, as conventional evaluation metrics are no longer valid in the presence of dependent censoring. We have tried to solve this problem by considering several metrics that serve different purposes, but further research is needed in this direction.

A related approach was recently proposed by Midtfjord et al. (2022), which employs the Clayton copula to account for dependent censoring and builds on the accelerated failure time model. However, this approach is based on a fixed dependency parameter which does not depend on the covariates. We allow our algorithm to select also variables for the dependency parameter, like the chemotherapy in our colon cancer application.

Overall, the favorable performance of our new approach motivates further research in this direction, for example, the extension to non-linear effects, left censoring and the use of semi-parametric margins.

Acknowledgments: The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number 428239776).

References

Czado, C. and Van Keilegom, I. (2023). Dependent censoring based on parametric copulas. *Biometrika*, **110**, 721 – 738.

Hans, N., Klein, N., Faschingbauer, F., Schneider, M. and Mayr, A. (2023). Boosting distributional copula regression. *Biometrics*, **79**, 2298 – 2310.

Huang, X. and Zhang, N. (2008). Regression survival analysis with an assumed copula for dependent censoring: A sensitivity analysis approach. *Biometrics*, **64**, 1090 – 1099.

Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012). Generalized additive models for location, scale and shape for high-dimensional data – a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C*, **61**, 403 – 427.

Midtfjord, A.D., Bin, R.D. and Huseby, A.B. (2022). A copula-based boosting model for time-to-event prediction with dependent censoring. *arXiv:2210.04869 [stat.ME].*

Bayesian hidden Markov models for early warning

Daniele Tancini[1], Francesco Bartolucci[1], Silvia Pandolfi[1]

[1] [1]Department of Economics, University of Perugia, Italy

E-mail for correspondence: `francesco.bartolucci@unipg.it`

Abstract: We show how Bayesian hidden Markov models may be employed to build early warning systems of particular risky events. The adopted model formulation assumes that every binary response variable depends only on the latent state further to the lagged covariates and response. A Markov chain Monte Carlo algorithm is proposed for estimation and forecasting, where the latter is based on the optimisation of the F-score. An application referred to banking crisis of countries based on an unbalanced panel dataset is used as an illustration.

Keywords: Discrete latent variable models; F-score; forecasting; Markov chain Monte Carlo.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

Mitigating spatial confounding: a unified approach integrating simplified spatial+ and restricted regression

Arantxa Urdangarin[1,2], Tomás Goicoa[1,2,3], María Dolores Ugarte[1,2,3]

[1] Department of Statistics, Computer Science, and Mathematics, Public University of Navarre, Pamplona, Spain.
[2] INAMAT[2] (Institute for Advanced Materials and Mathematics), Public University of Navarre, Pamplona, Spain.
[3] Centro Asociado de la UNED, Pamplona, Spain.

E-mail for correspondence: arantxa.urdangarin@unavarra.es

Abstract: Incorporating covariates into spatial models for areal count data is a common approach for studying linear associations between covariates and the dependent variable. However, if a covariate is spatially structured, the spatial random effects might compete to explain the spatial variability in the response, possibly resulting in biased fixed effect estimates. This issue is referred to as spatial confounding. While various methods have been developed to address this problem, they have achieved only partial success, managing to reduce bias but often falling short in providing appropriate coverage rates. In this study, we propose a combination of a simplified spatial+ method with restricted regression to mitigate bias and achieve coverage rates closer to the nominal value. To illustrate the effectiveness of our method, we analyse the association between dowry deaths, a form of crime against women, and several socio-demographic covariates in Uttar Pradesh, India. Additionally, we conduct a simulation study that shows the improvement in the coverage of credible intervals for fixed effects with the proposed combined approach.

Keywords: Crimes against women; Restricted regression; Spatial confounding; Spatial+.

1 Introduction

Spatial models for areal count data are essential to smooth standardized incidence or mortality ratios of a disease and represent their geographical distribution. Incorporating covariates into spatial models to evaluate potential risk factors is

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

usual in ecological regression. However, spatial confounding might appear. This refers to the difficulty in disentangling the effects of covariates and spatial random effects, ultimately resulting in biased fixed effect estimates.

Various procedures have been proposed in the literature to mitigate spatial confounding. An interesting approach is the simplified spatial+ method (Urdangarin et al., 2024), a modification of spatial+ model (Dupont et al., 2022), that reduces bias of the fixed effects estimates, though coverage rates are over the nominal value.

In this work we propose combining the simplified spatial+ method and restricted regression (RSR) (Reich et al., 2006) to obtain unbiased fixed effects estimates while achieving appropriate coverage rates. For illustration purposes, we assess the association between the socio-demographic indicator sex ratio and dowry deaths in Uttar Pradesh, India, in 2011 (Vicente et al., 2020). Finally, we conduct a simulation study to evaluate coverage rates of the credible intervals for fixed effects with the proposed method. Model fitting and inference is carried out from a full Bayes approach, using integrated nested Laplace approximations (Rue et al., 2009).

2 A unified approach combining simplified spatial+ and restricted regression methods

The simplified spatial+ method (Urdangarin et al., 2024) is a modified version of the spatial+ approach (Dupont et al., 2022) that reduces bias of the fixed effects, avoiding fitting a spatial model to the covariate. The first step of the simplified spatial+ approach removes the spatial dependence from the covariate. Then, in a second step, a spatial model is fitted replacing the covariate by its decorrelated version.

Let Y_i and e_i denote the number of observed and expected cases, respectively, in the ith small area ($i = 1, \ldots, S$). Conditional on the relative risk r_i, Y_i is assumed to follow a Poisson distribution

$$Y_i|r_i \sim Poisson(\mu_i = e_i r_i) \quad \text{and} \quad \log \mu_i = \log e_i + \log r_i,$$

where the log risk is modeled as

$$\log \boldsymbol{r} = \mathbf{1}_S \alpha + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi}. \tag{Spatial}$$

Here, α is an intercept, $\boldsymbol{r} = (r_1, \ldots, r_S)^{'}$ is the vector of relative risks, $\boldsymbol{X} = (x_1, \ldots, x_S)^{'}$ is the covariate of interest, $\boldsymbol{\beta}$ is the fixed effect coefficient and $\boldsymbol{\xi}$ is the vector of spatial random effects. We express $\boldsymbol{X}$ as a linear combination of the eigenvectors $\boldsymbol{U}_i$ ($i = 1, \ldots, S$) of the spatial precision matrix as follows

$$\boldsymbol{X} = a_1 \boldsymbol{U}_1 + \cdots + a_S \boldsymbol{U}_S.$$

Assuming unconfoundedness at high frequencies, we split the covariate into two parts $\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{Z}^*$ where $\boldsymbol{Z}^*$ comprises large-scale eigenvectors associated with the lowest eigenvalues, responsible for the collinearity between fixed and random effects, and $\boldsymbol{Z}$ contains the rest of eigenvectors.

Then, the simplified spatial+ model consists in modeling the log risks using the spatial model, but replacing the covariate $\boldsymbol{X}$ by its spatially decorrelated part $\boldsymbol{Z}$, i.e.

$$\log \boldsymbol{r} = \mathbf{1}_S \alpha + \boldsymbol{Z}\beta + \boldsymbol{\xi}. \tag{SpatPlus}$$

Restricting the spatial random effects to the space orthogonal to the fixed effects $\boldsymbol{Z}$ in the previous model, we obtain the combination of the simplified spatial+ method and restricted regression. Namely

$$\log \boldsymbol{r} = \mathbf{1}_S\alpha + \boldsymbol{Z}\beta + \hat{\boldsymbol{W}}^{-1/2}\boldsymbol{L}\boldsymbol{L}^{'}\hat{\boldsymbol{W}}^{1/2}\boldsymbol{\xi}, \qquad \text{(SpatPlus-RSR)}$$

where the columns of $\boldsymbol{L}$ are the eigenvectors having non-null eigenvalues of the projection matrix $\boldsymbol{I}_S - \hat{\boldsymbol{W}}^{1/2}\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{Z}}^{'}\hat{\boldsymbol{W}}\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}^{'}\hat{\boldsymbol{W}}^{1/2}$, $\boldsymbol{W}$ being a diagonal matrix of weights with $W_{ii} = \mu_i$, and $\tilde{\boldsymbol{Z}} = [\mathbf{1}_S, \boldsymbol{Z}]$.

3 Real data analysis

In this section we study the association between dowry deaths in Uttar Pradesh in 2011 and the covariate sex ratio, defined as the number of females per 1000 males. We fit the SpatPlus and SpatPlus-RSR removing six large scale eigenvectors (SpatPlus6 and SpatPlus6-RSR), and the classical spatial model using an intrinsic CAR prior for the spatial random effects. Table 1 shows that SpatPlus6 and Spatplus6-RSR estimate very similar fixed effects, however the restricted regression yields smaller standard error for the regression coefficient.

TABLE 1. Posterior means, standard errors and 95% credible intervals of β for dowry deaths data in Uttar Pradesh in 2011.

Model	Mean	SD	95% CI	
Spatial	-0.1916	0.0592	-0.3062	-0.0734
SpatPlus6	-0.0922	0.0404	-0.1711	-0.0122
SpatPlus6-RSR	-0.1153	0.0224	-0.1592	-0.0714

4 Simulation study

We simulate a spatial confounding scenario using the geographical setup of Uttar Pradesh. We consider the observed values of sex ratio ($\boldsymbol{X}_1$) and an additional simulated variable ($\boldsymbol{X}_2$) that plays the role of the unobserved covariate. Here, the sex ratio and the unobserved covariate have a moderate-to-strong correlation, $cor(\boldsymbol{X}_1, \boldsymbol{X}_2) = 0.7$. The data are generated as

$$\log \boldsymbol{r} = \mathbf{1}_S\alpha + \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2$$
$$\boldsymbol{Y}^l|\boldsymbol{r} \sim Poisson(\boldsymbol{\mu} = \boldsymbol{er})$$

where $l = 1, \ldots, 300$, $\boldsymbol{e}$ is the vector of expected cases taken from the real case study, $\alpha = -0.03$, $\beta_1 = -0.2$ and $\beta_2 = -0.3$. We consider 6 (SpatPlus6) and 11 (SpatPlus11) large-scale eigenvectors in $\boldsymbol{Z}^*$ and the models fitted to the data only include the covariate sex ratio. In contrast to the simplified spatial+ without restricted regression, the integrated procedure provides coverages rates close to the nominal value 95%. Both methods recover well the true fixed effects.

TABLE 2. Posterior means, standard deviations, 95% credible intervals and empirical 95% coverage probabilities of the true value of β_1 based on 300 simulated datasets.

Model	Mean	SD	95% CI		95% coverage
Spatial	-0.3846	0.0591	-0.5002	-0.2671	2.6667
SpatPlus6	-0.2288	0.0424	-0.3119	-0.1451	97.0000
SpatPlus11	-0.2133	0.0403	-0.2922	-0.1338	98.6667
SpatPlus6-RSR	-0.2104	0.0243	-0.2581	-0.1627	95.6667
SpatPlus11-RSR	-0.1905	0.0230	-0.2356	-0.1454	96.3333

Acknowledgments: This work has been supported by Projects PID2020-113125RB-I00/MCIN/AEI/10.13039/501100011033 and UNEDPAM/PI/PR24/05A.

References

Dupont, E., Wood, S.N. and Augustin, N.H. (2022). Spatial+: A novel approach to spatial confounding. *Biometrics.* **78**, 1279 – 1290.

Reich, B.J., Hodges, J.S. and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics.* **62**, 1197 – 1206.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, **71**, 319 – 92.

Urdangarin, A., Goicoa, T., Kneib, T. and Ugarte, M.D. (2024). A simplified spatial+ approach to mitigate spatial confounding in multivariate spatial areal models. *Spatial Statistics.* **59**, 100804.

Vicente, G., Goicoa, T., Fernandez-Rasines, P. and Ugarte, M.D. (2020). Crime against women in India: unveiling spatial patterns and temporal trends of dowry deaths in the districts of Uttar Pradesh. *Journal of the Royal Statistical Society: Series A.* **183**, 655 – 679.

An updated Wilcoxon–Mann–Whitney test

Paul Wilson[1]

[1] University of Wolverhampton, UK

E-mail for correspondence: `pauljwilson@wlv.ac.uk`

Abstract: The Wilcoxon–Mann–Whitney test, also known as the Wilcoxon rank−sum test and the Mann-Whitney U test, is a non−parametric method used to compare differences between two independent variables when the dependent variable is either ordinal or continuous. The associated test statistic follows a discrete distribution, the quantiles and corresponding p-values of which are traditionally determined using "traditional" definitions. We show that the utilisation of the mid−quantiles and mid p-values leads to a test with improved attainment rates and power.

Keywords: Mann-Whitney-Wilcoxon test; Mid p-values; Discrete test statistics.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Elicitation of priors for intervention effects in educational trial data

Qing Zhang[1], Germaine Uwimpuhwe[1], Dimitris Vallis[1,2], Akansha Singh[1], Tahani Coolen-Maturi[1], Jochen Einbeck[1]

[1] Durham University, Durham, UK
[2] The Policy Institute, King's College London, UK

E-mail for correspondence: `qing.zhang@durham.ac.uk`

Abstract: Effect sizes for educational interventions are commonly small, and hence decisions to re-grant efficacy trials (small trials with homogeneous populations under idealized conditions) as effectiveness trials (larger trials with heterogeneous populations) are often based on limited evidence from the efficacy trial itself. However, supplementary evidence may be available on how (past) effectiveness trials with similar outcomes tend to perform. This work proposes a Bayesian approach of making use of such evidence for re-granting decisions.

Keywords: Randomized controlled trial; Multilevel model; Meta-analysis.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

R package mult.latent.reg for multivariate response scenarios with latent structures

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: We developed an R package, now available on CRAN, that implements the fitting methodologies for the 1-level and 2-level variants of a recently introduced model for clustered and highly correlated multivariate data. The computations are based on an EM algorithm in the spirit of the Nonparametric Maximum Likelihood (NPML) approach for the estimation of mixture models. The implementation also features alternative choices of the starting values for the EM algorithm, which we discuss in this abstract and which have not been described elsewhere.

Keywords: Random effects; Multilevel model; EM algorithm; Starting values; Clustering.

1 Introduction

In recent work, Zhang and Einbeck (2024) proposed a modelling framework for multivariate data, which provides implicit dimension reduction by representing 1-dimensional latent linear structures (as occurring for highly correlated data) through a univariate random effect. The basic model takes the form $x_i = \alpha + \beta z_i + \varepsilon_i$, where $x_i \in \mathbb{R}^m$ and $\epsilon_i \sim N(0, \Sigma)$. Regression problems with multivariate response can be dealt with through the inclusion of a covariate term Γv_i, allowing for correlations between response variables to be taken into account. An extension enabling the handling of multivariate response scenarios with 2-level structure was provided in Zhang et al. (2023), allowing for covariates on both levels, $x_{ij} = \alpha + \beta z_i + \Gamma v_{ij} + \varepsilon_{ij}$, with lower-level units indexed by $j = 1, \ldots, n_i$, and upper-level units $i = 1, \ldots, r$. We use an NPML–type approach (Aitkin, 1996) for parameter estimation, implemented through the EM algorithm. This involves the fitting of a discrete mixture along the latent space which can be useful for clustering purposes.

2 R package

We developed an R package titled **mult.latent.reg** that implements the EM algorithm for the presented models. We use AIC and BIC for model selection. The main functions for the 1-level model are `mult.em_1level` and `mult.reg_1level`. The first function will run EM once with (by default) 20 iterations, producing output including parameter estimates, log-likelihood, disparity, AIC, BIC values and starting points; the second function can execute the EM multiple times (by default 10 runs) and outputs the result with the smallest AIC value (also giving the starting points that generate that result). We support four types of parameterizations for Σ: the same diagonal variance matrix for all mixture components, different diagonal variance matrices for different mixture components, the same full variance matrix for all components, and different full variance matrices for different components.

The main functions for the 2-level model are `mult.em_2level` and `mult.reg_2level`; the outputs are the same as the ones obtained from the functions for the 1-level model except we only AIC for model selection. The 2-level model offers only two choices for variance parameterization due to practical reasons: using the same diagonal variance matrix for all components of the mixture or using different diagonal variance matrices for different components.

Here, we present an example of the 2-level model function applied to bivariate trading data (OECD, 2023) on imports and exports collected between 2018 and 2022 across 44 countries, each with 3 to 5 annual observations available (no covariates). We use `option = 1` for the starting value (to be explained in Section 3) and adopt the second variance parameterization.

```
> set.seed(49)
> trade_res <- mult.em_2level(trading_data, K=4, steps=10, var_fun=2,
  option = 1)
```

then we obtain the estimates (only showing an excerpt), where `p` and `z` are estimated mixture parameters, `beta` corresponds to the β parameter from Section 1, and `W` is the matrix of responsibilities.

```
> trade_res$p
[1] 0.23645108 0.43082442 0.02272643 0.30999807
> trade_res$z
[1] -1.4015770  0.8455288  2.9333327 -0.3210802
> trade_res$beta
0.5073759         0.5381193
> trade_res$W
              [,1]         [,2]          [,3]         [,4]
 [1,] 9.786883e-01 5.870324e-04  0.000000e+00 2.072465e-02
 [2,] 1.355809e-06 9.525065e-01  0.000000e+00 4.749210e-02
 [3,] 1.251692e-12 9.999948e-01 8.906935e-141 5.227975e-06
 [4,] 9.999611e-01 4.111404e-06  0.000000e+00 3.478559e-05
 ...
```

3 Starting values

Using appropriate starting values for the parameters is beneficial for the EM to find the maximum likelihood parameter estimates. In R package **mult.latent.reg**, we provide four options of data-dependent starting values for the EM initialization, with the first based on Zhang and Einbeck (2024) and the other ones novel:

(i) `option=1`: For the mixture weights, we use $\pi_k^{(0)} = \frac{1}{K}$, where K is the number of components. We draw random numbers from a standard normal distribution as the starting values for the mass points $z_k^{(0)}$. We use column means for the line parameters $\alpha^{(0)}$, and $\beta^{(0)} = x_r - \alpha^{(0)}$, where $x_r \in \mathbb{R}^m$ is a randomly selected observation. For parameter Γ, we first fit separate linear models, each using one of the columns of x_i as response variable and v_i as predictor variables, then we use the coefficient estimates as the starting values, $\Gamma^{(0)}$. For all four variance parameterizations, we use a diagonal matrix $\Sigma^{(0)} \in \mathbb{R}^{m \times m}$, not depending on k, as the 'starting variance matrix': Denote s_j for $j = 1, 2, ..., m$ the sample standard deviation of the j-th variable. Then, for each diagonal element $(\sigma_j^{(0)})^2$ of $\Sigma^{(0)}$, one has the starting value $\sigma_j^{(0)} = \frac{s_j}{K}$, $j = 1, \ldots, m$.

(ii) `option=2`: We use a short run (5 iterations) of the EM process which uses option (i) with `var_fun=1` as the starting values, and then use the estimates as the starting values for a relatively larger number of iterations. This approach is motivated by Biernacki et al. (2003), where a short run of the EM is applied before running CEM runs.

(iii) `option=3`: The parameter β in our model plays a similar role to the rotation matrix in principal component analysis, specifically aligned with the first principal component. This observation motivated our choice of using the first principal component of the rotation matrix as the initial values for β, while keeping the starting values for the remaining parameters consistent with those described in (i).

(iv) `option=4`: In the application of clustering, a small number of observations in a dataset intended to form a distinct group may occasionally be assigned to a neighboring cluster. This inspired the idea that it would be better to use a more precise starting value for the mass points z_k. What we do is that first, take the scores of the first principal component of the data and perform K-means on these. Then the starting values for the parameter π_k are the proportions of the clustering assignments, and the starting values for z_k are the values of the K-means centers. The starting values for the rest of the parameters are the same as described in (i).

The performance of these options is illustrated in Figure 1 for two data sets, namely the trading data introduced in Section 2, and a 1-level data set where five fetal movement types serve as multivariate outcomes, and a variable indicating pre/post-Covid status as covariate (more detail on the modelling in Zhang and Einbeck, 2024). The left plot shows that, in terms of AIC values obtained from 50 applications of each starting value option, `option=3` tends to perform better than the other three options, with `option=2` showing the worst performance. Meanwhile, for the fetal data, `option=2` tends to perform best, emphasizing that different starting point choices may be successful for different data sets.

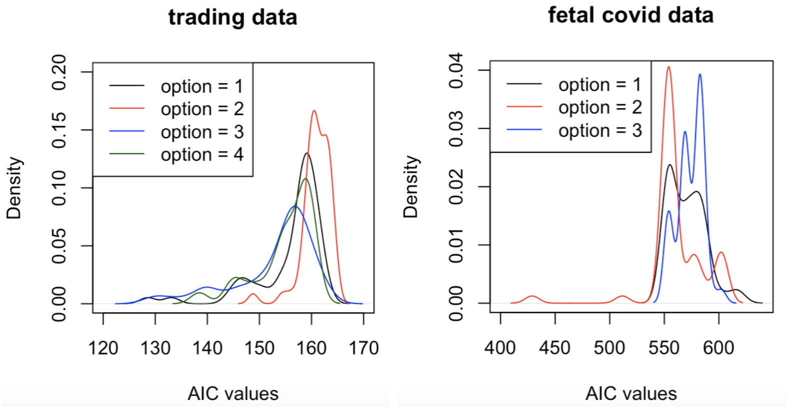


FIGURE 1. Distributions of AIC values from 50 runs for each starting value option, for the trading data (left) and the fetal data (right).

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251 – 262.

Biernacki, C., Celeux, G. and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, **41**, 561 – 575.

OECD (2023). Organisation for Economic Co-operation and Development. Trade in Goods and Services. https://data.oecd.org/trade/ trade-in-goods-and-services.htm. Accessed on 2023-05-29.

Zhang, Y., Einbeck, J. and Drikvandi, R. (2023). A multilevel multivariate response model for data with latent structures. In: *Proceedings of the 37th International Workshop on Statistical Modelling*, 343 – 348.

Zhang, Y. and Einbeck, J. (2024). A versatile model for clustered and highly correlated multivariate data. *Journal of Statistical Theory and Practice*, **18**, 5.

The underlap coefficient as a measure of a biomarker's discriminatory ability and covariate dependence in cluster analysis

Zhaoxi Zhang[1], Vanda Inácio[1], Miguel de Carvalho[1], Sara Wade[1]

[1] School of Mathematics, University of Edinburgh, UK

E-mail for correspondence: `Z.Zhang-156@sms.ed.ac.uk`

Abstract: We introduce the underlap coefficient as a novel metric for the statistical evaluation of medical tests across multiple disease stages, and for measuring covariate dependence in cluster analysis. Initially designed to quantify separation between distributions, we demonstrate its application through a covariate-specific variant estimated via a Bayesian approach. Utilizing a Bayesian nonparametric covariate-dependent mixture model with a logit stick-breaking prior, we conduct a case study on Alzheimer's disease (AD), aimed at evaluating the accuracy of potential biomarkers in distinguishing between normal cognition, mild cognitive impairment, and dementia, and examining how this accuracy is influenced by covariates. This metric is also useful in the context of precision medicine and targeted interventions, where it's crucial to assess the dependency of partitions obtained by mixture models on covariates.

Keywords: Underlap coefficient; Mixture models; Degree of separation; Covariate dependence.

1 The underlap coefficient

Let Y_d, where $d \in \{1, 2, \ldots, K\}$, denote independent continuous random variables representing the biomarker values across K distinct disease classes. The underlap coefficient (UNL) (Zhang et al 2024) was proposed to measure a biomarker's discriminative ability, and for K classes is defined as:

$$\text{UNL}(f_1, \ldots, f_K) = \int \max(f_1(y), \ldots, f_K(y))dy, \qquad (1)$$

where $f_d(y)$ is density of biomarker value y in group d.

The underlap coefficient can be intuitively interpreted as the "effective" number of populations of biomarkers for all groups, drawing an analogy to the effective

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

sample size in Markov chain Monte Carlo (MCMC). Its values range from 1 to K, with higher values indicating higher degree of separation. A value of K indicates complete separation of biomarker values across groups, signifying the presence of K distinct "effective" populations without any overlap. Conversely, a value of one for the UNL suggests that only one "effective" population exists, as all K populations are identical.

Covariates can impact the discriminatory power of a biomarker and ignoring covariate information may lead to erroneous inferences about a test's accuracy, and therefore a covariate-dependent structure should be included when modelling the underlap coefficient. Let $\mathbf{X}_d$, where $d \in \{1, 2, \ldots, K\}$, denote independent covariate vectors in group d. For a given covariate vector value $\mathbf{x}$, the covariate-specific underlap coefficient is defined as:

$$\mathrm{UNL}(f_1, \ldots, f_K|\mathbf{x}) = \int \max(f_1(y|\mathbf{x}), \ldots, f_K(y|\mathbf{x}))dy, \qquad (2)$$

where $f_d(y|\mathbf{x})$ denotes the conditional density of Y_d, for $d \in \{1, 2, \ldots, K\}$.

2 Bayesian estimator for the underlap coefficient

Within the Bayesian nonparametric framework, we consider the general class of covariate-dependent infinite mixture of normals model

$$f_d(y \mid \mathbf{x}) = \sum_{l=1}^{\infty} \omega_l(\mathbf{x})\phi(y|\theta_l(\mathbf{x})), \quad d \in \{1, 2, \ldots, K\}, \qquad (3)$$

where the mixing weights follow a a stick-breaking construction, i.e., $\omega_1(\mathbf{x}) = v_1(\mathbf{x})$, $\omega_l(\mathbf{x}) = v_l(\mathbf{x})\prod_{m=1}^{l-1}(1 - v_m(\mathbf{x}))$ for $l \geq 2$. Popular particular cases, mainly due to computational simplicity, of the model specification in (3) include the single-weights model ($\omega_l(\mathbf{x}) = \omega_l$) and the single-atoms model ($\theta_l(\mathbf{x}) = \theta_l$). However, the covariate-independent assumption for the mixing weights or the atoms might have limited flexibility in practice. With this in mind, we follow the logit stick-breaking prior formulation, recently proposed by Rigon and Durante (2021), which retains the computational simplicity but affords the necessary flexibility needed in many applications. Specifically, let $\theta_l(\mathbf{x}) = (\mu_l(\mathbf{x}), \sigma_l^2)$, where $\mu_l(\mathbf{x})$ is modelled as a linear combination of selected functions of the covariates $\lambda(\mathbf{x}) = \{\lambda_1(\mathbf{x}), \ldots, \lambda_M(\mathbf{x})\}^T$, thus leading to

$$\mu_l(\mathbf{x}) = \lambda(\mathbf{x})^T\boldsymbol{\beta}_l$$

A logit stick-breaking prior for the weights is employed, which is represented by a sequence of logistic regressions:

$$\eta_l(\mathbf{x}) = logit(v_l(\mathbf{x})) = \psi(\mathbf{x})^T\boldsymbol{\alpha}_l$$

where $\psi(\mathbf{x}) = \{\psi_1(\mathbf{x}), \ldots, \psi_R(\mathbf{x})\}^T$ are selected functions of the observed covariates. Note that $\eta_l(\mathbf{x})$ is interpreted as the log-odds of being allocated to component l in the continuation-ratio parameterization (Tutz 1995), conditionally on the event of surviving to the first $(1, \ldots, l-1)$ components. In practice, the infinite mixture in (3) is truncated to a finite number of components, say L, which shall be regarded as an upper bound on the number of occupied components. To complete the model specification, we should set prior distributions for the model parameters. For conjugacy reasons, we let

$$\boldsymbol{\alpha}_l \sim N_R(\mu_\alpha, \Sigma_\alpha), \quad \boldsymbol{\beta}_l \sim N_M(\mu_\beta, \Sigma_\beta), \quad \sigma_l^2 \sim IG(a_{\sigma^2}, b_{\sigma^2}),$$

where $IG(a, b)$ represents an inverse-gamma distribution with shape parameter a and rate parameter b. It is worth mentioning that Pólya-gamma data augmentation scheme (Polson et al., 2013) should be adapted to solve the difficulty of Bayesian inference in logistic regression, in order to get the full posterior conditional distributions of each $\boldsymbol{\alpha}_l$ in Gibbs sampling. For a detailed model specification and justification, please see Rigon and Durante (2021). Based on the estimated conditional densities, the integral in (1) and (2) can be approximated numerically by the Simpson's rule.

3 Examining the the discrminatory ability of 4 potential AD biomarkers by UNL

The dataset derived from the Alzheimer's Disease Neuroimaging Initiative consists of 1032 subjects, with 313 subjects in the cognitively normal group, 581 subjects in the mild cognitive impairment group, and 138 subjects in the Alzheimer's disease group. We aim to evaluate the age and gender effect on the accuracy's performance, as measured by the underlap coefficient, of CSF Abeta, CSF Tau, hippocampal volume, and hypometabolic convergence index (HCI) to distinguish (simultaneously) between the three groups.

In selecting the predictor functions for modeling the weights in our mixture model, we employed the 'absolute difference strategy' (Ariyo et al 2020). This approach entails choosing a more complex model only if it surpasses a simpler model by more than 5 units in the criterion value (WAIC/LPML value in our case). Accordingly, we implemented a B-splines formulation with no interior knots for the cognitively normal group in the HCI, the mild cognitive impairment group, and the Alzheimer's disease group in CSF Abeta, as well as for the cognitively normal group in hippocampal volume. For the other groups of the four biomarkers, raw predictors were used to model the mixing weights. Posterior inference was obtained using 5000 iterates after 5000 iterations were discarded as burn-in period.

In Figure 1 we show the estimated age and gender specific underlap coefficient. For the HCI, the underlap coefficient is consistently higher for females than for males of equivalent age, implying a marginally superior performance of HCI in females. The trend of UNL observed for CSF Abeta closely mirrors that of the HCI. However, it is noteworthy that the UNL values for CSF Abeta are consistently moderately lower than those for HCI across all age groups and in both genders. Moreover, the underlap coefficient for CSF Tau is higher for females than for males up to the age of 77, suggesting a more robust classifying capability for females within that age range. In terms of hippocampal volume, the rate of change in the underlap coefficient is more gradual compared to the other three biomarkers, with test performance slightly favoring males over females. Also, ignoring the age effect would lead to underestimate the performance of the biomarkers for younger individuals and to overestimate it for older individuals. The results indicate a moderate age & gender effect.

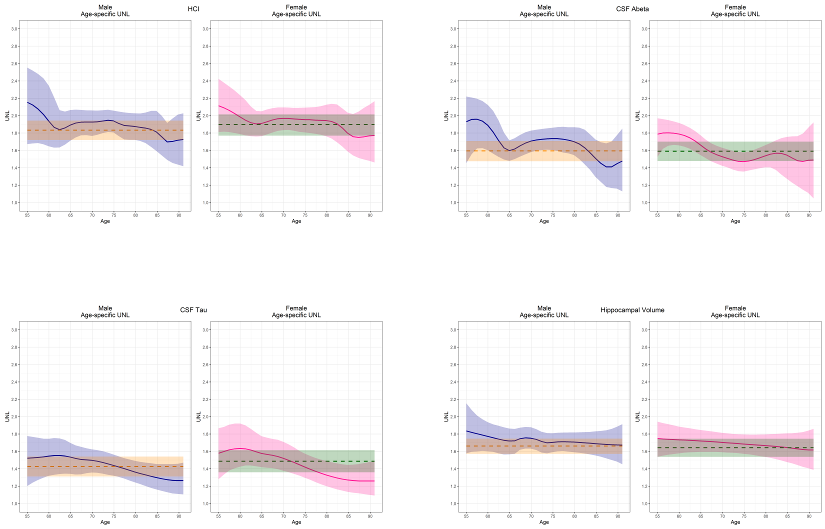


FIGURE 1. Blue/Pink line and ribbon: posterior mean and 95% credible intervals for the age and gender specific underlap coefficient for the four biomarkers (Male/Female). Orange/Green line and ribbon: posterior mean and 95% credible intervals for 3-class underlap coefficient for the four biomarkers ignoring age effect (Male/Female).

4 Measuring covariate dependence based on UNL

To measure the covariate dependence in cluster analysis, we would need to model the underlap of covariate rather than that of the response, which is

$$\mathrm{UNL}(f_1,\ldots,f_K)=\int \max(f_1(\mathbf{x}),\ldots,f_K(\mathbf{x}))d\mathbf{x}, \tag{4}$$

where $f_j(\mathbf{x})$ is density of covariate $\mathbf{x}$ in cluster j.

Given a partition with K clusters, if the clusters are equally presented across the covariate space, the UNL will be close to 1, whereas it will be k when all clusters correspond to distinct regions in the covariate space.

To illustrate how to use UNL for measuring covariate dependence, we consider one simulated data example with a quadratic regression function. For $i=1,\ldots,50$,

$$Y_i|\mathbf{X}_i \overset{\text{ind}}{\sim} N(\mathbf{X}_i^2,1); \quad \mathbf{X}_i \overset{\text{ind}}{\sim} U(-5,5)$$

we employed a linear dependent Dirichlet process Gaussian mixture (LDDP) model (formulated as $y \mid \mathbf{x}, P_{\mathbf{x}} \sim \sum_{j=1}^{J} w_j \mathrm{N}(y \mid \beta_j \mathbf{x}, \sigma_j^2))$ to model the data. Subsequently, we identified a representative partition of the posterior by minimizing the lower bound to the posterior expected Variation of Information from Jensen's Inequality (Wade and Ghahramani 2018).

Based on the representative partition, we modeled the covariate $\mathbf{x}$ within each cluster using an unconditional Dirichlet process Gaussian mixture (DPM) model to compute the underlap coefficient of all clusters together. As depicted in Figure

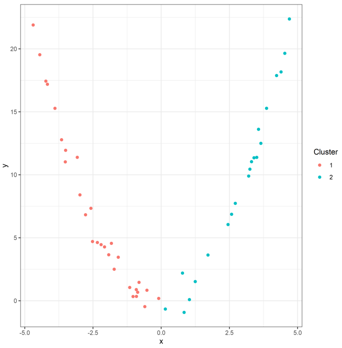


FIGURE 2. Representative partition of simulated data generated using the LDDP model

2, there are two clusters given by the LDDP model in this example, with the covariate regions of the two clusters are very distinct from each other. The posterior median of UNL is 1.90, and the 95% credible interval of UNL is $(1, 77, 1.96)$. Both the point estimate and the credible interval approximate 2 closely, suggesting that allowing the weights in LDDP model to depend on covariate $\mathbf{x}$ might enhance the predictive performance.

References

Rigon, T. and Durante, D. (2021). Tractable Bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference*, **211**, 131 – 142.

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, **13**, 559 – 626.

Zhang, Z., Inácio, V. and de Carvalho, M. (2024). The underlap coefficient as measure of a biomarker's discriminatory ability in a three-class disease setting. *Technical report.*

Polson, N. G., Scott J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American statistical Association*, **108**, 1339 – 1349.

Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics & Data Analysis* **11(3)**, 275 – 295.

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G. and Lesaffre, E. (2019). Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics* **47**, 890 – 913.

Spatial regression with misaligned covariates for soil moisture mapping

Weiyue Zheng[1], Andrew Elliott[1], Claire Miller[1], Marian Scott[1]

[1] School of Mathematics and Statistics, University of Glasgow, Glasgow, United Kingdom

E-mail for correspondence: weiyue.zheng@glasgow.ac.uk

Abstract: High-resolution soil moisture data have great value in many different application areas. Soil moisture can be measured in a variety of ways, including using in-situ sensors, which can provide accurate and stable long-term soil moisture values. However, typically the sensor data have limited spatial coverage and in some cases, soil moisture is indirectly measured through other related covariates. In this project, we develop a data fusion method using an SPDE (Stochastic Partial Differential Equation) approach to generate detailed soil moisture maps using measurements of Volumetric Water Content (VWC) and related covariates obtained from in-situ sensors. This work accommodates both misaligned and aligned covariates in a spatial perspective. The model is examined via simulation to explore how model performance scales with the number of sensors. The preliminary results are presented both in a detailed simulation and in a real data application from Elliot Water in Scotland, UK.

Keywords: Data fusion; Spatial misalignment; INLA-SPDE.

1 Introduction

Monitoring soil moisture can play an important role in helping to inform researchers, regulators and land owners about the available water content of the soil for agriculture and vegetation. However, the capacity to observe soil moisture is constrained by practical and financial limitations making it challenging to observe continuously across space and time. We can only monitor soil moisture at a finite number of spatial locations and time points. One of the most accurate methods for measuring soil moisture is using in-situ sensors. However, the high cost of deploying these sensors extensively means that soil moisture data tends to be collected from a sparse network of monitoring points. Given the limited in-situ sensor data, it becomes essential to explore the benefits of utilizing other data sources through developing and using data fusion techniques. Data fusion allows for the integration of different related determinands from different in-situ

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

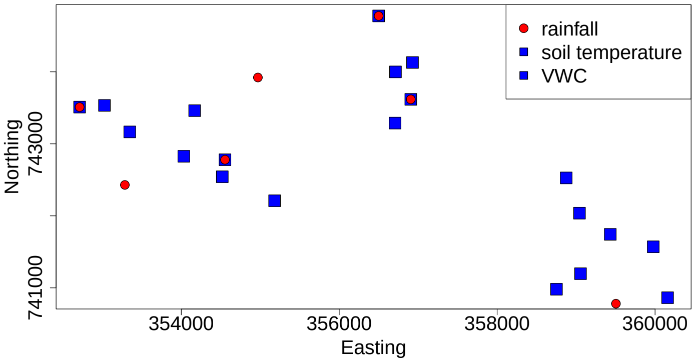


FIGURE 1. Locations for the response variable (VWC) and the aligned covariate (soil temperature) are represented by blue squares, while the misaligned covariate (rainfall) is represented by red circles.

sensors, enhancing the ability to make informed decisions and understand environmental phenomena with more precision, despite the limited direct monitoring of soil moisture.

Spatial misalignment, which here refers to the response variable and the covariates being observed at different spatial locations, is a common challenge in many environmental research studies. Scott (2023) mentioned that it is very challenging to deal with the data fusion of misaligned data. Spatial regression models, commonly employed to investigate the relationship between response variables and covariates while considering spatial correlation, often assume that these variables are observed at the same locations. However, this is not always true in the real world; with the development of new technology, it has become increasingly common for response variables and covariates to be collected from different locations and data sources, such as environmental sensors gathering information from different collection points. Figure 1 illustrates the spatial distribution of the direct measurements of soil moisture (Volumetric Water Content, VWC), along with two variables related to soil moisture: soil temperature and rainfall from sensors deployed by the Scottish Environment Protection Agency (SEPA) in the Elliot water catchment, which is located in the north-east of Scotland. In this context, rainfall is the misaligned covariate, whereas soil temperature is identified as the aligned covariate.

2 Methodology

This paper uses a spatial regression model to deal with misaligned covariates and generate a predictive soil moisture map for the Elliot water catchment.

The framework of the geostatistical model is described as follows:

$$\begin{aligned} y_1(\mathbf{s}^*) =&\ \alpha_1 + \mu_1(\mathbf{s}^*) + e_1(\mathbf{s}^*) \\ y_2(\mathbf{s}) =&\ \alpha_2 + \mu_2(\mathbf{s}) + e_2(\mathbf{s}) \\ y_3(\mathbf{s}) =&\ \alpha_3 + \beta_3 x(\mathbf{s}) + \beta_1(\alpha_1 + \mu_1(\mathbf{s})) + \beta_2(\alpha_2 + \mu_2(\mathbf{s})) + \mu_3(\mathbf{s}) + e_3(\mathbf{s}), \end{aligned} \tag{1}$$

where $y_k(\mathbf{s})$ denotes the realization of the spatial process $Y(\cdot)$ with $k = 1, 2, 3$ and $\mathbf{s}$ denoting the set of all locations. The notation $\mathbf{s}^*$ here denotes that the variable is collected at non-aligned locations with other variables, and $\mathbf{s}$ denotes that the variable is collected at the same locations with other variables. The α_k are the intercepts, $\mu_k(\mathbf{s})$ are random fields with Matérn covariance function, $x(\mathbf{s})$ introduces a covariate as a fixed effect. β_1 and β_2 are scaling parameters for the spatial effects, β_3 is the scaling parameter of the fixed effect and $e_k(\mathbf{s}) \sim N(0, \sigma^2_{ek})$ are uncorrelated error terms defined by a Gaussian white-noise process, and the error process is spatially uncorrelated. Further, the decision to model a covariate as either a fixed effect or a random effect depends on its availability at the predicted target locations.

3 Simulation study

In many applications, real data tends to be complex and hard to explain. Particularly in environmental applications, the monitoring network used for data collection can be sparse. So even if the variable of interest is continuous across space and through time, the scarcity of monitoring makes it challenging to model spatial correlations due to the small number of available locations. Therefore, a simulation study is employed to assess the effectiveness of the model with different numbers of sensors, where the simulated data are constructed to mimic the real data Elliot water case as far as possible.

For data simulation, the spatial process $\mu_k(\mathbf{s})$ is simulated by generating independent random field realizations from a Matérn Gaussian random field. The behaviour of the Matérn field is controlled through three parameters: range (ρ), marginal variance (σ), and smoothness (v). The trend covariate $x(\mathbf{s})$ is derived from a surface where values (from 0 to 3.5) exhibit an increasing pattern from the southwest to the northeast across the area.

For each scenario, 100 independent replications are performed to compute posterior quantities of interest which include the posterior mean, posterior median, and 95% credible intervals of the parameters within the model. As above, the viability of this approach in such a data constrained setting can be evaluated by fitting the model using a different number of locations. The number of locations for each variable can be found in Table 1, where n_1 is the number of misaligned covariate locations, n_2 is the number of aligned covariate locations, and n_3 is the number of response variable locations. The model is fitted using integrated nested Laplace approximation (INLA) and SPDE approaches (Lindgren and Rue, 2011). The details about the prior distribution and the parameter setting for the model fitting can be found in Lindgren and Rue (2015).

Table 1 presents the mean from the posterior distributions for a selection of the parameters within the model. It reveals consistent and accurate estimations for the fixed effect parameters regardless of the amount of available data. As the number of locations increases, the accuracy in estimating the scaling parameters β_1 and β_2 improves. However, for some remaining parameters, for example, the estimation of ρ_3 is not accurate no matter how many locations are available within our tested ranges. The simulation of model (1) indicates that 95% CIs (not shown) from the number of sensors in the real data (n_1 = 10, n_2 = 22, n_3 = 22) capture true fixed effects values only, whereas CIs from the large sample size accurately include most parameter values.

TABLE 1. Mean of the posterior parameter distribution as the number of locations varies, offering insights into how parameter estimates evolve across different scenarios.

n_1	n_2,n_3	α_1	α_2	α_3	β_1	β_2	β_3	ρ_1	ρ_2	ρ_3
10	22	0.46	0.76	0.76	-0.15	-0.09	-0.25	2.19	1.37	2.15
40	88	0.49	0.8	0.74	-0.22	-0.15	-0.19	1.68	1.68	3.51
200	440	0.46	0.79	0.94	-0.3	-0.35	-0.19	3.58	3.38	2.7
Actual Value		0.5	0.8	1	-0.3	-0.4	-0.2	4	3	2

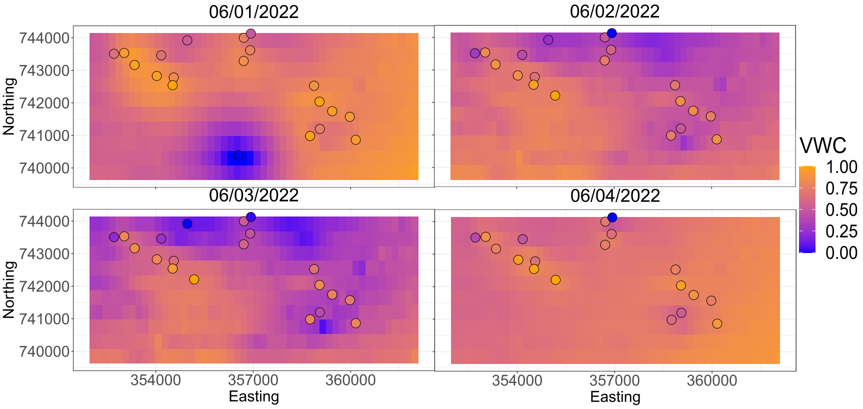


FIGURE 2. Prediction of VWC from sensors in the Elliot Water catchment on 06/01/2022 (top left), 06/02/2022 (top right), 06/03/2022 (bottom left) and 06/04/2022 (bottom right)

4 Real data application

Model (1) is implemented for the soil moisture dataset of the Elliot Water catchment, where y_1 represents rainfall, y_2 represents soil temperature, y_3 represents Volumetric Water Content (VWC) and x denotes high resolution elevation data. All the variables are normalized between 0 and 1 to prevent one variable from being overly influential when they are measured in different units. Figure 2 displays the predicted soil moisture map for the Elliot water catchment on 06/01/2022, 06/02/2022, 06/03/2022, and 06/04/2022 individually. Notably, the soil moisture map for 06/01/2022 shows more spatial variation, particularly in the middle-bottom region, which is directly linked to the availability of the number of in-situ sensors over time. The circles represent the actual VWC values measured by sensors. Due to the sparse monitoring network of the in-situ sensors, the predicted mean does not exhibit significant spatial variation. The elevation, which is available everywhere, accounts for the observed spatial patterns in the areas where there are no sensors.

5 Conclusion and future directions

In this paper we have proposed a spatial regression model to predict soil moisture from (mis)aligned covariates. The effectiveness of the model has been illustrated through simulation and real data indicating the information gained through the increase in in-situ sensor locations. The future work includes expanding the spatio-only model to a spatio-temporal model to improve the estimation of parameters through the utilization of temporal data as well as incorporating grid satellite data to enhance the spatial coverage of the soil moisture map (Moraga et al., 2017).

Acknowledgments: This project is supported by the CSC studentship and has developed from the UKRI-NERC digital environment project Feasibility Study - NE/T005564/1, NERC Constructing a Digital Environment. The authors gratefully acknowledge the Scottish Environment Protection Agency (SEPA) for providing the in-situ data.

References

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, **63**, 1 – 25.

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

Moraga, P., Cramb, S. M., Mengersen, K. L. and Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, **21**, 27 – 41.

Scott, E.M. (2023). Framing data science, analytics and statistics around the digital earth concept. *Environmetrics*, **34**, e2732.

A computationally efficient spatio-temporal fusion model for reflectance data

Zhaoyuan Zou[1], Ruth O'Donnell[1], Claire Miller[1], Duncan Lee[1], Craig Wilkie[1]

[1] University of Glasgow, UK

E-mail for correspondence: z.zou.2@research.gla.ac.uk

Abstract: Fusing remotely-sensed reflectance data from different sources at different spatial and temporal scales is useful to monitor lake water quality. The nonparametric statistical downscaling model (NSD) [Wilkie et al., 2019, *Environmetrics*] can account for a change of spatial and temporal support between two remote sensors, but it is computationally demanding for large datasets. This work proposes a method to improve the computational efficiency of the NSD model by endowing it with a Gaussian predictive process. The predictive performance and computational efficiency of both models are compared through simulation and using satellite reflectance data from Lake Garda.

Keywords: Nonparametric statistical downscaling; Gaussian predictive process; Reflectance; Lake water quality.

Full paper

This manuscript is available as part of the Springer volume Developments in Statistical Modelling using the direct link provided on the conference main page.

Thanks to:

Durham
University
Research Methods Centre