technische universität
dortmund

fakultät
statistik

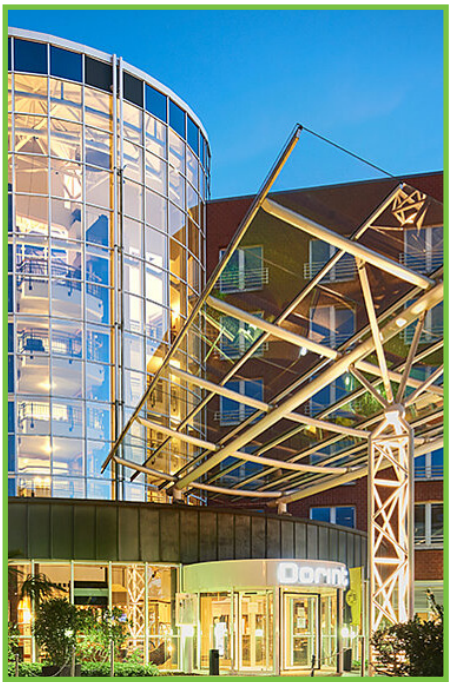
IWSM 2023

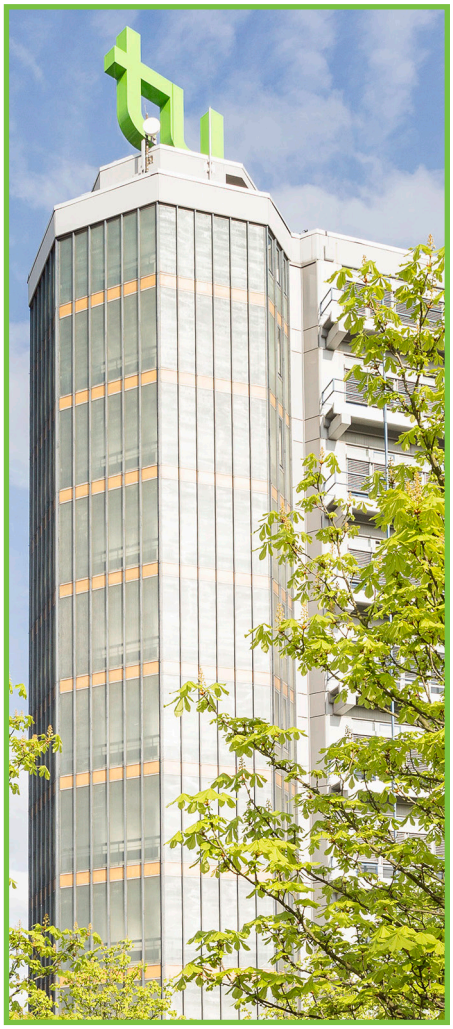
37th International Workshop on Statistical Modelling

16.07. – 21.07.2023

Dortmund

Proceedings book





Proceedings of the 37th International Workshop on Statistical Modelling

July 17-21, 2023 - Dortmund, Germany

Editors
Elisabeth Bergherr
Andreas Groll
Andreas Mayr

TU Dortmund University

International Workshop on Statistical Modelling (37°. 2023. Dortmund)

Proceedings of the 37th International Workshop on Statistical Modelling : July 17-21, 2022
Dortmund, Germany / Elisabeth Bergherr, Andreas Groll, Andreas Mayr (editors). – Dortmund : TU Dortmund University, 2023. – 1 copy online : PDF (693 S. : ill.)

ISBN: 978-3-947323-42-5

Authors:

Bergherr, Elisabeth
Groll, Andreas
Mayr, Andreas

Topics:

1. Statistics congress. 2. Econometrics models congress
330.015195 = Mathematical statistics

Editors:

ELISABETH BERGHERR
University of Göttingen, Chair of Spatial Data Science and Statistical Learning

ANDREAS GROLL
TU Dortmund University, Department of Statistics

ANDREAS MAYR
University of Bonn, Department of Medical Biometry, Informatics and Epidemiology

ISBN 978-3-947323-42-5 (online)

TU Dortmund University
Department of Statistics
Vogelpothsweg 78
44227 Dortmund
Germany

https://ub.tu-dortmund.de/
https://statistik.tu-dortmund.de/

Scientific Committee

Ruggero Bellio

University of Udine (Italy)

Elisabeth Bergherr (Co-Chair)

University of Göttingen (Germany)

Fernanda De Bastiani

University of Pernambuco (Brazil)

María L. Durbán Reguera

University of Madrid (Spain)

Jan Gertheiss

Helmut Schmidt University, Hamburg (Germany)

Andreas Groll (Chair)

TU Dortmund (Germany)

Thomas Kneib

University of Göttingen (Germany)

Dae-Jin Lee

IE University, School of Science and Technology, Madrid (Spain)

Andreas Mayr (Co-Chair)

University of Bonn (Germany)

Fulvia Pennoni

University of Milano-Bicocca (Italy)

María Xosé Rodríguez Álvarez

University of Vigo (Spain)

Gunther Schauberger

TU München (Germany)

Nicola Torelli

University of Trieste (Italy)

Lola Ugarte

University of Navarra (Spain)

Nikolaus Umlauf

University of Innsbruck, (Austria)

Helga Wagner

University of Linz, (Austria)

Local Organising Committee

Chiara Balestra

TU Dortmund University

Elisabeth Bergherr (Co-Host)

University of Göttingen

Guillermo B. Sánchez

TU Dortmund University

Jennifer Engel

TU Dortmund University

Alexander Gerharz

TU Dortmund University

Colin Griesbach

University of Göttingen

Andreas Groll (Host)

TU Dortmund University

Tobias Hepp

University Erlangen-Nürnberg

Hannah Klinkhammer

University of Bonn

Andreas Mayr (Co-Host)

University of Bonn

Hendrik van der Wurp

TU Dortmund University

Preface

Dear Participants,

we are more than happy to host the 37th International Workshop on Statistical Modelling in Dortmund, Germany! This is the second year to meet in person after the COVID break, and we hope to have a wonderful time like we did last year in Trieste.

This year we will have 54 contributed talks and more than 60 posters, and it was a tough challenge to pick among the many excellent submissions we had! Thanks again to the scientific committee for putting so much work into the selection process. But we obviously also want to give our thanks to all the researchers, who contributed with their great submissions and made it possible to put together such an excellent set of presentations. Having a special focus on students is a tradition of the Statistical Modelling Society, hence we are especially happy to welcome such a large number of younger researchers contributing to the conference. We are already excited to find out who will win the awards for best student paper, best student presentation and best student poster! The Statistical Modelling Society furthermore awarded travel grants to two students.

We will also have five great invited talks, from different areas in statistics: Brian Reich, Maria Iannario, Alexander Gerharz together with Matthias Kolodziej, Gillian Heller and Simon Wood agreed to give keynotes at the workshop. Furthermore, Andreas Bender and Fabian Scheipl will provide a short course about Piece-wise Exponential (Additive) Models (PEMs / PAMs) before the conference starts.

As always, the IWSM is a one-track conference, leading to a familiar atmosphere and to the possibility for communication between the different fields of statistical modelling.

Looking back at all the years we were participating in great workshops, hosted at so many different universities and all the amazing people we got to meet there, we are both humble and exited to welcome you all to enjoy the conference and your stay at the river Ruhr area with its long tradition of coal mining and steal production, beer brewing and, of course, its omnipresent football vibe.

Andreas Groll, Elisabeth Bergherr and Andreas Mayr

Dortmund, Göttingen and Bonn, July 2023

Contents

Part III

Part I

Data science meets football

Alexander Gerharz[1,2], Mathias Kolodziej[2]

[1] TU Dortmund University, Germany
[2] Borussia Dortmund, Germany

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

Abstract: Over time, the world of Statistics and Data Science has more and more found a way into the world of Sports. Professional football teams have started to rely more and more on the results of statistical analysis and statistical models to make decisions. One of the fields in which statistical analysis is used is about the physical performance of players and their risk to get injured. A lot of metrics are developed to express information found in the underlying data. Even though this field has gotten more attention in recent years, some questions have not found perfect answers yet and still need some more research regarding the statistical methods used. One of these questions to answer is about the physical development of a player. Can we give a definite answer if a player has statistically improved without having a good estimation for a measurement error and a variance? Also, another question to answer is about how to include rarely measured variables into a statistical injury model. Can we include variables with lots of missing data into our statistical models without being able to impute them properly?

Keywords: Sports Analysis; Injury Risk; Football; Statistical Modelling.

1 Data Analysis in Football

Professional football teams no longer rely on just the influence of the coaches and a few number of medical personal to influence the performance of the players in training and competitive matches. It is now common for elite clubs to operate with a more diverse range of support staff, who fulfil specialist roles related to the development of performance of both the individual, and the team (Drust, 2019).

Furthermore, football has undergone a significant transformation in recent years: it has become more dynamic, intense and complex. As a result of this development, however, football also entails a considerable risk of

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

injury. However, not only has the physical development of the game increased rapidly, but the technological progress in professional football is also impressive. New and complex technologies are finding their way into professional football through advances in sports science and data analysis (Seshadri et al., 2021). For this reason, there is an opportunity to better understand how biomechanical, physiological, and biochemical data relate to injury risk and performance development. In addition, there is the opportunity to use more advanced statistical approaches to better understand the extent to which collected data can be translated into training-relevant decisions. This will allow the multidisciplinary team to make more informed, comprehensive, and precise decisions for training than ever before.

In order to optimize the training process in professional football and to minimize the risk of injury, it is first necessary to understand the manner and magnitude in which a prescribed training dose produces a specific physiological response (training dose-response relationship/fitness-fatigue model). Due to the significant physiological and performance differences within a team, ensuring an optimal training load for each player (individualization) to maximize performance has become a critical part of daily monitoring (Scott, Lockie, Knight, Clark, & de Jonge, 2013). Objective measurement methods for monitoring external and internal load have become much more accessible in recent years due to technological advances and provide a good insight into external and internal responses to training load (Seshadri et al., 2021). A reliable recommendation requires a standardized framework in which collected data are organized, transformed, analyzed, and visualized (Lacome, Simpson, & Buchheit, 2018). From a sport science perspective, it is essential to understand when a change in a test procedure is significant for assessing physical performance ability (e.g., hip muscle strength). To do this, the measurement error of the variable must first be determined and the smallest significant change must be identified. In addition, z-scores or STEN scores are suitable for comparing and classifying a current measured value with historical measured values as part of an outlier analysis. Time series analyses (trend analyses) using linear models can also be used to assess the physical performance development of a player. More complex machine learning procedures, such as the optimization of a decision tree using the CART method, help to assess the risk of injury and to create injury risk profiles (Kolodziej, Nolte, Schmidt, Alt, & Jaitner, 2021). With the help of interpretable machine learning methods, rules can still be derived from complicated models that have a high practical relevance.

As part of the decision-making process in the training process, the role of sports science is to provide the multidisciplinary team with the right information at the right time. Among the various components of effective sports science support, three aspects are elementary:

1. adequate understanding and analysis of the data, i.e. using only the most relevant and useful metrics that have a direct impact on the training process.

2. interactive, intuitive and informative reports ("simple but powerful)" through measured visualization using business intelligence tools (e.g. Tableau).

3. communication skills and personality traits that help communicate data and reports to management, coaches, staff and players.

2 Data

In the field of physical performance, there are three main areas from which we get data for our statistical analysis:

1. Medical data, which is provided to us by our doctors and physicians. These contain data about the injuries and well-being of our players.

2. Monitoring data, which is measured on a regular basis about muscle strengths, jump heights, etc. Also, data from questionaires about the well-being of the players are collected here.

3. GPS data, which is collected by our players wearing GPS sensors or GPS data provided by the league.

The medical data contains a lot of information about the type of injuries that the players had and also about the treatments that were given. Even though this contains a lot of interesting information, as it is very individual for each and every type of injury it is still a challenge how to include them into a nice statistical analysis regarding the risk of an injury. However, to roughly assess what type of injuries the club has to deal with on a regular basis this data is used.

The monitoring data contains a lot of different types of information as different measurements are done over the course of a season. We use this kind of data to not only assess the physical development of the players, but also to screen the players regarding risk factors that might lead to injuries. First of all, the preparation phase of a new season starts with a performance measurement phase in which all players are tested regarding their body measurements, strengths, balances, stamina, etc. As this is a very time consuming process, usually this needs multiple days of time. Within the season there are some measurements that are performed more frequently (e.g. hip muscle strength, questionaires) and some that are measured less frequently (e.g. stamina) as they might be very taxing for the body or just take too long to measure them frequently. Also, sometimes measurements

are performed because of a previous injury of a specific player, so we might also have types of measurements for just one specific player only.

The GPS data contains all the information regarding the players moving on the pitch, even in training. The measurement is done with very precise instruments measuring on a frequency of about 100Hz (100 measurements per second). From this data, the running distance, the speed and the acceleration can be derived. All the information can then be aggregated into metrics regarding a whole training session or a complete match as we are interested in the complete running distances (with respect to different speed intervalls), the highest speed a player had within the session or the number of fast accelerations and fast decelerations a player had. As this kind of data is collected in every training session on the pitch and also in every match it is very helpful to measure the physical development of players and how taxing the training sessions and the matches are for the players.

3 Data Analysis

3.1 Basic Analysis

One of the biggest challenges of performing statistical analysis in a professional football club is to break all the information down into simple graphics and simple numbers, which other staff members can understand. Even though good communication can help, graphics and numbers should always be presented in a way, which other people could understand on their own. For this purpose, lots of information is presented in very basic ways (e.g. monitoring data with dot and line plots for a specific measurement taken multiple times over a long period). Also, it is of interest to detect trends within the physical performance of individual players over a specific period. To make this as understandable as possible just basic linear models are used to give a trend and also a statistical significance.

3.2 Evaluation of Physical Performance

Another one of the challenges in the statistical analysis comes with the nature of the data collection process. All measurements for a specific player have one starting point in time. This might be due to the fact that the measurement process has started then or because the player just joined the team on this point of time. Let's assume we measure a new kind of leg strength. Even if we have not measured this strength now, a player would have always had some kind of value for this measurement (see Figure 1, top left). If we would have started at the exemplary measurement timepoint 1, we can see in the top right that the player would not have a very significant development regarding this leg strength. In the bottom left of this Figure,

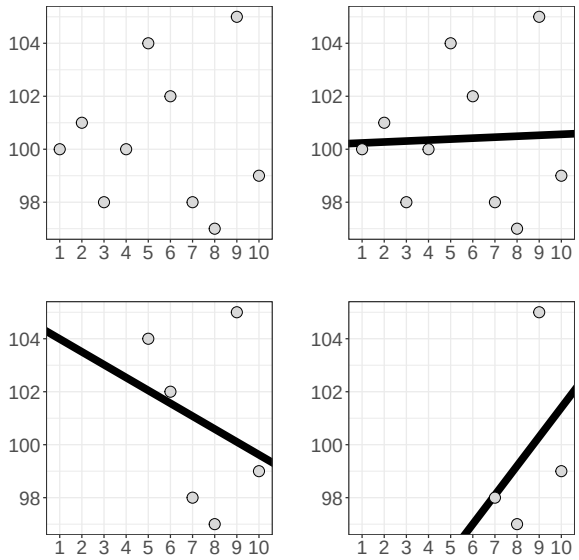


FIGURE 1. Exemplary measurements and trend calculations.

we do not have this value up until a very long time ago, but we just have the last 6 measurements and in the bottom right just 4 measurements. In the end, our conclusion can be extremely different depending on when we started to measure. One can also easily see that the computed trend will differ if, e.g., one measurement in between was skipped as with a low amount of data points each one can have a huge impact on the computation of a trend. This kind of issue can be addressed by good communication as, aside of giving information about the trends, we also need to communicate if we have enough data points to be sure of the trends or if the calculation of the trends are not significant.

Using Figure 1 as an example again, there is another challenge, which occurs on a daily basis. After taking a new measurement, it is of interest if it was an improvement, a stagnation or if the player has gotten worse (Paton & Hopkins, 2005). Of course, just checking if the value was higher or lower than the last measurement is a valid solution, but all measurements have some kind of a measurement error and we need to take this into account (Hopkins, 2004). For some measurements, this error also has the potential to vary over time and can be very different regarding the athlete. The estimation of this error is hard to compute in reality. If for each measurement we would like to estimate the variance of the measurement, we would need repeated measurements for each athlete. As this might be feasible for some measurements, stamina values are a great counter example. Stamina

measurements take a lot of time and are very taxing for the body, so even if a player would agree to repeatedly do stamina runs the measured variance would be bloated due to fatigue and decreasing ability. The resulting variance would not be representative for the error of the first measurement.

In Sports Science there exists an approach to estimate this variance as it is simply replaced by the variance of the past measurements of a player. While this actually gives a plausible variance, as the underlying player ability for the measurement varies over time the variance of a single measurement gets bloated by the variance of the underlying player ability.

3.3 Injury Risk Model with Missing Data

A completely different challenge arises when using monitoring data in statistical injury risk modelling. While GPS data is collected on a daily basis with almost no gaps at all monitoring data is collected with a much lower frequency, which leads to a lot of missing values in the covariates for an injury risk model. While some monitoring variables might not differ as much over time and can be imputed with a last-observation-carried-forward approach for a certain time, others can vary a lot on a daily basis and can not be imputed as good. One of our monitoring variables is Creatin Kinase (CK), which is a blood measurement value and indicates how badly the muscle cells of a player are damaged. If we measure this every day, then we can also see how well a player regenerates, so it is not just the absolute value, but also the change over time that contains important information. Missing values here can not just be imputed, because that would create a false impression of the regeneration process of a player. Also random events like tackles or bumping into each other can significantly impact this value, which also makes it hard for imputation.

In the end, we will end up with a data set in which some of our covariates contain a very high ratio of missing data. Most of the classical statistical models can not deal with missing data in the modelling process, which gives us basically four options. We could either just include all full observed observations, which will be less then 0.1% of all data points. The main problem here is that we would also need to have fully observed data points for our predictions, which in reality we rarely have. Another approach would be to fit one model for each combination of the covariates and include as much observations as possible for these kind of models, but over the course of a season there are a lot of different monitoring tasks, which are performed, which will lead to lots of different models. This might also lead to incomparable outputs. As mentioned before, imputation is hard, but we could try our best with different imputation approaches. It is possible to impute the missing values for all variables, but we could also narrow our

data set down to the variables that are easy to impute and just exclude the others.

The last option is to use methods that actually allow the underlying data set to contain missing data. One of these approaches could be a single decision tree using the CART method and allowing surrogate splits (Lewis, 2000). A surrogate variable functions as a replacement of a used splitting variable. After determining the best split of the data in a specific node, possible splits with other covariates are looked at, which can seperate the data as similar as possible. If now the variable that is used for the split is missing then the best surrogate split variable can be used as a substitute. If this variable is also missing then the next best substitute can be used and so on. This introduces a form of error within the model, but allows us to deal with missing data. Another approach could be a k-nearest-neighbor approach using the Gower's distance (Gower, 1971). Here each variable is evaluated on its observed range of values and produces a difference between 0 and 1. In the end an average distance is computed. If one of the two observations for which the distance is measured has a missing value for a specific variable, then this variable is excluded for the computation of the average distance. Even though, here the predictions are done with varying information and can result in different levels of confidence, we always include as much information as possible to determine the nearest neighbors to evaluate the injury risk for a new observation.

4 Conclusion

In a professional football club lots of different kinds of data are collected from a lot of different data sources and methods from the field of Statistics and Data Science are used to extract as much information as possible from these sources. The work of a Data Analyst with a focus on the physical performance of the players contains a lot of different questions that need to be answered with this kind of data. With the nature of the data collecting process and the daily necessities in a professional football club a very specific demand of Statistical methods arises. While some of the tasks can be solved with basic methods others need very specific sometimes taylor-made methods to provide valuable information.

It is a very difficult task to evaluate the improvement of a player with only a few datapoints, which makes the computation of trends very unstable. Also, when comparing different measurements of a player, for some tasks it is easy to account for measurement errors for some tasks it is actually an (almost) unsolvable problem. There exist approaches to roughly estimate the measurement errors, but these estimations are usually giving very generous upper bounds.

Another arising demand is for methods to assess the risk of an injury which allow for a high ratio of missing data. While some of the covariates for injury risk models are collected very frequently others are collected very rarely, which leads to big gaps in the data. Modelling methods that can actually account for missing data are thus in demand to not loose to much information.

References

Drust, B. (2019). Applied science and soccer: a personal perspective on the past, present and future of a discipline. *Sport Performance & Science Reports*, **56**(v1).

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, **27**, 4, 857 – 871.

Hopkins, W. G. (2004). How to interpret changes in an athletic performance test. *Sportscience*, **8**, 1 – 7.

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.

Lacome, M., Simpson, B., and Buchheit, M. (2018). Monitoring training status with player-tracking technology. *Still on the way to Rome. Aspetar. J*, **7**, 55 – 63.

Lewis, R. J. (2000). An introduction to classification and regression tree (CART) analysis. In: *Annual meeting of the society for academic emergency medicine in San Francisco, California. Vol. 14*, California.

Paton, C., and Hopkins, W. (2005). Competitive Performance of Elite Olympic-Distance Triathletes: Reliability and Smallest Worthwhile Enhancement. *Sportscience*, **9**, 1 – 5.

Scott, B. R., Lockie, R. G., Knight, T. J., Clark, A. C., and de Jonge, X. A. J. (2013). A comparison of methods to quantify the in-season training load of professional soccer players. *International journal of sports physiology and performance*, **8**(2), 195 – 202.

Seshadri, D. R., Thom, M. L., Harlow, E. R., Gabbett, T. J., Geletka, B. J., Hsu, J. J., ..., Voos, J. E. (2021). Wearable technology and analytics as a complementary toolkit to optimize workload and to reduce injury burden. *Frontiers in sports and active living*, **2**, 228.

Robust regression modelling for ordinal categorical data

Maria Iannario[1]

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

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

Abstract: Ordinal regression models are commonly implemented for analysing an ordinal response variable as a function of some explanatory covariates. The Maximum Likelihood (ML) estimators are used for estimating the unknown parameters of these models but gross-errors in the response, specific deviations due to the respondents' behaviour, and outlying covariates may affect their reliability. The paper emphasises that the choice of the link function can influence the robustness of inferential methods. In addition, robust M estimators are proposed as an alternative to ML estimators producing more reliable results.

Keywords: Anomalous data; Link functions; Ordinal response models; Robust inference.

1 Introduction

In recent years the analysis of ordinal response data has become a popular topic in mainstream research. Such data occur in many areas of scientific study, for example in psychology, sociology, economics, medicine, political science and many other disciplines, where the final response of a subject belongs to a finite number of ordered categories. Pioneering work in this field was carried out by McCullagh (1980), who advocates the use of a latent continuous variable that drives ordinal responses based on some unknown cut-off. This method has become popular because it allows us to treat the ordinal response pattern within the framework of the Generalised Linear Model (GLM) (Nelder and Wedderburn, 1972).
In this area, the false conjecture that the bounded support of the response variable (integer values between 1 and m) cannot generate anomalous data which jeopardize the reliability of the estimators and of the related tests has discouraged for long the analysis of the robustness of the inferential

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

methodologies in contrast with the rich literature on robustness for continuous and specific discrete data.

When ordinal data come from surveys, it has been recognised that respondents may thoughtfully or unconsciously choose the 'wrong' category. This phenomenon, in addition to the occurrence of gross errors (the probability of which is never negligible) or the irregular behaviour of some respondents, produces a contamination of the assumed distribution of the model, which may alter the reliability of the Maximum Likelihood (ML) estimators and of the related test procedures. Only lately studies have been carried out to improve the properties of estimators and tests in this field.

Lack of robustness for the ML estimators in the ordinal logistic regression model has been studied for instance in Croux *et al.* (2013), Iannario *et al.* (2017), and recently in Iannario and Monti (2023a, 2023b). Croux *et al.* (2013) studied a weighted maximum likelihood (WML) estimation method under the logit link function, through different choices of the weight function. Iannario *et al.* (2017) proposed a general M estimation procedure with the objective function chosen as a weighted likelihood function under different considerations of link functions. Unlike the approaches of Croux *et al.* (2013), where the weights are the function of robust Mahalanobis type distances, Iannario *et al.* (2017) considered Huber's weights for different link functions. They pointed out that a good weight function essentially controls the influential observations with respect to some reference models. In Iannario and Monti (2023b) robust estimation for unordered and ordered response models based on the logistic link function has been developed extending the results of Iannario *et al.* (2017) and Scalera *et al.* (2021) on the proportional odds model. In addition in Scalera *et al.* (2021) the impact of the chosen link on the estimators has been discussed, since contrasting results obtained by the links can highlight the occurrence of anomalous responses.

The paper summarises the main findings on the topic, analysing some robustness issues and diagnostic procedures for ordinal response models and exploiting some properties of the M estimator from Iannario *et al.* (2017), Iannario and Monti (2003b) which are able to produce reliable inference in case of data contamination. The contribute also provides some suggestions for new research topics.

The paper is organized as follows: the next Section deals with a brief overview of the methods based on the GLM framework for ordinal response models with special emphasis on the role of generalized residuals. Section 3 gives some insights on the robustness inference and highlights that M estimation can make a difference in fitting the models when anomalous responses occur. Diagnostic procedures along with empirical evidences are illustrated in Section 4. Concluding remarks have been made in Section 5.

2 Maximum likelihood inference

Let Y be a m-category ordinal response variable which represents the discrete measurement of an underlying (continuous) latent variable Y_i^* such that, for any i-th subject,

$$\alpha_{j-1} < Y_i^* \leq \alpha_j \qquad \Longleftrightarrow \qquad Y_i = j\,, \qquad j = 1, 2, \ldots, m\,, \qquad m > 2\,,$$

and $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_m = +\infty$ are the cut-off points in the continuous support of the latent variable Y^*.
The $i-$th copy of Y^* linearly depends on $p(\geq 1)$ covariate(s) through $\boldsymbol{x}_i = (x_{i1}, \ldots x_{ip})$ as $Y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$, $i = 1, 2, \ldots, n$, where $\epsilon_i \sim G(\cdot)$, which is assumed to have a probability density function $g(\cdot)$. If $G(\cdot)$ is the standardized normal cumulative distribution function (*cdf*) we have the probit link, while when $G(\cdot)$ is the logistic *cdf* we have the logit link. An alternative choice considered in the literature is the inverse of the *cdf* of the extreme values (or log-Weibull) distribution (Agresti, 2010) but it is also possible to consider the Cauchy link related to Cauchy *cdf*, rarely chosen, and even more rare the Student (or Gosset) link for Student *cdf* (Albert and Chib, 1993).
Under the above parametric set-up, the probability mass function of the cumulative link model is

$$\begin{aligned} Pr(Y_i = j \mid \boldsymbol{x}_i) &= Pr(\alpha_{j-1} < Y_i^* \leq \alpha_j) \\ &= G(\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta}), \quad j = 1, 2, \ldots, m. \end{aligned} \tag{1}$$

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{m-1})'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, then $\boldsymbol{\theta} \in \Omega(\boldsymbol{\theta})$; the latter is an open subset of $\mathbb{R}^{p+m-1}$. Given an observed random sample $(y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \ldots, n$, let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and let $\boldsymbol{X}$ be the matrix whose rows are given by $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. The log-likelihood function (McCullagh, 1980) of the sample is

$$\begin{aligned} \ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{X}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} I[y_i = j] \log Pr(Y_i = j|\boldsymbol{x}_i) = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} I[y_i = j] \log\left[G(\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta})\right] \end{aligned}$$

where $I[\omega]$ is the indicator function which takes value 1 if ω holds and 0 otherwise. The score function is $s(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} s(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)$ where $s(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i) = (s_{\alpha_1}, \ldots, s_{\alpha_{m-1}}, \beta_1, \ldots, \beta_p)'$. Here,

$$s_{\alpha_s}(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i) = \begin{cases} \dfrac{-g(\alpha_s - \boldsymbol{x}_i\boldsymbol{\beta})}{G(\alpha_{s+1} - \boldsymbol{x}_i\boldsymbol{\beta}) - G(\alpha_s - \boldsymbol{x}_i\boldsymbol{\beta})} & \text{if } s = j-1 \\ \dfrac{g(\alpha_s - \boldsymbol{x}_i\boldsymbol{\beta})}{G(\alpha_s - \boldsymbol{x}_i\boldsymbol{\beta}) - G(\alpha_{s-1} - \boldsymbol{x}_i\boldsymbol{\beta})} & \text{if } s = j \\ 0 & \text{if } s \neq j-1, j, \end{cases} \tag{2}$$

$$s_{\beta_r}(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i) = -\sum_{j=1}^{m} I[y_i = j] e_{ij}(\boldsymbol{\theta}) x_{ir} \tag{3}$$

and the quantities $e_{ij}(\boldsymbol{\theta})$ in (3) are the generalized residuals (Franses and Pap, 2004, Iannario and Monti, 2023a)

$$e_{ij}(\boldsymbol{\theta}) = \frac{g(\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}) - g(\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta})}{G(\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}) - G(\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta})}\,,\; i = 1, 2, \ldots, n\,; \quad j = 1, 2, \ldots, m\,. \tag{4}$$

The score function for the whole sample, related to the scalar regression coefficient β_r, is given by

$$\sum_{i=1}^{n} s_{\beta_r}(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i;) = -\sum_{i1}^{n}\sum_{j=1}^{m} I[y_i = j] e_{ij}(\boldsymbol{\theta}) x_{ir} \tag{5}$$

which has the same structure of the Gaussian (ML) equation in the linear model. Hence the combination of $e_{ij}(\boldsymbol{\theta})$ and $\boldsymbol{x}_i$ determines the impact of the observations on the estimators. This point has been extensively discussed in Iannario and Monti (2023a) where it is stressed that outlying covariates induce both the numerator and denominator of (4) to approach 0 and thus the final value of residuals depends on the speed of convergence of the two terms. This last point leads to suggest that the link function be chosen in such a way $e_{ij}(\boldsymbol{\theta})$ are bounded (see also Scalera *et al.* 2021, for further details).

Lastly, the generic term of the information matrix $\mathcal{I}(\boldsymbol{\theta}, \boldsymbol{X})$ for a single observation, conditionally on $\boldsymbol{X} = \boldsymbol{x}$, is given by

$$\begin{aligned} \mathcal{I}_{ls}(\boldsymbol{\theta}, \boldsymbol{x}) &= E_Y\left\{-\frac{\partial^2 \ell(\boldsymbol{\theta}, Y, \boldsymbol{X})}{\partial\,\theta_l\,\partial\,\theta_s}\,\middle|\,\boldsymbol{X} = \boldsymbol{x}\right\} \\ &= -\sum_{j=1}^{m} I[Y = j]\,\frac{\partial^2 \ell(\boldsymbol{\theta}, Y, \boldsymbol{x})}{\partial\,\theta_l\,\partial\,\theta_s}\,Pr(Y = j|\boldsymbol{x}), \end{aligned}$$

for $(l, s) = 1, 2, \ldots, m + p - 1$, and the elements of the unconditional information matrix $\mathcal{I}(\boldsymbol{\theta})$ are given by $\mathcal{I}_{ls}(\boldsymbol{\theta}) = E_{\boldsymbol{X}}\{\mathcal{I}_{ls}(\boldsymbol{\theta}, \boldsymbol{X})\}$.

3 Robust inference

A robust estimation requires the influence function of the estimator to be bounded (Hampel *et al.*, 1986; Huber and Ronchetti, 2009). When the ML estimators are used, the influence function is proportional to the score function, i.e. $IF_{ML}(y, \boldsymbol{x}_i; \boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\theta})^{-1} s(\boldsymbol{\theta}; y, \boldsymbol{x}_i)$ for $y = 1, 2, \ldots, m$. Consequently the sources of unboundness of the score functions should be investigated.

By focusing on (3) two important sources of unboundness appear: the regressors and the generalized residuals. Outlying regressors can occur anytime the covariates include unlimited variables. Large generalized residuals can be produced by two kinds of events: anomalous responses, as shown by Iannario *et al.* (2017, Section 2, example 1), and by outlying covariates (as reported in Iannario and Monti (2023b), Iannario *et al.* (2023)). As mentioned the impact of a large $\boldsymbol{x}_i$ on $e_{ij}(\boldsymbol{\theta})$ is filtered through $G(\cdot)$ and may be limited by an appropriate choice of the link function.
The latter choice is indeed crucial, because it determines also the behavior of the threshold score function (2). This choice is extensively discussed in Iannario *et al.* (2017, Section 3) where the nice robustness properties of the logit link, which is generally the most recommended and used, are emphasised.
For the cases links different from the logit one are applied or outlying regressors may occur, Iannario *et al.* (2017) propose an M estimator which is the implicit solution $\hat{\boldsymbol{\theta}}_M$ of

$$\sum_{i=1}^{n} \psi(y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_M) = \sum_{i=1}^{n} s(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)\, w(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) - a(\boldsymbol{\theta}) = 0, \qquad (6)$$

where $a(\boldsymbol{\theta}) = E\big\{s(\boldsymbol{\theta}; Y, \boldsymbol{X})\, w(Y, \boldsymbol{X}; \boldsymbol{\theta})\big\}$ and this term is required to achieve Fisher consistency. The weights $w(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ in (6) are designed to down-weight outlying observations in order to control their impact in the estimation. Of course, if $w(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) \equiv 1$ then $a(\boldsymbol{\theta}) \equiv 0$ and $\psi(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$ coincides with $s(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})$, by displaying the ML estimators as a special case of M estimators. Moreover, let $M(\boldsymbol{\theta}, \psi) = -E\left\{\dfrac{\partial}{\partial\, \boldsymbol{\theta}}\, \psi(Y, \boldsymbol{X}; \boldsymbol{\theta})\right\}$, the influence function of the M estimator is

$$IF\,(y, \boldsymbol{x}_i; \psi) = M^{-1}(\boldsymbol{\theta}, \psi)\psi\,(y, \boldsymbol{x}_i; \boldsymbol{\theta})\ \text{, for } y = 1, 2, \ldots, m \text{ and } \boldsymbol{x}_i \in \mathbb{R}^p\,.$$

The influence function is bounded if $\psi(.)$ is bounded, and this goal is achieved by choosing suitable weights.
To get an insight on what is an appropriate weight function, the attention should be focused on (3) which, as remarked in Section 2, recalls the normal equation in the linear model. If a large $\boldsymbol{x}_i$ is associated with a large generalized residual $e_{ij}(\boldsymbol{\theta})$, the corresponding statistical unit will have a dominating role in (5) when determining the estimate of the parameter. If instead a large generalized residual is associated to a small $\boldsymbol{x}_i$ or viceversa a large $\boldsymbol{x}_i$ is associated to a small $e_{ij}(\boldsymbol{\theta})$, then the impact of anomalous data in the estimation process is limited. In other words, observations characterized by a large product $e_{ij}(\boldsymbol{\theta})\boldsymbol{x}_i$ are to be considered and treated as leverage points in an analogous fashion to what happens in the linear model.
Consequently the Huber weights (Hampel *et al.*, 1986) are proposed, which are non-increasing functions of the magnitude of both the residuals and

the covariates, leading to $w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = min(1, c/\sum_{j=1}^{m} I[y_i = j] \mid e_{ij}(\boldsymbol{\theta}) \mid \|\boldsymbol{x}_i\|)$, where $||\boldsymbol{x}_i||$ is the norm of the covariates which needs to be based on robust estimators of location and scale and c is a suitable tuning constant. The value of c should be chosen so that the loss of efficiency incurred by M estimators, with respect to the ML estimators, does not exceed a given threshold (say 5% or 10%) when there is no contamination in the data. On the basis of a thorough investigation Iannario *et al.* (2017) suggest an appropriate values of c which vary roughly between 1 and 2 according to the trace criterion, whereas in Iannario and Monti (2023b) $c = \sqrt{\chi^2_{p,\xi}}$ where $\chi^2_{p,\xi}$ is the ξ-th percentile of the χ^2_p distribution.
In general we need small values of c for greater efficiency under the model and large values of c for greater stability away from it. The researcher usually does not know, a priori, the amount of contamination in the data. So a data driven selection of the 'optimal' tuning parameter is a remarkable topic. Among alternative existing approaches Warwick and Jones (2005) choose the optimum data-based tuning parameter by constructing an empirical version of the mean square error and minimizing it over the tuning parameter.
Croux *et al.* (2013) instead proposed alternative M type estimators with similar residuals but weights depending only on the covariates. This approach is also pursued in Iannario *et al.* (2023) where the weights are a decreasing function of $\|\boldsymbol{x}_i\|$ leading to the simplified version of $w(\boldsymbol{y}_i, \boldsymbol{x}_i) = \min(1, c/||\boldsymbol{x}_i||)$. When the tuning constant c increases the M estimators approach the ML estimators, whereas when c decreases extreme design points are strongly downweighted. The square of the Mahalanobis distance $||\boldsymbol{x}_i||^2$ can be compared with the percentiles of a χ^2_p distribution. These weights have the advantage that they are calculated only once at the beginning of the estimation process and do not have to be updated.
Continuing on with the inferential aspects the asymptotic variance-covariance matrix of the M estimator $\hat{\boldsymbol{\theta}}_M$ is

$$V(\boldsymbol{\theta}, \psi) = \boldsymbol{M}^{-1}(\boldsymbol{\theta}, \psi)\boldsymbol{Q}(\boldsymbol{\theta}, \psi)\boldsymbol{M}^{-1}(\boldsymbol{\theta}, \psi),$$

where $\boldsymbol{Q}(\boldsymbol{\theta}, \psi) = E\{\psi(Y, \boldsymbol{X}; \boldsymbol{\theta})\psi(Y, \boldsymbol{X}; \boldsymbol{\theta})'\}$. It can be estimated using the corresponding sample statistics $M(\boldsymbol{\theta}, \psi)$ and $Q(\boldsymbol{\theta}, \psi)$ (cfr Iannario *et al.* 2017). Under general regularity conditions (Huber, 1981), the M estimator $\hat{\boldsymbol{\theta}}_M$ is asymptotically normal, i.e.

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}\right) \rightarrow N(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\theta}, \psi)).$$

Finally robust testing procedures may be performed by means of a t-type statistics. Under the null $H_0^r : \beta_r = 0$, for $r = 1, 2, \ldots, p$,

$$t_r = \frac{(\hat{\beta}_r^M - \beta_r)}{\hat{V}_\beta^{rr}} \longrightarrow N(0, 1).$$

Here$\hat{V}_\beta^{rr}$ is the r-th element on the diagonal of $\hat{V}_\beta$ which is the submatrix of $\boldsymbol{V}_\psi$ related to regressors and $\hat{\beta}_r^M$ is the r-the element of the M estimator $\hat{\beta}_M$ of $\boldsymbol{\beta}$.

4 Real Data Analysis

The red wine quality data (`https://archive.ics.uci.edu/ml/datasets/wine+quality`) (Cortez *et al.* 2009) from the UCI Machine Learning Repository, contain physicochemical (quantitative covariates) and one sensory (the ordinal one) variables referring to white variants of the Portuguese 'Vinho Verde' wine. The ordinal categorical response variable with values ranging in $\{1, 2, \ldots, 6\}$ is reported in Figure 1 whereas Figure 2 shows the boxplots of the continuous covariates where various outlying points appear.

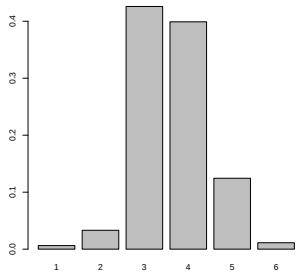


FIGURE 1. Frequency distribution of the ordinal variable Y.

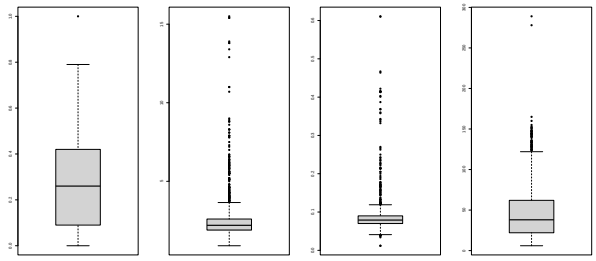


FIGURE 2. Boxplot of citric acid (first panel), residual sugar (second panel), chlorides (third panel), sulfur dioxide (fourth panel).

The tuning constant of the M estimators is computed as $c = \sqrt{\chi^2_{4,0.7}} = 2.209$, since there are four continuous covariates; it produces a limited loss

of efficiency in case of pure data (Iannario and Monti, 2023b). The Minimum Covariance Determinant estimators, which has a high breakdown point, have been applied for the estimators $\tilde{\boldsymbol{\mu}}_X$ and $\tilde{\boldsymbol{\Sigma}}_X$ that appear in the Mahalanobis distance given for the norm of the regressors. To simplify the analysis, the logit link is considered, bearing in mind that the score functions (2) of the thresholds are bounded and the generalized residuals vary within $(-1, 1)$.

The ML and the M estimates with weights depending on the covariates and on the residuals (similar results are obtained with M estimates with weights depending only on the covariates) in Table 1 are compared with the estimates from the ML estimator $\boldsymbol{\theta}_C$ related to the 'clean sample', i.e. a sample obtained by removing the statistical units corresponding to the outliers (whose norm is higher than $\chi^2_{4,0.95}$) replicating what was done in Iannario and Monti (2023b). The distance can be computed as $D(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_C) = \left\{(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_C)^T \hat{V}(\hat{\boldsymbol{\theta}}_{\boldsymbol{C}})^{-1}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_C)\right\}^{1/2}$ where $\hat{V}(\hat{\boldsymbol{\theta}}_C)$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\theta}}_C$. The ML estimate of *residual sugar* coefficient is not significant. In contrast the robust M estimates coefficients are all significant, since their standard errors are considerably smaller. Furthermore, the distance from the 'clean sample' estimates are $D(\hat{\boldsymbol{\theta}}_{ML}, \hat{\boldsymbol{\theta}}_C) = 4.314$ and $D(\hat{\boldsymbol{\theta}}_M, \hat{\boldsymbol{\theta}}_C) = 2.088$. The former is larger than the latter.

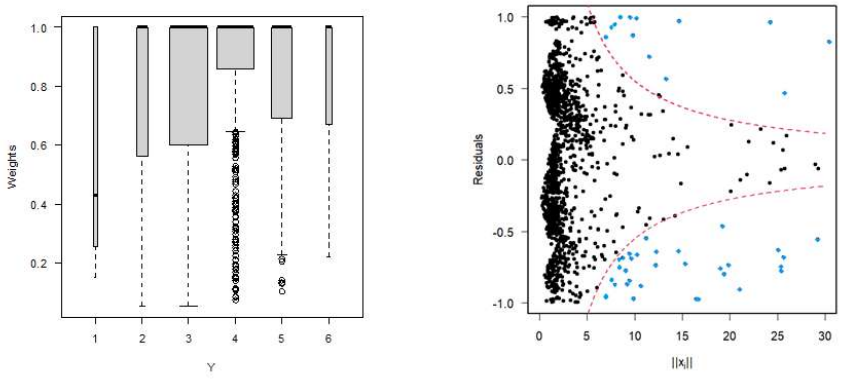


FIGURE 3. Left panel: Boxplots of the weights for each category of the response (the width of the box is proportional to the observed frequency of the corresponding category). Right panel: Residuals versus covariates (blue diamonds correspond to the influential data identified in the left panel).

Figure 3 (left panel) shows the boxplots of the weights for each category of the response. The weights associated to the statistical units display where M estimation applies a severe downweighting (i.e. $Y_i = 1$ has a larger

percentage of downweighted observations).
Figure 3 (right panel) presents the scatter plot of the residuals (derived from M estimation) versus the robust norm $||\boldsymbol{x}_i||$. The area delimited by the dashed lines incorporates data which are not severely influential. The influential observations are located outside this area and are characterized by a large product $e_{ij}(\boldsymbol{\theta})||\boldsymbol{x}_i||$.

TABLE 1. Estimates for the model for wine data.

	ML estimation			*Robust estimation*			*Clean data estimation*		
	Coef	St.Err	t-stat	Coef	St.Err	t-stat	Coef	St.Err	t-stat
α_1	-5.762	0.354	-16.294	-6.495	0.433	-15.016	-6.709	0.437	-15.353
α_2	-3.885	0.202	-19.238	-4.294	0.277	-15.493	-4.693	0.288	-16.275
α_3	-0.632	0.148	-4.263	-1.054	0.212	-4.975	-1.390	0.244	-5.694
α_4	1.563	0.157	9.977	1.215	0.218	5.587	0.837	0.245	3.409
α_5	4.273	0.276	15.503	3.899	0.325	12.003	3.575	0.338	10.571
$\beta_{1,citric\ acid}$	2.599	0.263	9.900	2.602	0.282	9.236	2.583	0.270	9.554
$\beta_{2,residual\ sugar}$	0.051	0.036	1.423	0.101	0.052	1.943	0.118	0.063	1.860
$\beta_{3,chlorides}$	-7.236	1.096	-6.604	-13.898	2.134	-6.514	-18.260	2.628	-6.949
$\beta_{4,sulfur\ dioxide}$	-0.014	0.002	-8.827	-0.014	0.002	-8.059	-0.015	0.002	-8.898

5 Conclusions

The lack of robustness in the likelihood based inferential procedures poses a major challenge in modelling ordinal response data. Here we explore a summary of alternative robust methodologies to estimate the parameters in such statistical models. The theory developed for the M estimators limits the impact of anomalous data on the fitted model, leading to a proper assessment of the effect of covariates. Theoretical results find a nice application in this article. Further studies concern performance indices for model comparison in order to achieve robustness of Akaike Information Criterion and the study of the impact of imputation methods on the outliers when missing values are analysed. Taking into account all possible challenges, we believe that the use of the estimators presented in the ordinal response models provides a useful tool for scientists working in the context of ordinal data.

Acknowledgments: I want to thank the IWSM2023 organizers and all the collaborators involved in the works reviewed in this short paper: Anna Clara Monti above all, with whom I co-authored most of the papers on the subject, and Domenico Piccolo, Elvezio Ronchetti, Valentino Scalera.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*. 2nd edition. Wiley, Hoboken

Albert, J.H., Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669 – 679.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009) Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, **47**, 547 – 553.

Franses, P.H., Paap, R. (2004). *Quantitative models in marketing research.* Cambridge University Press, Cambridge.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986). *Robust statistics: the approach based on infuence functions.* Wiley, New York.

Huber P.J. (1981). Robust Statistics, J. Wiley & Sons, New York.

Huber, P.J., Ronchetti, E.M. (2009). *Robust Statistics*, 2nd edition, J. Wiley & Sons, New York.

Iannario, M., Monti A.C. (2023a). Generalized residuals and outlier detection for ordinal data with challenging data structures. *Statistical Methods & Applications*. https://doi.org/10.1007/s10260- 023-00686-1

Iannario, M., Monti A. C. (2023b). Robust logistic regression for ordered and unordered responses. *Manuscript.*

Iannario, M., Monti A. C., Scalera D. (2023). Modeling financial risk attitude: the role of education and financial literacy *Manuscript.*

Iannario, M., Monti, A.C., Piccolo, D., Ronchetti, E. (2017). Robust inference for ordinal response models. *Electronic Journal of Statistics*, **11**, 3407 – 3445.

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

Nelder, J. A., Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A*, **135**, 370 – 384.

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

Warwick, J., Jones, M.C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation*, **75**, 581 – 588.

Back to the future: model what you measure

Gillian Heller[1]

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

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

Abstract: The evolution of statistical modelling has historically been constrained by the practical limitations of computation. As increased mathematical complexity often implies more intricate computation, over time statistical models have grown both mathematically and computationally more complex; but paradoxically sometimes conceptually simpler models present more computational challenges than complex ones. I shall be discussing two well-known examples of this phenomenon.

Keywords: distributional regression; logistic regression; relative risk regression; proportional hazards regression; Cox model.

1 Introduction

Statistical models are necessarily abstractions of the real world ("all models are wrong"). In the physical sciences the abstraction may be rather close to reality, when the phenomenon under study is well understood; in other areas such as social sciences the abstraction may be more speculative. In all cases, we observe data; and we formulate a statistical model to describe the data-generating mechanism, which is a mathematical abstraction of the real process.

When the purpose of the modelling is for prediction, the model's predictive ability is all that matters. Interpretability is not important, and in fact the model may be a "black box" (as in machine learning). However when the purpose of the modelling is exploratory or confirmatory, whatever the extent of the abstraction, it is generally accepted that the model should be as simple as possible while retaining interpretability and usefulness.

Before going further, we need to define what we understand by "simple", and to do this we need to distinguish between simplicity and mathematical

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

convenience. By simplicity we mean closeness to the truth of the data, or the evidence. For example, consider that we observe occurrences of a binary event. The most natural summary is the relative frequency, interpreted as a probability. Our contention is that this is as close as we can get to the evidence; and because it is close to the data, it is easily interpreted. We therefore regard the relative frequency or probability as a *simple* abstraction of the data. Another commonly-used summary of such data is the *odds*, defined as the relative frequency of occurrences to non-occurrences. This is not an intuitive concept, yet because of its mathematical convenience (discussed below), it is ubiquitous in the analysis of binary and categorical data. Despite the mathematical and computational convenience of statistical models for the odds, the odds is a *complex* abstraction of the data.

2 Binary outcomes

We consider the simplest situation of modelling a binary outcome as a function of a binary predictor (or risk factor or exposure or treatment allocation). The predictor at level 0 generally means the risk factor or exposure is absent, or the treatment allocation is to control; 1 means presence or active treatment. Standard notation and terminology is given in Table 1.

TABLE 1. Binary data.

		Event occurrence		
		No	Yes	Event rate
Predictor	0	a	b	$\mathrm{R}_0 = b/(a+b)$
	1	c	d	$\mathrm{R}_1 = d/(c+d)$

The event rates R_0 and R_1, for the predictor at levels 0 and 1, respectively, are alternatively referred to as *risks* of the event. To quantify the discrepancy between the event rates, two natural extensions are to define relative risk and risk difference:

- relative risk: $\mathrm{RR} = \dfrac{\mathrm{R}_1}{\mathrm{R}_0} = \dfrac{d/(c+d)}{b/(a+b)}$
- risk difference: $\mathrm{RD} = \mathrm{R}_1 - \mathrm{R}_0 = d/(c+d) - b/(a+b)$

both of which are intuitive quantities, in that, for example, a doubling of risk or a risk difference of 10% are concepts close to the data and unlikely to be misinterpreted. Clearly $\mathrm{RR} = 1$, or equivalently $\mathrm{RD} = 0$, indicate no difference in the risk for the predictor present or absent; there are simple statistical tests for this hypothesis.

We would generally want to extend the analysis of Table 1 to include multiple predictors, i.e. multiple regression with a binary outcome:

$$y|\boldsymbol{x} \sim \text{Bernoulli}(p); \qquad g(p) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} \ .$$

- *relative risk regression model:*
  $\log(p) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} \qquad \Rightarrow \qquad \exp(\beta_j)$ is the relative risk
- *risk difference regression model:*
  $p = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} \qquad \Rightarrow \qquad \beta_j$ is the risk difference

for a binary predictor x_j, or for a 1-unit increase in x_j, ceteris paribus. So by varying the link function $g(\cdot)$, we easily define regressions on the scale of relative risk and risk difference. And yet these regressions are infrequently used in the analysis of binary data. Estimation of the relative risk and risk difference regression models poses problems due to the fact that the log and identity link functions do not guarantee constraint of the fitted values $\hat{p}$ to $(0, 1)$, as required. Constrained optimisation for maximum likelihood estimates (MLEs) is intricate but the problems have largely been solved, using EM-type algorithms and an adaptive barrier approach which achieve stable convergence (Donoghoe and Marschner 2018).

Much better known for the analysis of binary outcomes is logistic regression, which is based on the logistic link function, giving effects on the odds:
$\log\left(\frac{p}{1-p}\right) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} \qquad \Rightarrow \qquad \exp(\beta_j)$ is the odds ratio (OR)
As is well known, the logistic link function is the canonical link for the Bernoulli distribution, so importantly delivers estimates $\hat{p} \in (0, 1)$ (as will the inverse of any sigmoid function). Computation of MLEs is straightforward; logistic regression is available in all statistical software packages and is the go-to method for binary outcome data. However the odds scale for regression effects is far less intuitive than relative risk, discussed by multiple authors (e.g. Knol et al 2011). OR and RR are approximately equal when R_0 is close to zero, or when $\mathrm{RR} < 1$. However when $\mathrm{RR} > 1$ and R_0 is not close to zero, OR exceeds RR, increasingly so as R_0 increases. So, for example, when $\mathrm{R}_0 = 0.4$ and $\mathrm{R}_1 = 0.8$, $\mathrm{RR} = 2$ and $\mathrm{OR} = 6$. While both measures indicate a substantial increase in risk, and would lead to qualitatively the same conclusion, $\mathrm{OR} = 6$ is a far more alarming statistic than $\mathrm{RR} = 2$, being understood as "six times the risk". In general the use of logistic regression will lead to qualitatively the same conclusions as relative risk regression, so in that sense its results are not misleading, but they are misleading if the quantification of the increase in risk is important. In summary, logistic regression is mathematically and computationally convenient, but is a complex abstraction of the data.

The left panel of Figure 1 shows the results of a search on PubMed Central (an archive of biomedical and life sciences journal literature) of the terms

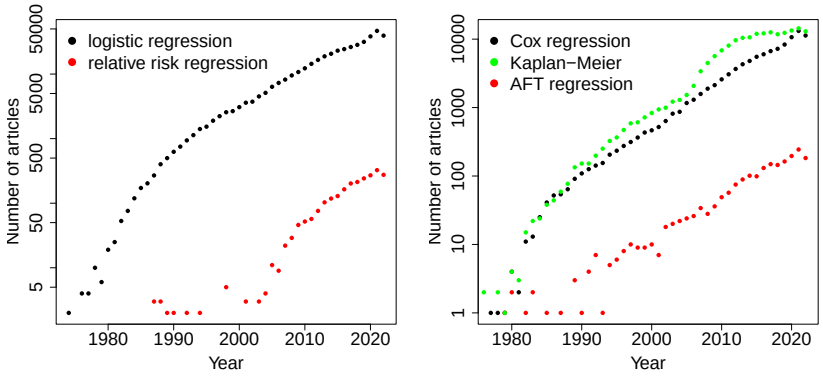


FIGURE 1. Number of journal articles on PubMed Central containing the search terms. Left panel: 'logistic regression', and ('relative risk regression' or 'log-binomial regression' or 'log binomial regression'). Right panel: ('proportional hazards regression' or 'cox regression'), 'Kaplan-Meier', and ('accelerated failure time' or 'parametric survival'). The y-axes are on the log scale.

'logistic regression', and ('relative risk regression' or 'log-binomial regression' or 'log binomial regression'), confirming the ubiquitous use of logistic regression, despite its interpretational difficulties discussed above; and the far more sparse use of relative risk regression.

3 Time-to-event outcomes

Time-to-event, or survival, outcomes have the typical feature of right-censoring, due to subjects either leaving the study or the study ending before observation of the event of interest. The simplest summary of the data, analogous to the computation of risks for binary data, is the estimated survival function, typically plotted as the Kaplan-Meier (KM) survival curve. This involves relatively straightforward computation of the sample probability of survival over time, dependent on the number of subjects at risk at any time point.

Proceeding to the next level of analysis, we incorporate multiple predictors into a model for survival time. A natural approach is a regression model for time to event t; were it not for the censoring issue, a GLM-like multiple regression model could look like

$$t_i|\boldsymbol{x}_i \sim \mathcal{D}(\mu_i, \sigma); \qquad \log(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \tag{1}$$

where $\mathcal{D}(\mu, \sigma)$ denotes a distribution with support on the positive real line; μ is a location parameter; σ is a scale/dispersion parameter; and the coefficients β_j are additive effects on $\log(\mu)$. Likelihood maximisation for such

models is straightforward; with censoring, estimation is somewhat more complicated, but generally feasible.

The development of regression models, and the ability to perform complex iterative computations was limited until the early 1970s, when the seminal paper of Nelder and Wedderburn (1972) introduced generalized linear models. These would have gone part of the way to solving (1); however this is not the trajectory that the analysis of survival data took. The closest model to (1) in the survival field is the *accelerated failure time model* (AFT, Kalbfleisch and Prentice 1980):

$$\log(t_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i \tag{2}$$

where $\epsilon_i \sim \mathcal{D}^*(0, \sigma)$ is a parametric distribution with support on the real line. Commonly-used choices for $\mathcal{D}^*$ are the normal, logistic and extreme value (Gumbel) distributions, which imply lognormal, log-logistic and Weibull regression models for survival time, respectively.

Yet historically things took quite a different turn. The ubiquitous approach to the modelling of survival data is the well-known proportional hazards (PH, or Cox) model:

$$h_i(t) = h_0(t)\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \tag{3}$$

where the *hazard function* $h(t)$ is the instantaneous probability of an event at time t, conditional on survival to time t; $h_0(t)$ is the "baseline hazard function" which is modified in (3) by the factor $\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ to get the hazard function for subject i; and $\exp(\beta_j)$ are multiplicative effects on the hazard function. The regression model (3) does not make any assumption regarding the distribution of the time to event; however implicit in (3) is the assumption that the hazard function has the same shape for all subjects ("proportional hazards").

The proportional hazards (PH) model (3) is very familiar to most statisticians; it is the go-to method for the analysis of survival data. And yet regression model (1) is a far more natural way of thinking about and modelling such data: survival time is observed, and effects on the mean or median survival time are simple and intuitive concepts. The hazard function, on the other hand, is not observed. It is a modelling abstraction and effects on it are, in the author's experience, not well understood by applied researchers.

To understand why a less-obvious model has come to dominate the field, we need to look at the history of survival analysis. In a wide-ranging interview with Nancy Reid (Reid 1994), Sir David Cox explained that he had been approached by "Quite a few people ... said they were getting a certain kind of data, censored survival data, with a lot of explanatory variables.

Nobody knew quite how to handle this sort of data in a reasonably general way, and there seemed to be dissatisfaction with assuming an underlying exponential distribution or Weibull distribution modified by some factor." Cox developed the PH model in response, with the breakthrough being the separation of the likelihood into a part that involved $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ and the part that involved $h_0(t)$, thus enabling maximisation of the partial likelihood and avoiding estimation of $h_0(t)$. This led to estimation which was feasible at the time, and avoided the need to specify the response distribution.

So by 1980 there were two competing regression models for survival data: the AFT (2) and PH (3) models. The PH model completely eclipsed the AFT model in popularity, and continues to do so: Cox's original paper (Cox 1972) is ranked 24th in *Nature*'s list of most cited papers of all time in all fields (Van Noorden et al. 2014). Citations are in fact an underestimate of the popularity of the method: it has become so mainstream that generally papers in applications journals use the "Cox model" without reference. A better indicator of usage of the models is the number of journal articles using the terms "Cox model" or "proportional hazards model". This is shown in the right panel of Figure 1, together with "accelerated failure time models" and "Kaplan-Meier" (obtained from PubMed searches). The pattern of PH vs AFT models is strikingly similar to logistic vs relative risk regression, albeit on a smaller scale. Note that Kaplan-Meier is even more widely used (according to this measure) than the PH model.

Sir David Cox appeared equivocal about the proliferation of his method. When asked about how he felt about the "cottage industry that's grown up around it" (Reid 1994), Cox replied "Don't know, really. In the light of some of the further results one knows since, I think I would normally want to tackle problems parametrically, so I would take the underlying hazard to be a Weibull or something. I'm not keen on nonparametric formulations usually .. if you want to do things like predict the outcome for a particular patient, it's much more convenient to do that parametrically ... another issue is the physical or substantive basis for the proportional hazards model. I think that's one of its weaknesses, that accelerated life models are in many ways more appealing because of their quite direct physical interpretation".

The AFT model (2) was developed in parallel to the GLM; and while it goes part of the way to addressing the specialised modelling needed for time-to-event outcomes, more general formulations are now possible. Following (1), we can specify a distributional regression (or Generalized Additive Models for Location, Scale and Shape, GAMLSS) model (Stasinopoulos et al. 2023):

$$\begin{aligned} t_i &\overset{\text{ind}}{\sim} \mathcal{D}(\theta_{i1}, \ldots, \theta_{iK}) && \text{for } i = 1, \ldots, n \\ g_k(\theta_{ik}) &= \boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta}_k && k = 1, \ldots, K \end{aligned} \tag{4}$$

where $\mathcal{D}(\theta_1, \ldots, \theta_K)$ is a K-parametric distribution with support on the postive real line; the $g_k(\cdot)$ are appropriate link functions; θ_1 (or μ) is a location parameter; $\theta_2, \ldots, \theta_K$ are shape parameters; and right censoring can be accommodated in the likelihood. The main advantages of the distributional regression model (4), as implemented in the **R** package **gamlss**, over the AFT model are: the large number of distributions available for modelling; the ability to model not just the location of the time distribution, but also its shape; and the availability of complex additive terms (e.g. smoothing splines, random effects, spatial effects) in the linear predictors. In the case of heavy right censoring, parametric models cannot reasonably be expected to do a good job of estimating the central tendency when the central portion and upper tail of the distribution are unobserved. In this case it makes sense to model the observed times, i.e. the left tail. Quantile regression, another member of the distributional regression family, is useful in this context, in which we model the lower quantiles of the survival time distribution using *censored quantile regression* (Koenker 2008).

4 Application

We will illustrate the alternative models discussed above with an observational dataset of head and neck cancer patients. Survival outcomes and several risk factors are observed.

5 Discussion

Statistical methods which are entrenched as the standard for analysis may not necessarily be based on conceptually simple abstractions of the data generating mechanism, and may have gained acceptance due to their mathematical or computational convenience. While odds ratios and hazard ratios qualitatively deliver the same information as more intuitive quantities such as relative risks and effects on the mean or median survival time, quantitatively they are not well interpreted.

Survival data is structurally different from non-temporal outcomes, because of the temporal nature of the outcome and possibly the covariate(s). Regression modelling may be accomplished on different scales:

- hazard function $h(t)$: PH (Cox) regression and its many variants
- survival function $S(t)$: generalized survival models (Liu et al 2018). (These are more recent and less well-known.)
- time t: AFT regression, distributional regression (GAMLSS, quantile regression)

There is a very rich body of survival modelling based on the hazard function. While the hazard function is not observable and perhaps less well understood than the survival function, it is informative of the course of a disease (when estimated). PH regression has the ability to incorporate the important feature of time-varying covariates and time-varying coefficients. It is difficult to overstate the pervasive nature of the hazard function and proportional hazards concept in survival analysis; and almost impossible to find a discussion of time-to-event data without mention of the hazard function. The AFT model, on the other hand, is the sadly neglected "poor relation" of survival analysis. And yet it was ahead of its time, foreshadowing the development of the rich framework of GAMLSS models, which have superseded it.

We urge applied researchers to be mindful of the interpretability of their analyses, and as far as possible to "model what you measure".

References

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187 – 202.

Donoghoe, M. W. and Marschner, I. C. (2018). logbin: an R package for relative risk regression using the log-binomial model. *Journal of Statistical Software*, **86**, 1 – 22.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Knol M.J., Duijnhoven R.G., Grobbee D.E., Moons K.G. and Groenwold R.H. (2011). Potential misinterpretation of treatment effects due to use of odds ratios and logistic regression in randomized controlled trials. *PLoS One*, **6**(6):e21248.

Koenker, R. (2008). Censored Quantile Regression Redux. *Journal of Statistical Software*, 27(6), 1 – 25. https://doi.org/10.18637/jss.v027.i06

Liu, X. R., Pawitan, Y. and Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5), 1531 – 1546.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370 – 384.

Reid, N. (1994). A conversation with Sir David Cox. *Statistical Science*, 9(3), 439 – 455.

Stasinopoulos, M.D., Kneib, T., Klein, N., Mayr, A. and Heller, G.Z. (2023). *Generalized Additive Models for Location, Scale and Shape: A Distributional Regression Approach, with Applications*. Cambridge University Press.

Van Noorden, R., Maher, B. and Nuzzo, R. (2014). The top 100 papers. *Nature*, 514(7524), 550.

Modeling extremal streamflow using deep learning approximations and a flexible spatial process

Reetam Majumder[1], Brian Reich[1], Banjamin Shaby[2]

[1] North Carolina State University, USA
[2] Colorado State University, USA

E-mail for correspondence: bjreich@ncsu.edu

Abstract: Quantifying changes in the probability and magnitude of extreme flooding events is key to mitigating their impacts. While hydrodynamic data are inherently spatially dependent, traditional spatial models such as Gaussian processes are poorly suited for modeling extreme events. Spatial extreme value models with more realistic tail dependence characteristics are under active development. They are theoretically justified, but give intractable likelihoods, making computation challenging for small datasets and prohibitive for continental-scale studies. We propose a process mixture model (PMM) which specifies spatial dependence in extreme values as a convex combination of a Gaussian process and a max-stable process, yielding desirable tail dependence properties but intractable likelihoods. To address this, we employ a unique computational strategy where a feed-forward neural network is embedded in a density regression model to approximate the conditional distribution at one spatial location given a set of neighbors. We then use this univariate density function to approximate the joint likelihood for all locations by way of a Vecchia (1988) approximation. The PMM is used to analyze changes in annual maximum streamflow within the US over the last 50 years, and is able to detect areas which show increases in extreme streamflow over time.

Keywords: Gaussian process; Max-stable process; Neural networks; Spatial extremes; Vecchia approximation.

1 Introduction

The Intergovernmental Panel on Climate Change released its Sixth Assessment in 2021 and projected an increased frequency of hydroclimatic

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

extremes. In addition to changes in the mean of climate variables, the impact of climate change is more severe with changes in the frequency and magnitude of hydroclimatic extremes. For example, Hirsch and Ryberg (2012) found a significant change in annual maximum streamflow (a key measure of flood risk) at 48 of 200 US Geological Survey (USGS) gauges and spatial clustering in the direction and magnitude of the changes. As a result, there is a need to account for spatial and temporal variability (i.e., nonstationarity) in flood frequency patterns when assessing current and future risk

A spatial extreme value analysis (EVA) models exceedances or pointwise maxima as a stochastic process over space. Modeling spatial dependence allows for predictions at ungauged locations and the estimation of the joint probability of extremes at multiple locations. It also facilitates the borrowing of information across locations to estimate the marginal distribution at each location, which is particularly useful for EVA where data are sparse and low-probability events are of interest, and gives valid statistical inference for model parameters by properly accounting for spatial dependence. We focus on the modeling of block maxima of streamflow with the help of the max-stable process (MSP; De Haan et al., 2006). MSPs are a limiting class of models for spatial extremes, featuring strong forms of tail dependence. In practice, MSPs pose two challenges. First, the analytic forms of (censored) MSP densities are computationally intractable for all but a small number of spatial locations A second challenge posed by MSPs is that they are restrictive in the class of dependence types they can incorporate. Environmental data often has weakening spatial dependence with increasing levels of extremeness; however, MSPs are unable to accommodate this behavior. A more general approach was taken in Huser and Wadsworth (2019) which combined a Pareto random variable with a Gaussian process (GP) resulting in a hybrid model with perfect dependence and asymptotic independence, indexed similarly by a mixing parameter. This flexible model can establish asymptotic dependence or asymptotic independence from the data without needing a prior assumption. A limitation of this model is that the Pareto random variable is shared by the spatial locations, inducing dependence between distant sites. This might be unrealistic for an analysis over a large spatial domain.

In this paper, we propose a spatial EVA model and an associated computational algorithm to address the aforementioned limitations of the MSP and related approaches. The EVA model is specified as a convex combination of an MSP and a GP for residual dependency, and has Generalized Extreme Value (GEV) margins with spatiotemporally varying coefficients (STVC). We refer to it as the process mixture model (PMM). From a modeling perspective, the mixture of the two spatial processes allows asymptotic dependence or independence for locations separated by distance h, with asymptotic independence as $h \to \infty$. Furthermore, the STVC can account for temporal nonstationarity which is key for large-scale climate studies.

This flexibility comes at a computational cost: the model has hundreds of parameters and even bivariate PDFs do not have a closed form. Therefore we develop a new computational algorithm that uses a feed-forward neural network (FFNN) embedded in a density regression model of Xu and Reich (2023) to approximate the conditional distribution at one spatial location given a set of neighbors. Following this, the univariate density functions are used to approximate the joint likelihood for all locations by means of a Vecchia approximation (Vecchia, 1988). Parameter estimation is carried out using Markov Chain Monte Carlo (MCMC). This computational framework is quite general. Unlike many of the approaches mentioned above, it can be applied to virtually any spatial process (e.g., GP, MSP, and mixtures), can accommodate high-dimensional STVC margins, as well as missing and censored data. We use the PMM to analyze changes in annual maximum streamflow within the US over the past 50 years.

2 A Process Mixture Model for Spatial Extremes

Let $Y(\boldsymbol{s})$ be the extreme observation at spatial location $\boldsymbol{s}$. We assume a potentially different marginal distribution for each spatial location $\boldsymbol{s}$ and denote $F_{\boldsymbol{s}}$ as the marginal cumulative distribution function (CDF) for site $\boldsymbol{s}$. For example, we assume that $F_{\boldsymbol{s}}$ is the generalized extreme value (GEV) distribution with location $\mu(\boldsymbol{s})$, scale $\sigma(\boldsymbol{s})$ and shape $\xi(\boldsymbol{s})$ so that marginally

$$Y(\boldsymbol{s}) \sim \text{GEV}\{\mu(\boldsymbol{s}), \sigma(\boldsymbol{s}), \xi(\boldsymbol{s})\}.$$

Then the transformed variables

$$U(\boldsymbol{s}) = F_{\boldsymbol{s}}\{Y(\boldsymbol{s})\} \tag{1}$$

share common uniform marginal distributions across the spatial domain. This transformation separates residual spatial dependence in $U(\boldsymbol{s})$ from the spatial dependence induced by spatial variation in the GEV parameters, which we model using GP priors over $\boldsymbol{s}$.
We define our spatial dependence model on $U(\boldsymbol{s})$ by taking $U(\boldsymbol{s}) = G\{V(\boldsymbol{s})\}$, such that

$$V(\boldsymbol{s}) = \delta g_R\{R(\boldsymbol{s})\} + (1-\delta) g_W\{W(\boldsymbol{s})\}, \tag{2}$$

where $R(\boldsymbol{s})$ is a max-stable process (MSP), $W(\boldsymbol{s})$ is a Gaussian process (GP), g_R and g_W are transformations that guarantee $g_R\{R(\boldsymbol{s})\}$ and $g_W\{W(\boldsymbol{s})\}$ follow the standard exponential distribution, and $\delta \in [0,1]$ is the weight parameter to control relative contribution of the two spatial processes. Mixing the asymptotically dependent MSP with the asymptotically independent GP provides a rich model for spatial dependence. This generalizes Huser and Wadsworth (2019), who assumed a standard Pareto random variable R common to all locations, by replacing it with an MSP. Since (2) mixes two processes, we refer to it as the process mixture model.

By construction, $V(\boldsymbol{s})$ marginally follows the two-parameter hypoexponential distribution with CDF

$$G(v) = 1 - \frac{1-\delta}{1-2\delta} e^{-\frac{1}{(1-\delta)}v} + \frac{\delta}{1-2\delta} e^{-\frac{1}{\delta}v}. \tag{3}$$

Without loss of generality, we assume that $R(\boldsymbol{s})$ has a marginal GEV(1,1,1) distribution and $W(\boldsymbol{s})$ has a standard normal marginal distribution. In this case, the transformations are $g_R(r) = -\log\{1-\exp(-1/r)\}$ and $g_W(w) = -\log\{1-\Phi(w)\}$ for standard normal CDF $\Phi(w)$. Although other options are possible, we model the correlation of the GP $W(\boldsymbol{s})$ using the isotropic powered-exponential correlation function $\text{Cor}\{W(\boldsymbol{s}_1, \boldsymbol{s}_2)\} = \exp\{-(h/\rho_W)^{\alpha_W}\}$ with distance $h = ||\boldsymbol{s}_1 - \boldsymbol{s}_2||$, smoothness $\alpha_W \in (0, 2)$, and range $\rho_W > 0$. The MSP $R(\boldsymbol{s})$ is assumed to have isotropic Brown-Resnick spatial dependence defined by the variogram $\gamma(h) = (h/\rho_R)^{\alpha_R}$ for smoothness $\alpha_R \in (0, 2)$ and range $\rho_R > 0$.
Extremal spatial dependence of the process at locations $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ is often measured using the conditional exceedance probability

$$\chi_u(\boldsymbol{s}_1, \boldsymbol{s}_2) := \text{Prob}\{U(\boldsymbol{s}_1) > u | U(\boldsymbol{s}_2) > u\}, \tag{4}$$

where $u \in (0, 1)$ is a threshold. The random variables $U(\boldsymbol{s}_1)$ and $U(\boldsymbol{s}_2)$ are defined as asymptotically dependent if the limit

$$\chi(\boldsymbol{s}_1, \boldsymbol{s}_2) = \lim_{u \to 1} \chi_u(\boldsymbol{s}_1, \boldsymbol{s}_2) \tag{5}$$

is positive and independent if $\chi(\boldsymbol{s}_1, \boldsymbol{s}_2) = 0$. Since we assume both $R(\boldsymbol{s})$ and $W(\boldsymbol{s})$ are isotropic processes, we simply write $\chi_u(h)$ and $\chi(h)$ as a function of the distance between locations.

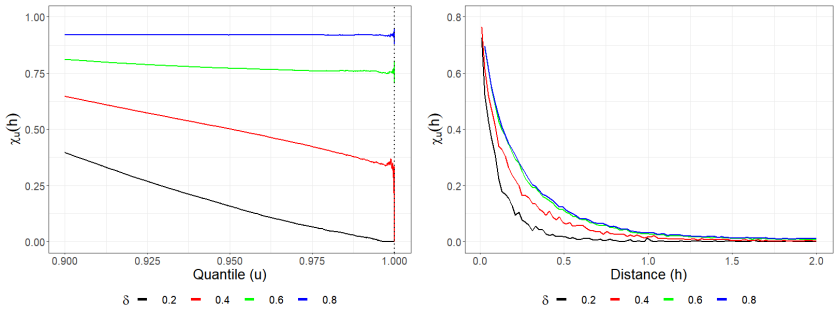


FIGURE 1. **Behavior of the empirical conditional exceedance:** Approximate $\chi_u(h)$ for the process mixture model plotted as a function of threshold u, distance h, and asymptotic dependence parameter δ. Smoothness parameters $\alpha_W = \alpha_R = 1$, GP range $\rho_W = 0.5$, and MSP range $\rho_R = 0.1$ are fixed for both plots. $\chi_u(h)$ as a function of u and δ at distance $h = 0.8$ (left) and as a function of h and δ, for threshold $u = 0.99$ (right).

Figure 1 plots Monte Carlo approximations of $\chi_u(h)$ as a function of u and h for the process mixture model. As a function of the threshold u, the limit tends to zero for $\delta < 0.5$ and to non-zero values for $\delta > 0.5$; for small h the MSP $R(\boldsymbol{s})$ is approximately the same for both sites and thus the univariate R result of Huser and Wadsworth (2019) that the process is asymptotically dependent if and only if $\delta > 0.5$ emerges. As the distance h increases, $\chi_u(h)$ converges to zero for all δ because both $R(\boldsymbol{s})$ and $W(\boldsymbol{s})$ have diminishing spatial dependence for long distances. We note that $\chi_u(h)$ does not converge to zero for large h under the common R model of Huser and Wadsworth (2019), which is unrealistic for studies on a large spatial domain.

3 Deep Learning Vecchia Approximation for the Process Mixture Model

Fitting the process mixture model introduced in Section 2 poses computation challenges, especially for large datasets. The joint distribution for the GP $W(\boldsymbol{s})$ is available in closed form but is cumbersome for large datasets; the joint distribution of the MSP $R(\boldsymbol{s})$ is available only for a moderate number of spatial locations, and the joint distribution of the mixture model is more complicated that either of its components. Below we develop a surrogate likelihood based on a Vecchia decomposition (Vecchia, 1988) and deep learning density regression.

Assume the process is observed at n locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$. Partition the parameters into those that affect the marginal distributions, denoted $\boldsymbol{\theta}_1$, and those that affect the spatial dependence, denoted $\boldsymbol{\theta}_2$. For the model in Section 2, $\boldsymbol{\theta}_1$ includes the GEV parameters $\boldsymbol{\theta}_1 = \{\mu(\boldsymbol{s}_i), \sigma(\boldsymbol{s}_i), \xi(\boldsymbol{s}_i); i = 1, ..., n\}$ and $\boldsymbol{\theta}_2 = \{\delta, \rho_R, \alpha_R, \rho_W, \alpha_W\}$. Let $Y(\boldsymbol{s}_i) \equiv Y_i$ and $U_i = F(Y_i; \boldsymbol{\theta}_1)$ be the transformation of the response so that the distribution of $U_i \in [0, 1]$ does not depend on $\boldsymbol{\theta}_1$. We approximate the spatial model on this scale and use the standard change of variables formula to define the joint likelihood on the original scale

$$f_y(y_1, ..., y_n; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f_u(u_1, ..., u_n; \boldsymbol{\theta}_2) \prod_{i=1}^{n} \left| \frac{dF(y_i; \boldsymbol{\theta}_1)}{dy_i} \right|. \tag{6}$$

We approximate the joint likelihood in (6) using a Vecchia approximation (Vecchia, 1988),

$$f_u(u_1, ..., u_n; \boldsymbol{\theta}_2) = \prod_{i=1}^{n} f(u_i | \boldsymbol{\theta}_2, u_1, ..., u_{i-1}) \approx \prod_{i=1}^{n} f_i(u_i | \boldsymbol{\theta}_2, u_{(i)}) \tag{7}$$

for $u_{(i)} = \{u_j; j \in \mathcal{N}_i\}$ and neighboring set $\mathcal{N}_i \subseteq \{1, ..., i-1\}$, e.g., the k locations in $\mathcal{N}_i$ that are closest to $\boldsymbol{s}_i$. Here, we use the notation that

the collection of variables z_i over the neighboring set is denoted $z_{(i)} = \{z_j; j \in \mathcal{N}_i\}$. Of course, not all locations that are dependent with location i need be included in $\mathcal{N}_i$ because distant observations may be approximately independent after conditioning on more local observations.

The conditional distributions for the process mixture model do not have closed-form expressions. We approximate the n conditional density functions separately, each using the density regression model introduced by Xu and Reich (2023):

$$f(u_i|\boldsymbol{x}_i, \mathcal{W}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i, \mathcal{W}) B_k(u_i) \tag{8}$$

where $\boldsymbol{x}_i = (\boldsymbol{\theta}_2, u_{(i)})$, $\pi_k(\boldsymbol{x}, \mathcal{W}) \geq 0$ are probability weights with $\sum_{k=1}^{K} \pi_k(\boldsymbol{x}) = 1$ that depend on the parameters $\mathcal{W}$ and $B_k(u) \geq 0$ are M-spline basis functions that, by definition, satisfy $\int B_k(u)du = 1$ for all k. By increasing the number of basis functions K and appropriately selecting the weights $\pi_k(\boldsymbol{x})$, this mixture distribution can approximate any continuous density function. The weights are modeled using a feed-forward neural network (FFNN) with H hidden layers with L_h neurons in hidden layer h and multinomial logistic weights. The model is

$$\begin{aligned} \pi_k(\boldsymbol{x}, \mathcal{W}) &= \frac{\exp\{\gamma_{Hk}(\boldsymbol{x}, \mathcal{W})\}}{\sum_{l=1}^{K} \exp\{\gamma_{Hl}(\boldsymbol{x}, \mathcal{W})\}} \\ \gamma_{hk}(\boldsymbol{x}) &= W_{hk0} + \sum_{j=1}^{L_h} W_{hkj} \psi\left\{\gamma_{h-1,j}(\boldsymbol{x}, \mathcal{W})\right\} \quad \text{for} \quad h \in \{1, ..., H\} \\ \gamma_{0k}(\boldsymbol{x}, \mathcal{W}) &= W_{0k0} + \sum_{j=1}^{p} W_{0kj} v_j \end{aligned} \tag{9}$$

where $\boldsymbol{x} = (x_1, ..., x_p)$, $\mathcal{W} = \{W_{hkj}\}$ are the parameters to be estimated and ψ is the activation function. Building on the universal approximation theorem for FFNNs, Xu and Reich (2023) argue that (9) with $H = 1$ and large K and L_1 can approximate any conditional density function that is smooth in its arguments.

Within this framework, approximating the conditional distributions is equivalent to estimating the weights $\mathcal{W}$. Unlike a typical statistical learning problem, observational data are not used to estimate $\mathcal{W}$. Rather, the weights are learned from training data generated from the process mixture model with parameters $\boldsymbol{\theta}_2 \sim p^*$, and then a realization from the process over sites i and $\mathcal{N}_i$ from the model conditioned on $\boldsymbol{\theta}_2$. Specifically, we generate data at the observed spatial location with the same neighbor sets to be used in the analysis. We select the design distribution p^* with support covering the range of plausible values for $\boldsymbol{\theta}_2$. Given these values, we generate $U(\boldsymbol{s})$ at $\boldsymbol{s} \in \{\boldsymbol{s}_i, \boldsymbol{s}_{(i)}\}$. The feature set $\boldsymbol{x}_i$ for modeling u_i at location i

thus contains the spatial parameters $\boldsymbol{\theta}_2$, process values at the neighboring locations $U(\boldsymbol{s}_{(i)})$, as well as the spatial configuration of the neighboring set, $\{(\boldsymbol{s}_{(i)} - \boldsymbol{s}_i)\} \equiv \{(\boldsymbol{s}_j - \boldsymbol{s}_i); j \in \mathcal{N}_i\}$, where the sites in $\mathcal{N}_i$ are ordered by the distances to $\boldsymbol{s}_i$. Algorithm 1 outlines the procedure. Therefore, all that is required to build the approximation is the ability to generate small datasets from the model. The size of the training data is effectively unlimited, meaning the approximation can be arbitrarily accurate. Once the weights have been learned, applying the FFNN to the approximate likelihood is straightforward, and the Vecchia approximation ensures that the computational burden increases linearly in the number of spatial locations.

Algorithm 1 Local SPQR approximation

Require: Locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ with neighbor locations $\boldsymbol{s}_{(1)}, \ldots, \boldsymbol{s}_{(n)}$
Require: Design distribution p^*, training sample size N
  $i \leftarrow 2$
  while $i \leq n$ **do**
    $k \leftarrow 1$
    while $k \leq N$ **do**
      Draw values of $\boldsymbol{\theta}_{2k} \sim p^*$
      Generate $U_k(\boldsymbol{s})$ at $\boldsymbol{s} \in \{\boldsymbol{s}_i, \boldsymbol{s}_{(i)}\}$ given $\boldsymbol{\theta}_{2k}$ using (2)
      Define features $\boldsymbol{x}_{ik} = (\boldsymbol{\theta}_{2k}, u_{(i)k})$, where $u_{(i)k} = \{U_k(\boldsymbol{s}); \boldsymbol{s} \in \boldsymbol{s}_{(i)}\}$
      $k \leftarrow k + 1$
    end while
    solve $\hat{\mathcal{W}}_i \leftarrow \underset{\mathcal{W}}{\operatorname{argmax}} \prod_{k=1}^{N} f(u_{ik}|\boldsymbol{x}_{ik}, \mathcal{W})$ for $f(u|\boldsymbol{x}, \mathcal{W})$ defined in (8) using `SPQR`
    $i \leftarrow i + 1$
  end while

Given the approximate model in (6) for f_y with an SPQR approximation for f_u, a Bayesian analysis using MCMC methods is straightforward. We use Metropolis updates for both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. For a spatially-varying coefficient model with local GEV coefficients for location i, we update parameters $\{\mu(\boldsymbol{s}_i), \sigma(\boldsymbol{s}_i), \xi(\boldsymbol{s}_i)\}$ as a block sequentially by site, and exploit the Vecchia approximation to use only terms in the likelihood corresponding to sites $j \ni i \in \mathcal{N}_j$, i.e., sites for which site i is included in the neighboring set. All Metropolis updates are tuned to give acceptance probability 0.4, and convergence is diagnosed based on the visual inspection of the trace plots. Additional computational details are given for specific analyses below, and MCMC code is available at `https://github.com/reetamm/SPQR-for-spatial-extremes`.

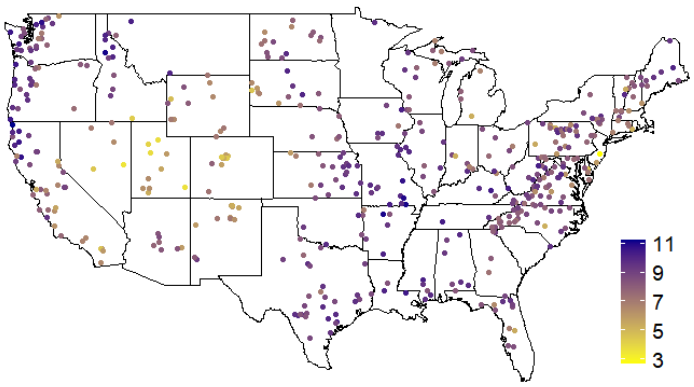


FIGURE 2. **HCDN annual maxima**: Sample 0.9 quantile of the log annual streamflow (cubic ft/sec) maxima $Y_t(\boldsymbol{s})$ at each of the 489 gauges.

4 Analysis of Extreme Streamflow in the US

We apply the methods to model streamflow data from USGS Hydro-Climatic Data Network (Lins, 2012), which is designed to monitor streamflow in locations that are unaffected by human activities. We analyze data from 1972–2021 at 489 stations across the US with complete data. Our goal is to identify regions where the distribution of extreme streamflow is changing over time. For each year and station, we take as the response $Y_t(\boldsymbol{s})$ the logarithm of the annual maximum of daily streamflows. Figure 2 plots the sample 0.9 quantile of the observations at each station.

For the marginals at each location, we assume GEV distributions with spatio-temporally varying parameters,

$$Y_t(\boldsymbol{s}) \sim \text{GEV}\left[\mu_0(\boldsymbol{s}) + \mu_1(\boldsymbol{s})X_t, \exp\{\sigma(\boldsymbol{s})\}, \xi(\boldsymbol{s})\right], \tag{10}$$

where $X_t = (\text{year}_t - 1996.5)/10$ for $\text{year}_t = 1972 + t - 1$. This parameterization attempts to capture changes in the location parameter in the past 50 years due to changing climate; positive values of $\mu_1(\boldsymbol{s})$ would suggest an increase in the magnitude of the annual extremal streamflow. The marginal GEV parameters for each location are assigned GP priors with nugget effects which allows local heterogeneity, and the hyperpriors are uninformative. Once the local SPQR models have been fitted, we run two MCMC chains for 20,000 iterations each, with two different starting values of δ. The first 5000 iterations from each chain are discarded as burn-in

The posterior mean (standard deviations) of the spatial dependence parameter is $\hat{\delta} = 0.47\,(0.02)$, which puts the process in the asymptotic independence regime with high probability, but the posterior mean is near

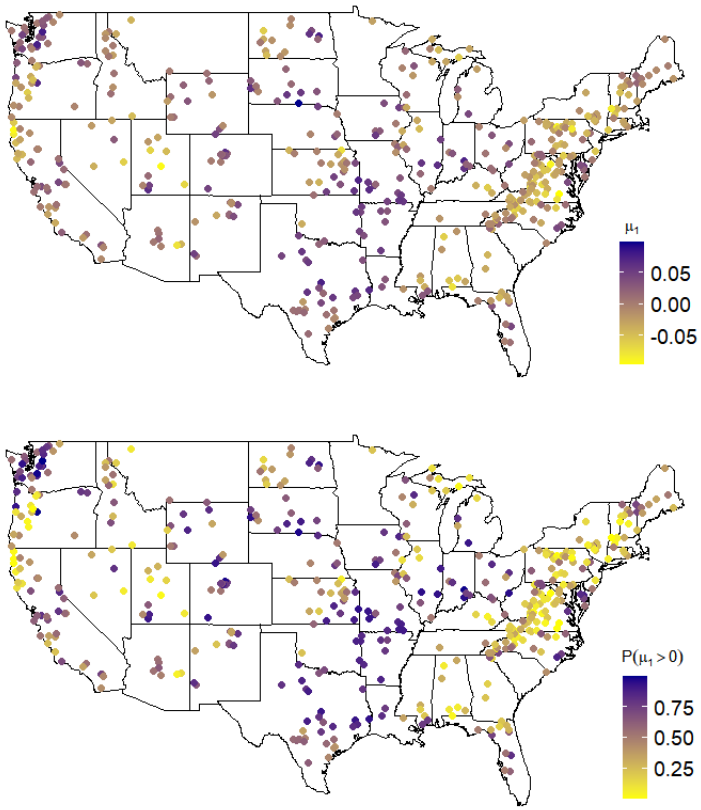


FIGURE 3. **HCDN GEV parameter estimates**: Posterior mean of $\mu_1(\boldsymbol{s})$ (left) and posterior probability that $\mu_1(\boldsymbol{s})$ is positive (right).

the 0.5 boundary. Figures 3 plots the estimated slope $\mu_1(\boldsymbol{s})$ of the location parameters with respect to time (left) and $Pr[\mu_1(\boldsymbol{s}) > 0]$ (right). Positive slope estimates indicate an increase in extreme streamflow over time. The majority of the positive slope parameters are concentrated in central and south USA. On the east coast, the stretch between Delaware and North Carolina contain several areas with positive slopes. Similarly on the west coast, Washington has a high concentration of positive slope parameter estimates, as does California, with higher values inland and away from the coast. The states of Wyoming, Colorado, and New Mexico are also of interest since these have relatively low 0.9 quantile values in Figure 2, suggesting that extreme streamflow is starting to have large impacts in these areas.

Acknowledgments: This work was supported by grants from the Southeast National Synthesis Wildfire and the United States Geological Survey's National Climate Adaptation Science Center (G21AC10045), the National Science Foundation (CBET2151651, DMS2152887, and DMS2001433) and the National Institutes of Health (R01ES031651-01).

References

De Haan, L., Ferreira, A. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer.

Hirsch, R. M. and Ryberg, K. R. (2012). Has the magnitude of floods across the USA changed with global CO2 levels? *Hydrological Sciences Journal*, **57**, 1–9.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, **114**, 434–444.

Lins, H. F. (2012). USGS hydro-climatic data network 2009 (HCDN-2009). *US Geological Survey Fact Sheet*, 3047.

Xu, S. G. and Reich, B. J. (2023). Bayesian non-parametric quantile process regression and estimation of marginal quantile effects. *Biometrics*, **79**, 151-164.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B*, **50**, 297–312.

On Covid, dynamic models and inferring smooth functions

Simon Wood, Thea Abou Jawad, Lauren Corcoran, Beth Flood, Danshu Hu

[1] University of Edinburgh, UK

E-mail for correspondence: simon.wood@ed.ac.uk

Abstract: Epidemic dynamic models played a large part in the scientific management of the UK Covid-19 epidemic. However, the models were not validated for prediction, neglected individual heterogeneity of profound dynamic importance, ignored the nosocomial transmission that accounted for a high proportion of serious Covid cases, and employed restrictive parametric assumptions that introduced substantial artefacts when the models were used for statistical inference. In particular, over-restrictive assumptions about how contact rates changed over time entirely drove inferences about the timing of infection waves relative to lockdowns. In fact the dynamic models could be re-formulated using smooth functions to represent such contact rates. A simple empirical Bayes approach to inference is then possible, in which an extended Fellner-Schall approach to smoothing parameter estimation is employed. The framework allows inference with complex dynamic models, provided that the first derivatives of the model can be obtained. The methods are used to investigate the timing of incidence (new infections per day) relative to lockdowns in several European countries.

Keywords: smoothing parameters; Fellner-Schall; epidemic model.

1 Introduction

> ... a substantial number of people still do not feel sufficiently personally threatened; it could be that they are reassured by the low death rate in their demographic group... the perceived level of personal threat needs to be increased among those who are complacent, using hard hitting emotional messaging.

This quote is extracted from the 22 March 2020 recommendations on Covid-19 from the UK government advisory Scientific Pandemic Influenza Group

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

on Behaviour (SPI-B). It's an unusual approach to medical risk communication, but one that dominated the UK government management of the pandemic until late 2021. The demographic risk profile that perhaps formed the basis for being 'complacent' is shown in Figure 1.

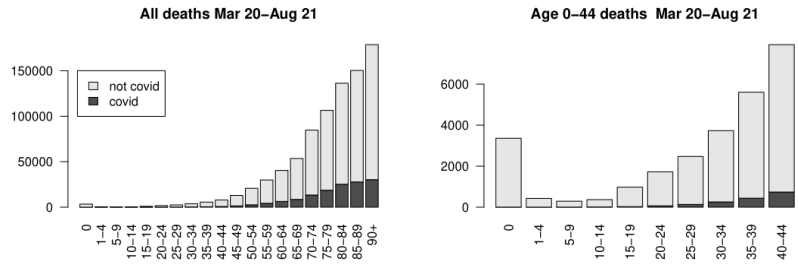


FIGURE 1. UK deaths with and without Covid-19 from March 2020 until august 2021, by age group. Data from the UK Office for National Statistics.

A major part of the scientific justification for the Covid response came from epidemic dynamic modelling. Some of this modelling followed the traditional physics model approach, in which parameters are estimated separately from the model, and the model then makes predictions given these parameter estimates. Other modelling used modern Bayesian methods to update parameters, given data on observed epidemic dynamics, albeit usually relying also on some externally estimated parameters. None of the models were seriously validated for prediction, in the way that one would expect for a weather of climate model, for example.

Soon after the first wave, there was some evidence that lack of validation might matter. The leading modelling group in the UK, at Imperial College, had published model predictions of Covid deaths under various social distancing measures, for a variety of countries (Walker et al., 2020). For Sweden they predicted about 35 thousand first wave deaths without full lockdown, but with 'social distancing of the whole population' – the scenario closest to what Sweden actually did. Sweden actually had about 6 thousand first wave Covid deaths.

Perhaps such a mismatch should not be surprising. Prediction with an epidemic model not validated for prediction amounts to long range extrapolation with a non-linear model. But there were also extreme simplifications in the models that reduced the likelihood of accuracy. The most obvious is the extent to which people were modelled as passive clones, rather than variable individuals with agency. This meant that the consequences of spontaneous behavioural change in response to perceived risk were not captured by the models. That is perhaps inevitable given the difficulty of modelling such responses. What was not inevitable was the neglect of most person-to-person variability in susceptibility and contact rates. The sub-

stantial impact of such heterogeneity on epidemic dynamics and size has been well understood mathematically since at least Novozhilov (2008). It is also easy to understand: more susceptible and connected individuals are infected first, so that transmission rates decline much faster than would be implied by the depletion of a susceptible population of clones (see McKeigue and Wood, 2022, for a concise introduction to the maths). That this effect was significant for Covid was demonstrated by Gomes and co-workers early in 2020 (eventually published as Gomes et al., 2022). Realistic levels of heterogeneity can easily reduce epidemic size by a half.

The models also omitted hospital acquired infection, despite Wang et al. (Feb. 2020) reporting a suspected 41% nosocomial infection rate in Wuhan as a key finding, a feature that would be repeated in the first wave in Lombardy in Italy where Boccia et al. (2020) note that "SARS-CoV-2 became largely a nosocomial infection". Later analysis showed that within Scotland, for example, the proportion of serious Covid cases that were hospital acquired peaked at around 60% (McKeigue et al., 2021).

One might hope to be on firmer ground when considering the inferences made with epidemic models that had their parameters fitted or updated using modern statistical methods. But here another problem was apparent. To achieve computational tractability, simple parametric formulations were employed, which had the potential to introduce serious artefacts. Around the issue of lockdown's impact on transmission rates a particularly insidious problem occurred. If lockdown reduces transmission rates then at lockdown a sort of partitioning of the population occurs. There is a low transmission locked down sub-population and a higher transmission 'key-worker' sub-population. The epidemiologist's key measure of transmission rates is R, the average total number of new infections caused by each existing infection. The average here is over infections, not people, and that has consequences. In particular, immediately after lockdown, most infections are in the locked down sub-population, so average R is low. Over time the proportion of infections in the key-worker sub-population must grow as the growth rate is higher there, and that means that R must grow too, since an increasing proportion of *infections* are in the higher transmission sub-population. Any model that does not allow for this dip and recovery in R will obviously suffer from artefacts.

In fact the most influential analyses, suggesting that lockdowns were essential, used models that could not capture this dip and recovery, because of using very simple parametric models for how contact rates varied over time. Replacing these over-simplified parametric representations with more flexible splines changed the conclusions entirely. Such models turn out to be quite easy work with, by adapting the extended Fellner-Schall approach to smoothness selection (Wood and Fasiolo, 2017).

2 Extended Fellner Schall methods

Fellner (1986) introduced a straightforward update formula for variance components of simple independent random effects in a linear model, which Schall (1991) generalized to the GLM setting. Given the long established duality between spline type smooths and random effects (e.g. Kimeldorf and Wahba, 1970), the method can be used to estimate smoothing parameters of smooth functions, when these are controlled by a single smoothing parameter. The original derivations are slightly mysterious, as they involve an equation in which the same variance term appears on both sides, and it is then decided to substitute an estimate on the right hand side, while treating the left hand side version as its update. However, it can be shown that: the update is indeed in a direction that increases the (Laplace approximate) marginal likelihood of the model; it tends to make larger update steps than the EM algorithm; it can be generalized to spline like smooths with multiple smoothing parameters, such as adaptive or tensor product splines, and to model likelihoods beyond GLMs (Wood and Fasiolo, 2017). Generically, consider a model for n data, $\mathbf{y}$, with coefficient vector β and log likelihood $l(\beta)$, to be estimated by penalised likelihood maximisation

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ l(\beta) - \frac{1}{2}\beta^{\mathrm{T}}\mathbf{S}_\lambda\beta, \text{ where } \mathbf{S}_\lambda = \sum_j \lambda_j \mathbf{S}_j, \tag{1}$$

the $\mathbf{S}_j$ are known positive semi-definite matrices and the λ_j are positive unknown smoothing parameters. $\hat{\beta}$ is a posterior mode, if the penalty is induced by an improper Gaussian *smoothing prior* $\beta \sim N(\mathbf{0}, \mathbf{S}_\lambda^-)$. In that case, provided $\dim(\beta) = o(n^{1/3})$, we also have the $n \to \infty$ result

$$\beta|\mathbf{y} \sim N(\hat{\beta}, \mathbf{H}_\lambda^{-1})$$

where $\mathbf{H}_\lambda$ is the negative Hessian of the penalized log likelihood. Denoting this approximate posterior as $\pi_g(\beta|\mathbf{y})$ then the log *Laplace Approximate Marginal Likelihood* (LAML) is

$$\begin{aligned} l_r(\lambda) \equiv \log \pi_\lambda(\mathbf{y}) &= \log\left\{\pi(\mathbf{y},\hat{\beta})/\pi_g(\hat{\beta}|\mathbf{y})\right\} \\ &= l(\hat{\beta}) - \hat{\beta}^{\mathrm{T}}\mathbf{S}_\lambda\hat{\beta}/2 + \log|\mathbf{S}_\lambda|_+/2 - \log|\mathbf{H}_\lambda|/2 + \text{const.} \end{aligned}$$

where $|\cdot|_+$ denotes a product of strictly positive eigenvalues. Differentiating w.r.t. λ_j we have

$$\frac{\partial l_r}{\partial \lambda_j} = -a + b - c$$

where $a = \hat{\beta}^{\mathrm{T}}\mathbf{S}_j\hat{\beta}$, $b = \operatorname{tr}(\mathbf{S}_\lambda^-\mathbf{S}_j) - \operatorname{tr}(\mathbf{H}_\lambda^{-1}\mathbf{S}_j)$, $c = \operatorname{tr}(\mathbf{H}_\lambda^{-1}\partial\mathbf{H}/\partial\lambda_j)$ and $\mathbf{H}$ is the negative Hessian of the unpenalized log likelihood at $\hat{\beta}$. The terms a and b are positive. $c = 0$ in the Gaussian case and for several other GLM

distribution-link combinations. Otherwise c is of either sign, but typically small. The standard generalized Fellner-Schall update therefore neglects c and uses the update

$$\lambda_j^* = \lambda_j \frac{b}{a} \text{ for all } j.$$

These updates are alternated with updates of $\hat{\beta}$, given the current λ_j estimates. If $c = 0$ it can be shown that the updates always result in smoothing parameter changes in the direction of improving l_r. If c is non-zero then the process will only approximately optimize l_r. Note that the neglect of c is exactly what is done by PQL (Breslow and Clayton, 1993), for example. In practice $\hat{\beta}$ is usually found by Newton's method, which itself requires evaluation of $\mathbf{H}_\lambda$. The fact that the smoothing parameter update requires no more than was anyway needed to find $\hat{\beta}$ is then quite convenient. Note also that the expression $\text{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j)$ is purely formal, and one would obviously not compute it by forming an unstructured pseudo-inverse of $\mathbf{S}_\lambda$. For any single λ_j smooth, $\text{tr}(\mathbf{S}_\lambda^- \mathbf{S}_j) = \text{rank}(\mathbf{S}_j)/\lambda_j$, the required rank being fixed and known. Only for a smooth with multiple λ_j is an explicit pseudoinverse required, and then only for the diagonal block of $\mathbf{S}_\lambda$ corresponding to the smooth (in fact with some upfront re-parameterization the pseudoinverse can be replaced by a regular inverse).

If we do not want to neglect c then it can be computed, by applying implicit differentiation to find $\partial\hat{\beta}/\partial\lambda_j$ and then using the chain rule to compute the derivative of $\mathbf{H}$. Doing so typically requires 3rd derivatives of the log likelihood to be evaluated, and the update then becomes

$$\lambda_j^* = \begin{cases} \lambda_j(b-c)/a & c \leq 0 \\ \lambda_j b/(a+c) & c > 0. \end{cases}$$

A less implementationally tedious alternative is to start off updating all smoothing parameters at each iteration using the $c = 0$ update, but near convergence to switch to only updating one λ_j at a time, so that each update can also provided a finite difference estimate of $\partial\mathbf{H}/\partial\lambda_j$ and hence $c_j = \text{tr}(\mathbf{H}_\lambda^{-1}\partial\mathbf{H}/\partial\lambda_j)$. This c_j can be carried forward as the estimate of c for the next time λ_j is updated. As the iteration converges, so the carried forward c_j values converge to the correct values. Either the exact or finite difference scheme exactly optimizes l_r.

2.1 EFS computation for complicated models

The above discussion assumes that we are happy to evaluate $\mathbf{H}_\lambda$, but that may not always be practical. For example, in order to check and replicate Knock et al. (2020), Wood and Wit (2021) fitted a Covid epidemic dynamic model with some 7 hundred state variables to various health data streams, representing a key contact rate modifier with an adaptive smoothing spline. Simply to obtain first derivates of the model log likelihood w.r.t.

model coefficients required a system of over 65 thousand ordinary differential equations. The second derivative system required for direct evaluation of $\mathbf{H}_\lambda$ would have been entirely intractable.
The solution is to solve (1) by a quasi-Newton method, requiring only first derivatives to be computed. $\mathbf{H}_\lambda$ can then be approximated by finite differencing the first derivatives around $\hat{\beta}$, and used in a generalized Fellner-Schall update of the smoothing parameters. Note that if applied to a standard generalized additive model this approach would have the $O(np^2)$ cost of the usual methods, where $p = \dim(\beta)$. In that case derivative computations are each $O(np)$, so the $O(p)$ of these required for the finite differenced Hessian leads to $0(np^2)$ cost.

3 Inferring Covid Incidence

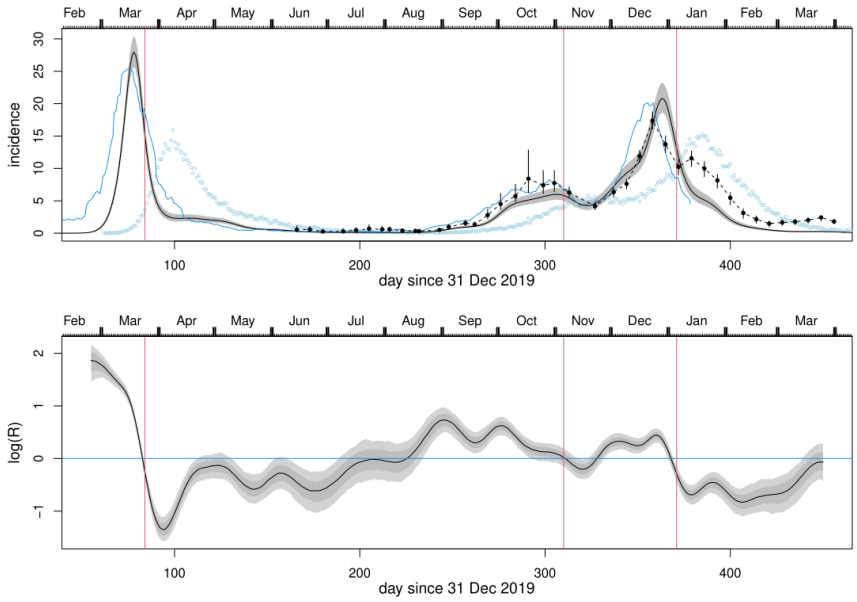


FIGURE 2. Top: Covid-19 incidence reconstructions for England (scaled to fatal incidence scale). The grey 95% confidence band is the reconstruction from the NHS hospital daily death with covid data shown as blue circles. The blue line is the REACT-2 reconstruction from symptom onset dates of a random sample of antibody positive subjects. The black dots and CIs jointed by the dashed line are ONS incidence estimates from statistical surveillance sampling. Lockdown dates are at vertical red lines. Bottom: pathogen reproductive number, R, corresponding to the grey incidence band, assuming a simple SEIR model.

The preceding approach makes it relatively straightforward to work with

non-standard models containing unknown smooth functions. A straightforward example is a deconvolution model for inferring Covid fatal incidence (new, ultimately fatal, infections per day), from data on deaths with Covid by exact day of death. For example,

$$\mathbb{E}(y_i) = \sum_{d=0}^{D_i} \exp\{f(t_i - d)\}\pi(d) \quad y_i \sim \text{negative binomial}$$

where y_i is the number of deaths with Covid on day t_i, $\pi(d)$ is the probability of a fatal disease duration of d days and $\exp\{f(t)\}$ is the number of new (fatal) infections on day t. D_i is the maximum lag considered at day t_i: it might typically be set to e.g. 80 days, but to lower values at the start of the epidemic, for statistical stability reasons and reflecting the fact that at the start, short disease durations will be seen first. The formulation in terms of log incidence ensures that incidence remains positive. Smoothness on the log scale also implies smoothness of the epidemiologist's intrinsic rate of increase parameter, r.

Figure 2 shows the results of applying this model to English National Health Service (NHS) data on daily deaths with Covid, assuming a fatal disease duration distribution $\log(d) \sim N(3.151, 0.469^2)$ based on the meta-analysis of McAloon et al. (2020) for time from infection to first symptoms, and data on over 24000 fatal cases from Pritchard et al. (2020) for onset to death. This distribution is similar to what is reported in Verity et al. (2020), Linton et al. (2020) and Wu et al. (2020). The pathogen reproductive number, R, consistent with the reconstructed incidence, assuming a simple SEIR model, is also shown (see Wood, 2021, for details). The results for the first wave are essentially the same as those obtained by this approach in early May 2020 (Wood, 2020). Subsequently, direct estimates of incidence based on statistical survey methods confirmed the pattern of incidence decline preceding lockdowns. The blue line on figure 2 shows incidence according to the REACT-2 study: antibody positive subjects, in their randomized statistical sample, where asked when their symptoms started (Ward et al., 2021). What is plotted is lagged by 5.8 days (McAloon et al., 2020) to allow for the delay from infection to first symptoms. The black dots with confidence intervals, joined by a dashed line, are reconstructions of incidence from the ONS statistical surveillance survey. Both reconstructions are rescaled for plotting on the fatal incidence scale.

3.1 Checking epidemic model based reconstructions

Analyses from Imperial College (Flaxman et al., 2020; Knock et al., 2020, 2021), and the MRC unit in Cambridge (Birrell et al., 2021) were influential in promoting the idea that lockdown was essential for turning around the first wave of infection. All were based on fitting epidemic models and all showed incidence continuing to surge until the eve of lockdown.

In the case of (Birrell et al., 2021), transmission rates were controlled by a step function that changed weekly, except prior to lockdown where it was constant: such a model has no way of accommodating a drop in incidence prior to lockdown. The surging incidence result was built in. Flaxman et al. (2020) used an epidemic renewal model in which R was controlled by a step function that changed when government policy changed, which it did frequently up until lockdown, but not for weeks thereafter. The model was fitted to daily death data, similarly to the simple deconvolution model. Recasting this epidemic model to use a penalized spline to model $log(R)$, estimation is fairly straightforward using the methods of section 2. This more data driven version of the model gives results very similar to figure 2, rather than surging incidence up until lockdown. The change relates to removal of the assumption that R is essentially constant after lockdown, which is not possible if lockdown is effective at limiting transmission.
Knock et al. (2020) used a more elaborate age structured model of the epidemic and health service to infer R and incidence from multiple streams of health data from each of the 7 English health service regions. The model for each region has some 700 state variables. The key transmission rate modifier controlling overall dynamics was a piecewise linear function: inference targeted the function values at its 12 breakpoints. Again the model setup allowed insufficient flexibility post lockdown. Knock et al. (2020) used particle filtering for inference, apparently with only 96 particles. Wood and Wit (2021) re-implemented the model, replacing the piecewise linear function with a more data driven adaptive spline. Inference used the approach outlined in section 2.1. Again the results then align with figure 2.

3.2 International results

The correspondence between incidence trajectories directly estimated by statistical sampling methods and the deconvolution approach, and the agreement with epidemic model fitting approaches once over-restrictive parametric assumptions are relaxed, all suggest that the deconvolution method is sufficiently trustworthy for application to other countries. We were able to obtain data by exact day of death for 9 additional countries. Figure 3 shows the incidence reconstructions for each, with dates of full national stay at home lockdowns shown as vertical red lines. For the most part the pattern of incidence decline preceding lockdowns is repeated.
This obviously does not imply that lockdowns had no effect. For example, both Sweden and Switzerland experienced broader waves, with subsidiary peaks, when they did not impose full lockdowns, but lesser measures. Countries that locked down tended to have shorter waves of infection. But the results do not support the idea that lockdowns were essential to turning around infection waves.

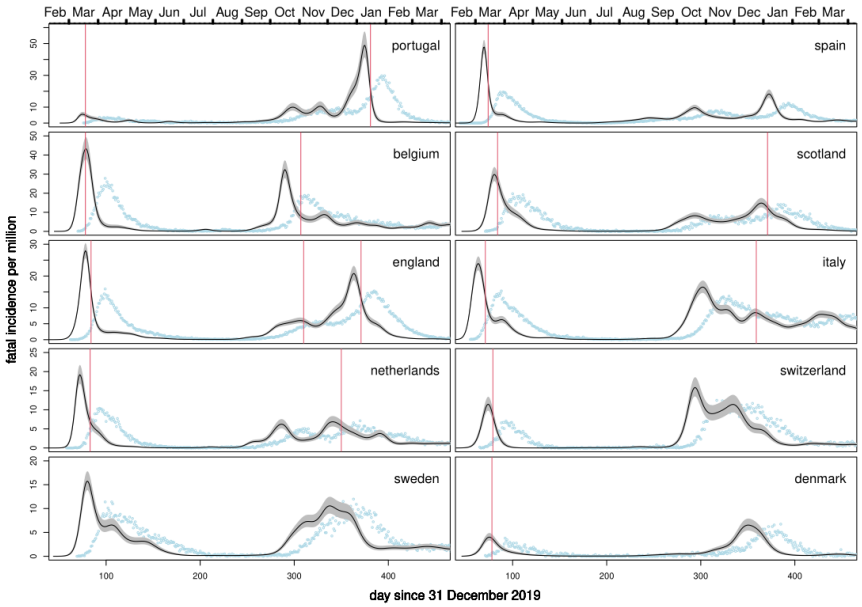


FIGURE 3. Fatal Covid-19 incidence rates in relation to lockdown timings inferred from daily death data for countries where these are available by exact day of death (rather than reporting day).

4 Conclusions

The combination of Newton or quasi-Newton methods to optimize penalized likelihoods with respect to model parameters, and simple generalized Fellner-Schall type updates to optimize smoothing parameters, facilitates the use of smooth functions within a variety of non-linear model structures. The resulting semi-parametric models may be less prone to severe model mis-specification bias than more parametric alternatives. The example of Covid peak incidence timing relative to lockdowns illustrates that the consequences of model mis-specification biases can potentially be quite grave. It seems unlikely that the reduction in suffering brought about by lockdowns was, or could be, optimally balanced against the profound damage caused by lockdowns, given a false belief that lockdowns were essential to turning around infection waves. At least in the UK, the lockdown damage includes an exacerbation of economic deprivation of the sort clearly linked to substantial life loss/ early death (Marmot et al., 2020).

References

Birrell, P., J. Blake, E. Van Leeuwen, N. Gent, and D. De Angelis (2021). Real-time nowcasting and forecasting of COVID-19 dynamics in Eng-

land: the first wave. *Philosophical Transactions of the Royal Society B 376*(1829), 20200279.

Boccia, S., W. Ricciardi, and J. P. Ioannidis (2020). What other countries can learn from Italy during the COVID-19 pandemic. *JAMA internal medicine 180*(7), 927–928.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

Fellner, W. H. (1986). Robust estimation of variance components. *Technometrics 28*(1), 51–60.

Flaxman, S., S. Mishra, A. Gandy, H. J. T. Unwin, et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature 584*(7820), 257–261.

Gomes, M. G. M., M. U. Ferreira, R. M. Corder et al. (2022). Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold. *Journal of theoretical biology 540*, 111063.

Kimeldorf, G. S. and G. Wahba (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics 41*(2), 495–502.

Knock, E. S., L. K. Whittles, J. A. Lees, P. N. Perez Guzman, et al. (2020). Report 41: The 2020 SARS-CoV-2 epidemic in England: key epidemiological drivers and impact of interventions. *Imperial College London*.

Knock, E. S., L. K. Whittles, J. A. Lees, P. N. Perez Guzman, et al. (2021). Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. *Science Translational Medicine 13*(602), eabg4262.

Linton, N. M., T. Kobayashi, Y. Yang, K. Hayashi, et al. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. *Journal of clinical medicine 9*(2), 538.

Marmot, M., J. Allen, T. Boyce, P. Goldblatt & J. Morrison (2020). *Health Equity in England: The Marmot Review 10 Years On*. The Health Foundation.

McAloon, C., Á. Collins, K. Hunt, A. Barber, A. W. Byrne, et al. (2020). Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. *BMJ open 10*(8), e039652.

McKeigue, P. M., D. McAllister, D. Caldwell, C. Gribben, et al. (2021). Relation of severe COVID-19 in scotland to transmission-related factors and risk conditions eligible for shielding support: REACT-SCOT case-control study. *BMC Medicine 19*(1), 1–13.

McKeigue, P. M. and S. N. Wood (2022). Limitations of models for guiding policy in the covid-19 pandemic. *medRxiv*.

Novozhilov, A. S. (2008). On the spread of epidemics in a closed heterogeneous population. *Mathematical biosciences 215*(2), 177–185.

Pritchard, M., E. A. Dankwa, M. Hall, J. K. Baillie, G. Carson et al. (2020). ISARIC clinical data report 4 October 2020. *medRxiv*.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika 78*(4), 719–727.

Verity, R., L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases 20*(6), 669–677.

Walker, P., C. Whittaker, O. Watson, M. Baguelin, K. Ainslie et al. (2020). Report 12: The global impact of COVID-19 and strategies for mitigation and suppression. *Imperial College London*.

Wang, D., B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang et al. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA 323*(11), 1061–1069.

Ward, H., G. Cooke, M. Whitaker, R. Redd et al. (2021). REACT-2 Round 5: increasing prevalence of SARS-CoV-2 antibodies demonstrate impact of the second wave and of vaccine roll-out in England. *medRxiv*.

Wood, S. N. (2020). Did COVID-19 infections decline before UK lockdown? *arXiv preprint ArXiv:2005.02090*.

Wood, S. N. (2021). Inferring UK COVID-19 fatal infection trajectories from daily mortality data: were infections already in decline before the UK lockdowns? *Biometrics 78*(3), 1127–1140.

Wood, S. N. and M. Fasiolo (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics 73*(4), 1071–1081.

Wood, S. N. and E. C. Wit (2021). Was R<1 before the English lockdowns? On modelling mechanistic detail, causality and inference about Covid-19. *PloS one 16*(9), e0257455.

Wu, J. T., K. Leung, M. Bushman, N. Kishore, R. Niehus et al. (2020). Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine 26*(4), 506–510.

Part II

State-switching decision trees

Timo Adam[1], Marius Ötting[2], Rouven Michels[2]

[1] University of Copenhagen, Denmark
[2] Bielefeld University, Germany

E-mail for correspondence: tiad@math.ku.dk

Abstract: Decision trees are simple yet powerful machine learning tools. However, when applied to time series data, they do not accommodate for features that are often found in such data, such as state-switching over time. To account for time-varying functional relationships between the response variable and covariates, we propose state-switching decision trees, where, at any time point, a Markov chain determines the tree that generates the corresponding outcome. The suggested approach is illustrated using American football data, where we predict whether a team attempts to reach the opposing team's end zone by either running or passing the ball conditional on covariates, such as the current quarter and score, and demonstrate how the states can be linked to the current level of the team's risk-taking. R code that implements the proposed methods is available at https://github.com/timoadam/MarkovSwitchingDecisionTrees.

Keywords: Decision trees; EM algorithm; Hidden Markov models; Time series modelling.

1 Introduction

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

- a hidden state process $\{S_t\}_{t=1,\dots,T}$ (e.g., the current level of a team's risk-taking);
- an observed state-dependent process $\{Y_t\}_{t=1,\dots,T}$ (e.g., a binary variable indicating whether the team attempts to reach the opposing team's end zone by either running or passing the ball).

The state process is modelled by a discrete-time, N-state Markov chain with initial probabilities $\delta_i = \Pr(S_1 = i)$, $i = 1, \dots, N$, and transition

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

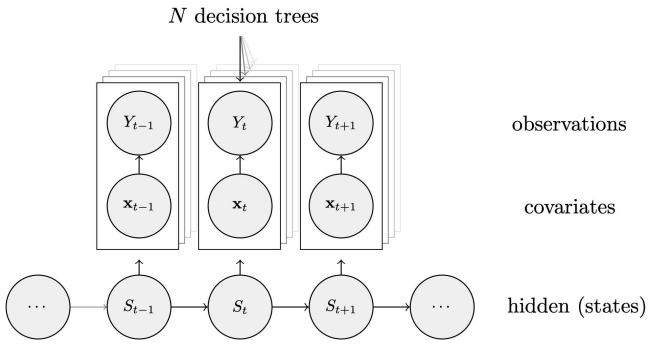


FIGURE 1. Dependence structure of state-switching decision trees. The state of the Markov chain that is active at time point t, S_t, selects one of N possible trees that generates the corresponding outcome Y_t depending on covariates $\mathbf{x}_t$.

probabilities $\gamma_{i,j} = \Pr(S_t = j \mid S_{t-1} = i)$, $i, j = 1, \ldots, N$. The state-dependent process is modelled by N trees, where the state of the Markov chain that is active at time point t, S_t, selects the tree that generates the corresponding outcome Y_t depending on covariates $\mathbf{x}_t$ (see Figure 1 for an illustration of the dependence structure).

2 Model fitting

Model fitting is conducted using the EM algorithm, where the joint log-likelihood of the outcomes and states is obtained by representing the state sequence by the binary variables $u_i(t) = I(S_t = i)$, $i = 1, \ldots, N, t = 2, \ldots, T$, and $v_{i,j}(t) = I(S_{t-1} = i, S_t = j)$, $i, j = 1, \ldots N$, $t = 1, \ldots, T$, such that

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log\left(\delta_{s_1} \prod_{t=2}^{T} \gamma_{s_{t-1},s_t} \prod_{t=1}^{T} \Pr(Y_t = y_t \mid S_t = s_t)\right) \\ &= \sum_{i=1}^{N} u_i(1) \log(\delta_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T} v_{i,j}(t) \log(\gamma_{i,j}) \\ &\quad + \sum_{i=1}^{N}\sum_{t=1}^{T} u_i(t) \log\big(\Pr(Y_t = y_t \mid S_t = i)\big), \end{aligned}$$

where

$$\Pr(Y_t = k \mid S_t = i) = \frac{1}{n_{\tilde{m}_i}} \sum_{\substack{j=1,\ldots,T: \\ \mathbf{x}_j \in R_{\tilde{m}_i}}} I(y_j = k),$$

with $\tilde{m}_i \in 1, \ldots, M_i$ being the node for which $\mathbf{x}_t \in R_{\tilde{m}_i}$ and $n_{\tilde{m}_i}$ denoting the number of observations in region $R_{\tilde{m}_i}$ for the i-th tree. The EM

algorithm alternates between the E-step, which involves the estimation of the $u_i(t)'s$ and $v_{i,j}(t)'s$ given the current estimates, and the M-step, which involves the maximisation of the joint log-likelihood with respect to the parameters, until convergence (Zucchini *et al.*, 2016). Note that only the third term of the joint log-likelihood depends on the trees' parameters. For maximising that part within the M-step, we use the CART algorithm (Breiman, 2017; Therneau and Atkinson, 2019), where the outcomes are weighted by the current estimates of the $u_i(t)'s$, and the Gini index is used as impurity measure to select the splitting variables and split points.

3 Application to American football data

In American football, the possession team (i.e., the offense) attempts to reach the opposing team's (i.e., the defense) end zone by either running or passing the ball. For the defense, it is thus of interest to predict the possession team's play (Joash Fernandes *et al.*, 2020). We use play-by-play data covering all NFL seasons from 2012–13 to 2018–19, where the states serve as proxies for the current level of a team's risk-taking. More risky styles of play are usually aligned with a higher propensity to throw a pass (as opposed to performing a run). As covariates, we use the current quarter (*qtr*), the score difference (*score_differential*), the *down* (e.g., one corresponds to the first of four possible attempts to reach the new first down), the yards to go for a new first down (*ydstogo*), whether the quarterback is in shotgun formation (*shotgun*), and whether the match is played at *home*.

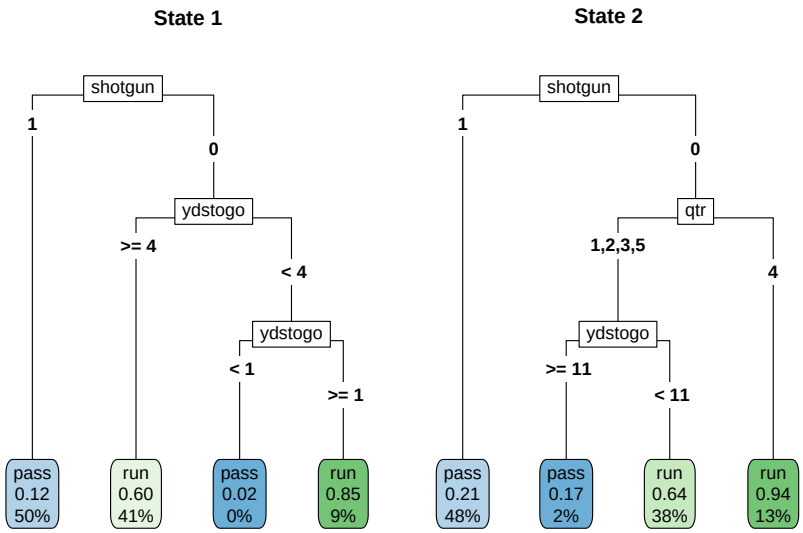


FIGURE 2. Fitted trees for state 1 (left) and 2 (right).

Figure 2 shows the fitted trees for the New England Patriots (to facilitate the interpretation, the maximum depth of each tree was chosen to be 3). The tree associated with state 2 is more likely to predict a pass, e.g., when the team is in shotgun formation (leaf node 1), but also when the team is in the first, second, third, or fourth quarter and there are at least 11 yards to go for a first down (leaf node 2), indicating a more risky style of play. In contrast, the tree associated with state 1 only predicts a pass when the team is in shotgun formation (leaf node 1) or when the team is in the first, second, or third quarter and there is less than 1 yard to go for a first down (leaf node 3), indicating a less risky style of play.

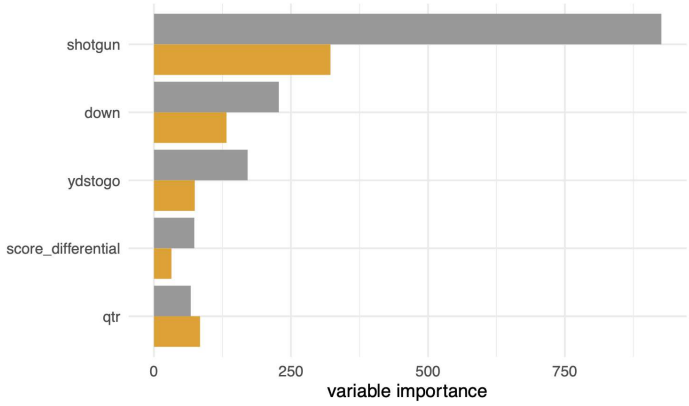


FIGURE 3. Variable importance for state 1 (grey) and 2 (orange).

To further investigate the fitted trees, we use a variable importance plot (see Figure 3). In both states, being in shotgun formation and down are the most important predictors. For the remaining predictors, the importance differs across states: in state 1, yards to go is the third most important predictor, while in state 2, quarter is more important.

4 Discussion

State-switching decision trees account for time-varying functional relationships between the response variable and covariates. Although we used binary outcomes, any time series of categorical outcomes can, in principle, be modelled. Our approach could also be adapted to other machine learning techniques such as state-switching regression trees, or more powerful ensemble methods such as state-switching random forests. State-switching decision trees thus provide a starting point for future research exploring the combination of machine learning with statistical modelling.

References

Breiman L. (2017). *Classification and regression trees*. Routledge.

Joash Fernandes, C., Yakubov, R., Li, Y., Prasad, A.K., and Chan, T.C. (2020). Predicting plays in the National Football League. *Journal of Sports Analytics*, 6(1), 35 – 43.

Therneau, T. and Atkinson, B. (2019). *rpart: recursive partitioning and regression trees*. https://CRAN.R-project.org/ package=rpart.

Zucchini W., MacDonald I.L., and Langrock R. (2016). *Hidden Markov models for time series: an introduction using R, 2nd edition*. Chapman and Hall/CRC.

Efficient stochastic learning of graphical structures for large-scale mixed data surveys

Giuseppe Alfonzetti[1], Ruggero Bellio[1], Yunxiao Chen[2], Irini Moustaki[2]

[1] University of Udine, Italy
[2] London School of Economics, United Kingdom

E-mail for correspondence: giuseppe.alfonzetti@uniud.it

Abstract: When dealing with high-dimensional multivariate data, it is often of interest to learn the dependence structure among the variables in the dataset while inducing sparsity to enhance the interpretability of the discovered patterns. Survey data are a typical example of this scenario, but the presence of continuous and discrete items aggravates the computational complexity of the learning task. Therefore, we propose a proximal stochastic gradient method which, by taking advantage of the structure of the pseudo-likelihood of the graphical model of interest, provides affordable but still efficient approximations of the conditional dependence patterns among the items.

Keywords: Graphical models; Stochastic optimisation; Structure learning;

1 Introduction

In the following, we consider the problem of learning undirected graphical structures from large-scale surveys with both categorical and continuous items. Namely, we want to learn which pairs of variables are conditionally independent given the rest of the survey and which are not. From an optimisation point of view, the problem can be tackled via proximal algorithms to account for a regularisation component to induce sparsity in the edge structure, as outlined in Lee and Hastie (2015). However, the per-iteration complexity of numerical optimisers grows linearly with the sample size, such that it might be computationally helpful to rely on stochastic gradient algorithms when dealing with large-scale problems.
In this contribution, we propose an efficient stochastic version of the proximal gradient algorithm which takes advantage of the structure of Besag's

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

pseudo-likelihood (Besag, 1974) to improve on the efficiency of stochastic gradient approximations based on mini-batches of observations (see Bottou et al., 2018 for an in-depth overview). We apply it to estimate the network structure of data from the US National Health and Nutrition Examination Survey (NHANES) for the 2017-2018 cohort.

2 Proximal stochastic gradient for mixed graphical models

Let $Y = (Y_1, \ldots, Y_p)^\top$ be the multivariate random variable corresponding to the graph of interest, with density $p(Y;\theta)$ and $\theta \in \mathbb{R}^d$. Let $\mathcal{C} = \{1, \ldots, p_c\}$ be the set of indexes corresponding to continuous nodes, while $\mathcal{D} = \{p_c + 1, \ldots, p\}$ the one for the categorical ones, for a total of p nodes. Rather than optimising $\log p(y;\theta)$, which requires the computation of the partition function of the graph here denoted as $Z(\theta)$, we follow Lee and Hastie (2015) and focus on the maximisation of the average log pseudo-likelihood

$$p\ell_n(\theta; y) = \frac{1}{n}\sum_i^n p\ell(\theta; y_{i.}) = \frac{1}{n}\sum_i^n \left\{ \sum_{j\in\mathcal{C}} \ell_j(\theta; y_{i.}) + \sum_{j'\in\mathcal{D}} \ell_{j'}(\theta; y_{i.}) \right\}, \quad (1)$$

where $\ell_j(\theta; y_i) = \log p(y_{ij}|y_{i\setminus j};\theta)$, $y_{i.}$ is the realisation of Y on the i-th unit, y_{ij} its j-th element and $y_{i\setminus j}$ is the collection of elements in $y_{i.}$ except for y_{ij}. The benefit of Besag's pseudo-likelihood is that it allows frequentist estimation of graphical structures by specifying only the full conditional margin of each node, not their joint distribution. Hence, it does not need to compute $Z(\theta)$, whose cost grows exponentially in p. For the sake of brevity, we refer to Lee and Hastie (2015) for the exact parameterisation of $\ell_j(\theta; y_i)$, j$\in \mathcal{C}, \mathcal{D}$. In order to induce sparsity in the estimation of the edge parameters, the negative log pseudo-likelihood can be augmented with $g(\theta)$, a non-smooth regularisation term composed of lasso and group-lasso penalties such that the optimisation problem to consider becomes

$$\min_{\theta\in\mathbb{R}^d} -p\ell_n(\theta; y) + \lambda g(\theta), \quad (2)$$

where λ is a scalar regularisation parameter. A natural approach to deal with the minimisation in (2) is to take advantage of proximal algorithms. In particular, given an initial value θ_0 and a stepsize scheduling η_t, the generic t-th update of a proximal gradient method can be written as

$$\theta_t = \text{Prox}_{\eta_t,\lambda g}\left\{\theta_{t-1} + \eta_t \nabla p\ell_n(\theta_{t-1})\right\},$$

with $\text{Prox}_{\eta,\lambda g}(\theta)$ being the proximal map associated to $\lambda g(\theta)$, as reviewed in Parikh and Boyd (2014). When dealing with large-scale problems, it

can be the case that $\nabla p\ell_n(\theta)$ is too expensive to compute exactly since its complexity grows with $O(np)$. Atchadé et al. (2017) propose replacing the gradient of the objective function with an unbiased stochastic approximation $U(\theta;\xi)$. Thus, in our case, we need it to satisfy $\int U(\theta;\xi)d\xi = \nabla p\ell_n(\theta)$. However, while Atchadé et al. (2017) directly optimise $\log p(y;\theta)$ by constructing gradients based on the stochastic approximation of $Z(\theta)$, here we focus on the minimisation of the negative of the pseudo log-likelihood outlined in (1). A straightforward strategy would be to use a standard mini-batch stochastic gradient, $U'(\theta) = m^{-1}\sum_k^m p\ell(\theta; y_{k\cdot})$, where m is the number of observations drawn uniformly at random from the dataset. The complexity of $U'(\theta)$ becomes $O(mp)$, which is independent of n.
However, while $U'(\theta)$ can be much cheaper to evaluate than $\nabla p\ell_n(\theta)$, its variability increases the number of iterations the algorithm needs to converge. Intuitively, the lower the noise $U'(\theta)$ injects in the optimisation, the faster the convergence of the algorithm. In this regard, by taking advantage of the log pseudo-likelihood sum structure, we propose defining $U(\theta;\xi)$ as the gradient of the combination of a random subset of conditional margins drawn from the available dataset. Namely, at each iteration t, the algorithm alternates

$$\xi_{tij} \overset{i.i.d.}{\sim} \text{Bernoulli}(\gamma), \text{ for } i = 1,\dots,n \text{ and } j = 1,\dots,p; \tag{3}$$

$$U^*(\theta_{t-1};\xi_t) = \frac{1}{n\gamma}\sum_i^n \nabla\left\{\sum_{j\in\mathcal{C}} \xi_{tij}\ell_j(\theta_{t-1};y_{i\cdot}) + \sum_{j'\in\mathcal{D}} \xi_{tij'}\ell_{j'}(\theta_{t-1};y_{i\cdot})\right\};$$

$$\theta_t = \text{Prox}_{\eta_t,\lambda g}\left\{\theta_{t-1} + \eta_t U^*(\theta_{t-1},\xi_t)\right\}.$$

In principle, it might seem that an equivalent algorithm can be obtained by using the mini-batch approximation. Nevertheless, because of the unique nature of the pseudo-likelihood, the variance of a mini-batch stochastic gradient can be much larger than that of $U^*(\theta;\xi)$, even when setting $n\gamma = m$ to match the two computational costs.
An intuitive explanation for the better efficiency of $U^*(\theta;\xi)$ stems from investigating how it compares to the standard mini-batch approach when fixing $n\gamma = m = 1$. Since the standard mini-batch approach considers m units, in this scenario, it only accounts for one single observation, with the cost of the gradient being $O(p)$. Regardless, the algorithm in (3) constructs $U^*(\theta;\xi)$ by pooling together $O(p)$ conditional contributions from different observations rather than just one. Depending on the problem, the dependence structure among the nodes can be more or less tight. Thus, choosing p full-conditional components from the same observation might lead to a certain overlapping of statistical information. In this regard, $U^*(\theta;\xi)$ makes better use of the available computational resources since it spreads the p contributions among many observations.

3 Real data application

To illustrate the effectiveness of the proposed methodology, we analyse data from the NHANES survey for the 2017-2018 cohort. Because of the inability of the algorithm in its current form to deal with missing values, let us focus on a subset of the complete pool of items available. In particular, we consider questionnaire responses on multiple areas, such as mental health, smoking-behaviour, sleep disorders and alcohol use. In addition to questionnaire responses, we account for demographic variables related to age, income, education level and marital status. Finally, the survey also includes medical measurements and examination data, among which we retained some elemental body indicators such as weight, height, blood pressure and heart rate. The final dataset accounts for $n = 1527$ adults (20 years old or more) and $p = 36$ variables (18 continuous, 2 binary, 6 with three categories and 10 with four) for a total $d = 3178$ parameters to estimate. The high ratio d/n makes the setting a relevant application for proximal methods.
The parameter γ is set to account for $1/10$ of the train set per iteration, and, for illustration purposes, the algorithm stops after one single pass through the train data. The regularisation parameter λ is chosen by minimising the negative log-likelihood of a 20% holdout set over 100 equispaced points in the interval $[10^{-3}, 1]$. Figure 1 shows that the optimal choice for λ decreases as the optimisation proceeds in terms of complete passes through the train set. Such behaviour is expected since the number of iterations plays a role similar to the sample size. In fact, as the optimisation proceeds, the algorithm accounts for an increasing share of the data, and the need for regularisation decreases.
After a complete pass through the training partition, the identified graph selects almost the 10% of the edges as active. Figure 2 shows the estimated structure, highlighting with colours the area to which each item belongs. The size of nodes and edges are proportional, respectively, to the number of edges connecting that node to the rest of the graph and the norm of the edge analysed. The figure outlines some remarkable patterns. For example, the items aiming to identify depression are strongly interdependent but connect to the rest of the graph mainly through the node *DPQ090*. Such label refers to a very crucial item for the mental-health questionnaire, namely "Thought you would be better off dead". The figure highlights how this node connects the depression-related questionnaire to some of the other areas, via the items "Home owned, bought, rented" (*HOQ065*), "Marital status" (*DMDMARTL*), "Born in US" (*DMDBORN4*) and "Sleep hours, weekends" (*SLD013*). Furthermore, it follows that the depression-related items are independent of body and blood measurements and smoking and alcohol behaviours when conditioned on sleep disorders, housing status and demographic variables.
As for all statistical methods, such models must be intended as a simplification of the complex reality they aim to describe. However, highlighting

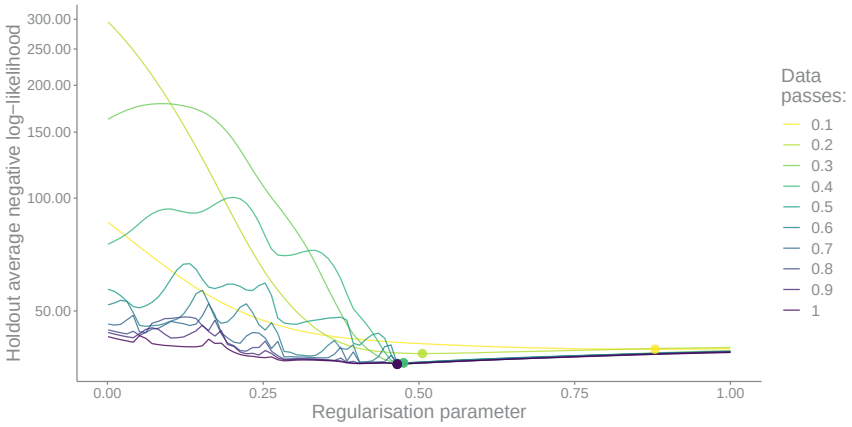


FIGURE 1. Holdout average negative log-likelihood with varying regularisation parameter. Colours refer to the number of complete passes through the dataset or the number of iterations (each step of the optimisation accounts for 10% of the data). Points are located at the optimal value of λ for each share of the dataset.

the statistical dependencies observed in the data can still provide some valuable insights to deepen the understanding of such delicate issues.

4 Discussion and ongoing work

Structure learning techniques have advanced notably in the last years, and the combination of proximal methods and stochastic approximations is a promising direction to address the scalability of such algorithms. In particular, with extensive population surveys, they can be extremely helpful in investigating the dependence structure among the variables of interest.
In order to extend the estimation to a larger pool of items from the NHANES survey, we plan to work out an imputation mechanism for the missing values. Finally, an open-source software package is in development as a companion to the proposed method.
For reproducibility purposes, the code used in Section 3 is available at `https://github.com/giuseppealfonzetti/iwsm2023`.

Aknowledgments

This work was supported by the Departmental Strategic Plan (PSD) of the University of Udine, Interdepartmental Project on Public Administration (2022-2025).

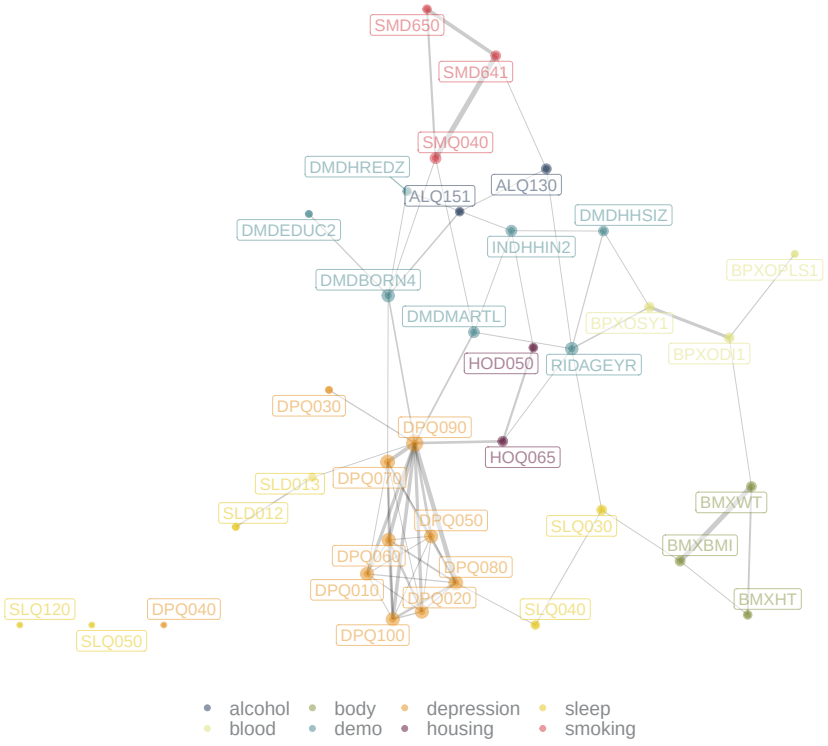


FIGURE 2. Network structure for the NHANES survey (adults, 2017-2018).

References

Atchadé, Y. F., Fort, G. and Moulines, E. (2017) On perturbed proximal gradient algorithms. *The Journal of Machine Learning Research*, **18**(1), 310–342.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, **36**(2), 192-225.

Bottou, L., Curtis, F. E. and Nocedal, J. (2018) Optimisation methods for large-scale machine learning. *SIAM Review*, **60**(2), 223-311.

Lee, J. D. and Hastie, T. J. (2015) Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, **24**(1), 230-253.

Parikh, N. and Boyd, S. (2014) Proximal algorithms. *Foundations and Trends in Optimization*, **1**(3), 127-239.

Flexible habitat selection analysis with generalized additive models

Rafael Arce Guillen[1], Jennifer Pohle[1], Björn Reineking[2], Ulrike Schlägel [1]

[1] Institute of Biochemistry and Biology, University of Potsdam, Germany
[2] INRAE, Grenoble, France

E-mail for correspondence: `arceguillen@uni-potsdam.de`

Abstract: Habitat selection analysis is a statistical framework used to understand sequential movement decisions of animals based on spatial features. In addition, this approach specifies the availability of sequential locations through a movement kernel. This movement kernel is defined as the product of parametric distributions for the step lengths and turning angles based on sequential animal locations. However, this assumption is not plausible for real data. The objective of this paper is to relax the need for parametric distributions with help of *Generalized Additive Models* (GAM) and the `R`-package `mgcv`. For this, we propose to specify the movement kernel as a bivariate tensor product instead of specifying parametric distributions for the movement kernel.

Keywords: animal movement; step selection analysis; generalized additive models.

1 Introduction

With the advances in tracking systems technology (every few seconds between sequential locations), a very common statistical approach to understand animal movement is called *Integrated Step Selection Analysis* (iSSA) (Avgar et al. 2016). With this approach, researchers try to understand movement decisions based on spatial features in the study area. Here they can account for the fact that sequential locations may be not independent to each other by including exponential family distributions of distances/step lengths (SLs) and turning angles (TAs) between locations. Typically, a gamma distribution and a von Mises distributions are assumed for the SLs and TAs respectively. Thus, Avgar et al. (2016) defined the movement kernel ϕ as the product of these two distributions, assuming that step lengths

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

and turning angles are independent to each other.

Despite this being an interesting approach, the movement kernel parameters estimates are usually not of interest. In addition, in real data analysis, the movement kernel may not be well represented by these parametric distributions. Another disadvantage is that in iSSA analyses, the step lengths and turning angles are assumed to be independent from each other. This assumption is not necessarily plausible for real data. For this reason, we propose to fit telemetry data with help of *Generalized Additive Models* (GAM). Doing this, the movement kernel can be specified as a two-dimensional tensor product with help of the `R`-package `mgcv` (Wood 2011). This can be done with the implementation of Muff et al. (2020) since it uses the GLM framework needed for the GAM approaches by including time-specific intercepts modelled as random intercept with a large fixed variance parameter. Thus, similar to Arce Guillen et al. (2023), in this paper we will interpret the movement process as sequence of Nonhomogeneous Point processes (NHPPs) which can be approximated to a single Poisson GLM model. We name our approach GAM-SSA.

2 Model formulation

Conceptually, we use the same spatial density as Forester et al. (2009). Thus, the spatial density for observing location s_t at time t given the last two observed locations s_{t-1} and s_{t-2} is given by:

$$f(s_t|s_{t-2}, s_{t-1}; \boldsymbol{\beta}) = \frac{\overbrace{\phi(s_{t-2}, s_{t-1}, s_t)}^{\text{Movement kernel}} \overbrace{\omega(\boldsymbol{X}(s_t); \boldsymbol{\beta})}^{\text{RSF}}}{\underbrace{\int_{q_t \in S} \phi(s_{t-2}, s_{t-1}, q_t)\omega(\boldsymbol{X}(q_t); \boldsymbol{\beta})\partial q_t}_{\text{Normalizing constant}}} \tag{1}$$

However, rather than defining the movement kernel as a product of parametric distributions, we define it as a two-dimensional positive function that depends on the step lengths and turning angles:

$$\phi(s_{t-2}, s_{t-1}, s_t) = \exp\left(f(SL_t, TA_t)\right) \tag{2}$$

The function $f()$ reflects the animal preferences for SL and TA combinations. In addition, the exponential function ensures positivity and since it is an increasing monotonic function, the preference relation remains the same. Using the trick of Muff et al. (2020), we specify the corresponding joint log-likelihood as the sum of a sequence of unconditional NHPPs (Arce Guillen et al. 2023). This model specification is a generalization of the iSSA approach. With the presented specification, parametric distributions coming from the exponential family are also included in this framework.

The selection of spatial features is expressed by a RSF ω:

$$\omega(\mathbf{X}(s_t);\boldsymbol{\beta}) = \exp(\beta_1 X_1(s_t) + ... + \beta_p X_p(s_t) + u(s_t)) = \exp(\eta(s_t)) \ . \quad (3)$$

where β_j for $j = 1, ..., p$ represents the strength of spatial feature $X_j(s_t)$. In addition, u represents all the missing spatial variation not explained by the spatial features.

Thus, the sequence of NHPPs is approximated by sampling at each time point integration points uniformly over a disk of ratio equal to at least the maximum observed step length (Figure 1).

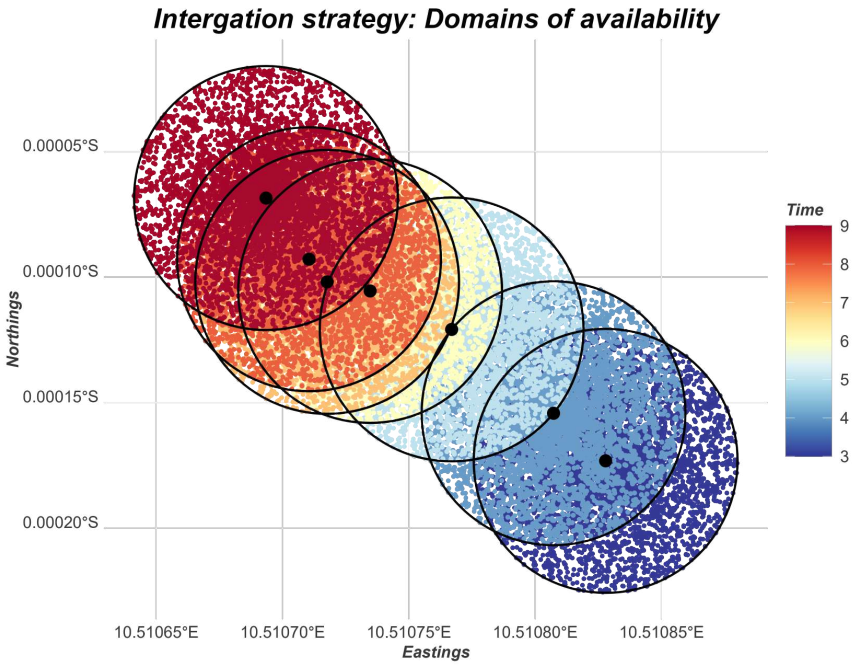


FIGURE 1. Integration strategy: We sample at each time point integration points uniformly within the respective disks of availability.

3 Simulation

We have simulated animal tracks using different movement kernel specifications. In these settings, the animals move according to three spatial

covariates "x1", "x2" and "cen". The latter represents a centralizing tendency. For each setting, we produced 100 animal tracks consisting of 1000 locations. We simulated the four different scenarios. For the first scenario, we assume a bimodal distribution for the SLs and a von Mises distribution for the TAs. In the second case, we use a bivariate copula distribution using a Weibull distribution for the SLs and a wrapped Cauchy distribution for the TAs. we assume a Gamma and a Von Mises distribution for the SLs and TAs respectively. The third scenario consists of the classical Gamma/von Mises setting. In the last scenario, we use a uniform movement kernel. These tracks have been fitted using the GAM-SSA and the iSSA using the same integration points. In these simulations, for simplicity, we have no missing spatial variation.

4 Results

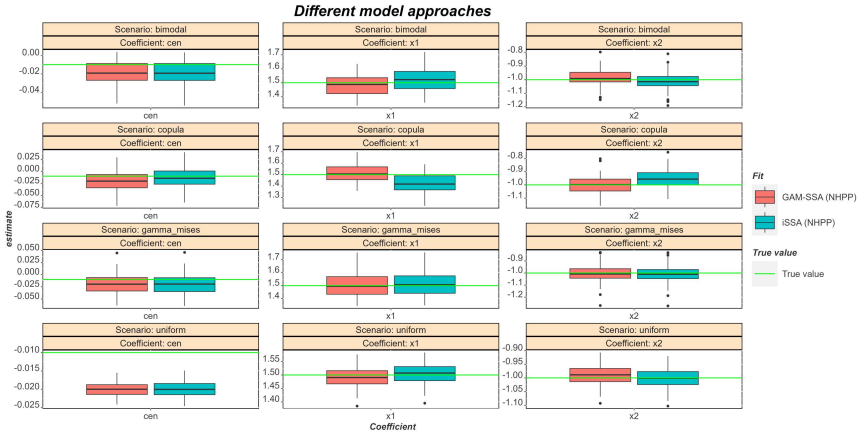


FIGURE 2. Fixed effects results for the simulation scenarios. Each box-plot represents the fixed effects estimates of 100 simulations.

With the exception of the centralizing tendency "cen" for the uniform movement kernel case, the GAM-SSA estimated the effects of the spatial covariates without any noticeable bias in any direction (Figure 3). However, the centralizing tendency was also biased for the classical iSSA approach. In the copula case, our method seemed to be more accurate than the iSSA approach. For the other cases, the GAM-SSA returned very similar fixed effects as the iSSA approach. Thus, the GAM-SSA can be used without any concern.

As observed in Figure 3, the usage of splines is leading to estimating the movement kernel accurately for all four scenarios with exception of the

Coefficient	Fit	Scenario	Mean	True value	0.05-quantile	0.95-quantile	Coverage
x1	GAM-SSA (NHPP)	Bimodal	1.48	1.50	1.37	1.60	0.94
x1	iSSA (NHPP)	Bimodal	1.52	1.50	1.41	1.64	0.96
x2	GAM-SSA (NHPP)	Bimodal	-0.99	-1.00	-1.10	-0.90	0.94
x2	iSSA (NHPP)	Bimodal	-1.01	-1.00	-1.12	-0.92	0.93
cen	GAM-SSA (NHPP)	Bimodal	-0.02	-0.01	-0.03	0.00	0.91
cen	iSSA (NHPP)	Bimodal	-0.02	-0.01	-0.04	-0.00	0.91
x1	GAM-SSA (NHPP)	Copula	1.51	1.50	1.39	1.64	0.98
x1	iSSA (NHPP)	Copula	1.42	1.50	1.30	1.56	0.84
x2	GAM-SSA (NHPP)	Copula	-1.00	-1.00	-1.11	-0.91	0.96
x2	iSSA (NHPP)	Copula	-0.95	-1.00	-1.06	-0.85	0.91
cen	GAM-SSA (NHPP)	Copula	-0.02	-0.01	-0.05	0.01	0.91
cen	iSSA (NHPP)	Copula	-0.01	-0.01	-0.04	0.01	0.96
x1	GAM-SSA (NHPP)	Gamma-Mises	1.50	1.50	1.37	1.63	0.97
x1	iSSA (NHPP)	Gamma-Mises	1.51	1.50	1.38	1.64	0.97
x2	GAM-SSA (NHPP)	Gamma-Mises	-1.00	-1.00	-1.10	-0.90	0.95
x2	iSSA (NHPP)	Gamma-Mises	-1.01	-1.00	-1.10	-0.90	0.94
cen	GAM-SSA (NHPP)	Gamma-Mises	-0.02	-0.01	-0.05	0.01	0.94
cen	iSSA (NHPP)	Gamma-Mises	-0.02	-0.01	-0.05	0.01	0.94
x1	GAM-SSA (NHPP)	Uniform	1.49	1.50	1.44	1.54	0.95
x1	iSSA (NHPP)	Uniform	1.50	1.50	1.45	1.56	0.96
x2	GAM-SSA (NHPP)	Uniform	-0.99	-1.00	-1.05	-0.94	0.96
x2	iSSA (NHPP)	Uniform	-1.00	-1.00	-1.07	-0.95	0.94
cen	GAM-SSA (NHPP)	Uniform	-0.02	-0.01	-0.02	-0.02	0.00
cen	iSSA (NHPP)	Uniform	-0.02	-0.01	-0.02	-0.02	0.00

TABLE 1. Fixed effects: Summary results of the simulation study. The coverage represents the percentage for which the true values were covered by the 95% confidence intervals.

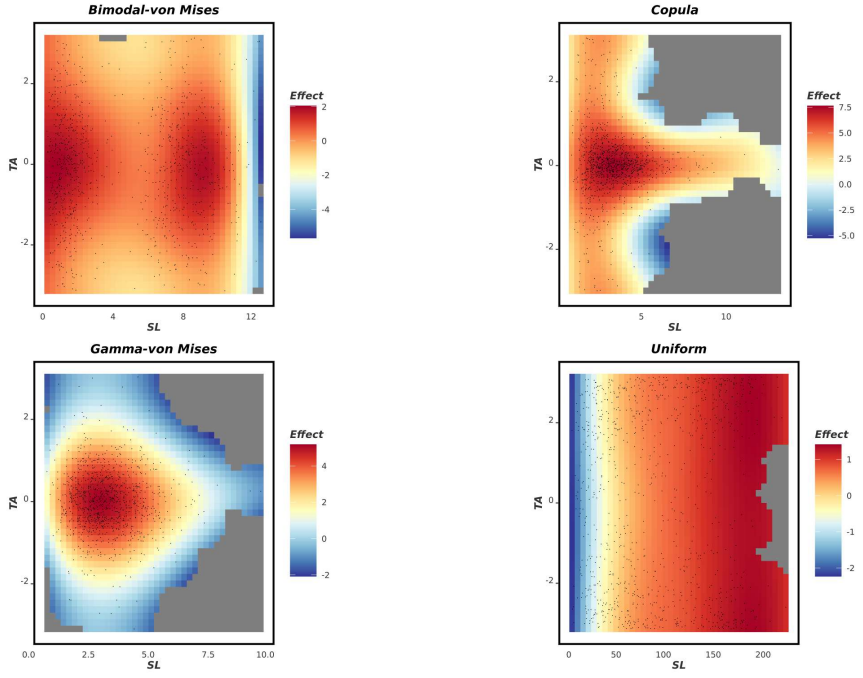


FIGURE 3. GAM-SSA: Estimated movement kernel for one simulated track. The black dots represent the observed animal locations.

uniform movement kernel, which does not represent a uniform movement behaviour. However, this is the case since the model is overestimating the coefficient of the centralizing tendency "cen" and accounts for this through the movement kernel.

References

Arce Guillen, R., Lindgren, F., Muff, S., Glass, T. W., Breed, G. A. and Schlaegel, U (2023). *bioRxiv*.**p. 2023.01. 17.52436**,

Avgar, T., Potts, J. R., Lewis, M. A. and Boyce, M. S. (2016). Integrated step selection analysis: bridging the gap between resource selection and animal movement. *Methods in Ecology and Evolution* **7(5)**, 619 – 630.

Forester, J. D., Im, H. K. and Rathouz, P. J. (2009). Accounting for animal movement in estimation of resource selection functions: sampling and data analysis. *Ecology* **90(12)**, 3554 – 356.

Muff, S., Signer, J. and Fieberg, J. (2020). Accounting for individual–specific variation in habitat–selection studies: Efficient estimation of mixed–effects models using bayesian or frequentist computation. *Journal of Animal Ecology*, **89(1)**, 80 – 92.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *ournal of the royal statistical society: Series b (statistical methodology)*, **73(1)**, 3 – 36.

An information-theoretic perspective on double descent in flooded boosting

Chiara Balestra[1], Andrés Madariaga[1], Emmanuel Müller[1], Christian Staerk[2], Andreas Mayr[2]

[1] Department of Computer Science, TU Dortmund, Dortmund, Germany
[2] Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Bonn, Germany

E-mail for correspondence: chiara.balestra@cs.uni-dortmund.de

Abstract: Boosting models are often used in statistics and machine learning, typically yielding good prediction accuracy on test data due to a relatively slow overfitting behavior. On the other hand, in special cases, they can even lead to a double descent phenomenon of the test risk, i.e., the performance on the test set first gets worse and then improves again as the number of boosting rounds increases. A strategy to enforce this double descent is to use a *flooded* version of the loss function. We explore the behavior of boosting combined with flooding from an information-theoretic perspective, giving insights on the double descent phenomenon with potentially interesting implications about the bias-variance trade-off in statistical modelling.

Keywords: Boosting; Flooding; Double descent; Information plane.

1 Introduction and related work

Statistical models play a fundamental role in classification and regression. The stopping criterion of the models' parameters learning phase (or the regularization parameter) influences the prediction accuracy. The aim is to identify the *sweet spot* between under- and over-fitting the data. The test error trajectory can follow the classical U-shape or a more fancy W-shape, involving multiple descending curves, known as double descent (Figure 1). Boosting techniques combine weak learners and increase the model accuracy iteratively; after several *boosting rounds*, the learning stops. Ferreira and Figueiredo (2012) studied the occurrence of double descent in several prediction models; some years later, Belkin et al. (2019) showed that

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

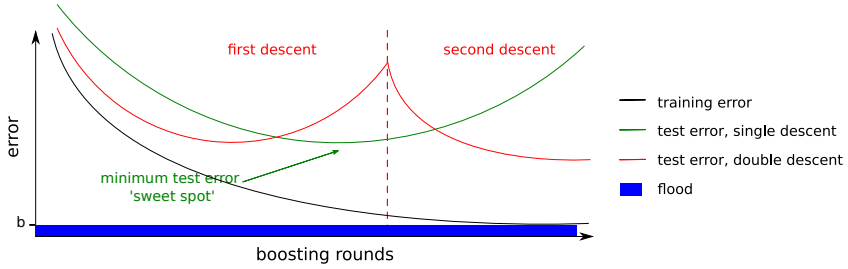


FIGURE 1. Representative plot of the training and test errors for increasing boosting rounds; in green color, the typical one descent, and in red color, the double descent behavior. The flood (in blue) impedes the training error from decreasing below b.

double descent is induced by changing the type of weak learners added at each boosting round from single trees to forests. Furthermore, Ishida et al. (2020) proposed a technique to induce double descent in prediction models by *flooding* the loss function.

We investigate conditions for observing double descent in boosting without changing the weak learners' complexity. To this purpose, we integrate a *flooded loss function*, thus increasing the chance of observing a W-shape of the test error (see Figure 1). We aim to explain the appearance of double descent from an information-theoretic perspective using the lossless compression criteria shown in Nikolaou (2021).

2 Methods and Results

A new weak learner meant to minimize a loss function J is integrated into the prediction model at each boosting round. We stick here to classical linear trees and forests of trees, such that each weak learner is a weak classifier prediction model with a pre-specified number of leaves and depth. Ishida et al. (2020) argued that minimizing the loss function might not be beneficial below a certain threshold. The authors propose to add a fixed *flood* $b > 0$ to the loss function; thus, instead of minimizing J, the aim is minimizing the modified loss defined as $\tilde{J} = |J - b| + b$. When the training error is above the threshold b, J and $\tilde{J}$ coincide; instead, when the error drops below the flood level b, the modified loss $\tilde{J}$ increases again. Hence, the training error is constrained to wiggle around the flood level b.

This flooded loss function can also lead to a W-shaped test error in boosting: we observed this phenomenon in illustrative examples (MNIST data and California housing data) and various simulations. Using an information-theoretic approach, we further explore the model's behavioral changes after adding the flood level. We need some notions beforehand; we will use the notation D_X and D_Y, respectively, to indicate the distribution of the explanatory variable X and the outcome Y of the training data. Information

theory introduces a new perspective on the relationship between the distribution of the training data and the model $\hat{f}$. The mutual information I measures the amount of information shared by two random variables and how much knowing one of the variables reduces uncertainty about the other. If the two variables are independent, knowing one of them gives no information about the second one; thus, the mutual information is zero. Mutual information definition relies on the Shannon Entropy H or equivalently on the Kullback-Liebler divergence; further details are found in Cover and Thomas (2005). With the introduced notation, the mutual information can be directly applied to study the relationship between the model $\hat{f}$ and the training data distribution. *Noiseless* datasets are defined as *ideal* datasets, where the explanatory variables D_X contains all causal and necessary information to explain the outcome D_Y. It is possible to find a model $\hat{f}$ which is *lossless*, i.e., perfectly discriminating the training data and achieving zero classification error. Furthermore, we say that a model $\hat{f}$ *maximally compresses* the data if it only captures the relevant information from D_X for describing D_Y. Returning to the prediction model $\hat{f}$, mutual information can be used to study the information shared between the model and the distributions D_X and D_Y of the training data. A lossless model satisfies $I(\hat{f}; D_Y) = I(D_Y; D_X)$ and it satisfies $I(\hat{f}; D_X) = I(\hat{f}; D_Y)$ if it maximally compresses. Both equations are satisfied in the *lossless maximal compression* case. Interpreting the equations, a lossless model contains as much information about the label as the information shared between the input features and the labels themselves. Similarly, for maximally compressed models, $\hat{f}$ is characterized by sharing the same amount of information with the labels and the input features distributions; in this contest, losslessness means that it does lose information from D_X when learning D_Y.

Inspired by Nikolau (2021), we track the model behavior for increasing boosting rounds by evaluating the entropy-normalized mutual information among $\hat{f}$ and D_X and the one among $\hat{f}$ and the target D_Y. Finally, we can plot the trajectory on the *information plane*, where the y-axis represents the entropy-normalized mutual information among model and covariates $I(\hat{f}; D_X)/H(D_X)$ and the x-axis is the entropy-normalized mutual information among model and outcome $I(\hat{f}; D_Y)/H(D_Y)$. We expect to observe two phases. First, the empirical risk minimization, when both $I(\hat{f}; D_Y)$ and $I(\hat{f}; D_X)$ increase; the model learns during this phase to better fit the training data by extracting information from D_X. If the data D_X, D_Y are noiseless, $\hat{f}$ potentially achieves zero training error, i.e., losslessness. In the second phase, the *compression*, $I(\hat{f}; D_X)$ starts decreasing, and the model reduces the information learned from D_X, while $I(\hat{f}; D_Y)$ remains stable. The model reaches the *lossless maximal compression* point after the compression phase if both noiselessness and losslessness are achievable.

Bashir et al. (2020) did not provide evidence of why using the flooded version of the loss $\tilde{J}$ potentially causes a double descent behavior of the test

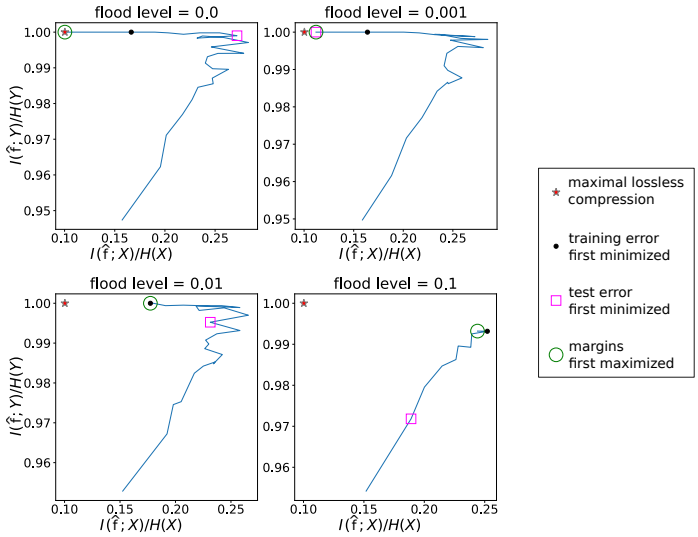


FIGURE 2. Boosting trajectories in the information plane for various flood levels.

error. We study how substituting J with $\tilde{J}$ influences the boosting model trajectory in the information plane. We run experiments on multiple synthetic and real-world datasets; Figure 2 and the analysis presented here refer to the results obtained using a synthetic dataset with 2000 samples, 20 input variables, among which 15 informative and a classification task with a binary response; the training is composed of 100 boosting rounds, with learning rate 0.1, and at each boosting round, a tree with maximum depth 6 is being added. Figure 2 shows the trajectories for increasing values of flood levels on the information plane: We have highlighted the first moments when the training and test error are minimized, the margins are maximized, and the point of maximal compression. The latter can only be reached after several boosting rounds if the model both maximal compresses and is lossless. Without flooding ($b = 0$), the model minimizes the test and training errors; as expected, the best generalization is obtained before "overfitting" the training data. Furthermore, maximal compression and margin minimization are obtained almost synchronously. When adding a flood level $b > 0$, we see a drastic change in the model's behavior: The maximal compression is never reached. This is a natural consequence as the model cannot thoroughly learn the training samples because of the flood. Furthermore, the other points change their respective positions when increasing the parameter b. The different behaviors on the information planes for the various levels of b give insights into the substantial change of the flooded boosting model. With higher flood levels, the model training is stopped before compression starts. As described, during the learning process, the non-flooded model grows along the y-axis (the learning phase)

and then decreases along the x-axis (the compression phase). However, keeping the boosting rounds fixed, when $b > 0$, this behavior stops. We still have a learning phase, while the compression phase is missing. All relevant points, the maximum margins point, and the minimum errors are nevertheless reached during the learning phase; additionally, the lack of the compression phase implies that the model cannot achieve the maximal lossless compression.

3 Conclusion

We used boosting models and investigated the relationship between boosting and flooding; in particular, we analyzed the influence of flooding in boosting models from an information-theoretic perspective. The information plane allows us to study how the model behaves through the learning process and whether there are structural changes explaining the different behaviors, i.e., a double-descent of the test error. We believe that further research on regression models as weak learners combined with flooding is warranted to investigate potential implications for the bias-variance trade-off in statistical boosting models.

Acknowledgments: This research was funded by the research training group *Dataninja* funded by the German federal state of NRW.

References

Belkin M., Hsu D., Ma S. and Mandal S. (2019). Reconciling modern machine learning practice and the classical bias–variance trade-off. *PNAS*, **116**, 15849–15854.

Nikolaou N. (2021). Lossless Compression and Generalization in Overparameterized Models: The Case of Boosting. In: *Neural Compression Workshop @ ICLR21.*

Ishida T., Yamane I., Sakai T., Niu G. and Sugiyama M. (2020). Do we need zero training loss after achieving zero training error?. In: *ICML.*

Bashir D. Montañez G., Sehra S., Segura P. S. and Lauw J. (2020). An Information Theoretic Perspective on Overfitting and Underfitting. In: *AI 2020: Advances in Artificial Intelligence.*

Ferreira A. J. and Figueiredo M. A. T. (2012). Boosting algorithms: A review of methods, theory, and applications. *Ensemble machine learning: Methods and applications*, 35–85.

Cover T. M. and Thomas J. A. (2005). *Elements of information theory* John Wiley & Sons, Ltd.

Adaptive random forests for high-dimensional regression

Moritz Berger[1], Christian Staerk[1]

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

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

Abstract: Random forests are a flexible tool for nonparameteric regression modelling. In classical random forests, for each split a subset of all explanatory variables is drawn uniformly at random. We propose adaptive random forests (AdaForest), where the probability (weight) of each explanatory variable to be considered as a splitting variable is not fixed, but is sequentially adapted based on its selection frequency in previous iterations of the algorithm. Based on a simulation study and high-dimensional gene expression data we illustrate that AdaForest can improve the prediction accuracy of classical random forests, particularly when only a subset of variables (genes) is informative. Furthermore, the adapted weights can be used to build sparser models with competitive prediction performance.

Keywords: High-dimensional data; Random forests; Regression analysis; Predictive modelling; Variable selection

1 Introduction

Classical regression approaches that relate an outcome variable Y to a set of p explanatory variables $\mathbf{X} = (X_1, \ldots, X_p)$ include linear and generalized linear models as well as their extensions to additive models. An important nonparametric alternative to these classical approaches is recursive partitioning or tree-based modelling. The main advantage of trees is that they are able to capture higher-order interactions between the explanatory variables and non-linear effects in a data-driven way. As single trees are often affected by a large variance, it can be beneficial to stabilize the results by applying ensemble methods such as random forests (Breiman, 2001). Random forests are increasingly used in practice and are subject of

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

ongoing methodological research. Even though random forests are a very flexible tool, particularly when the focus is on prediction, the selection of the most important variables to build a more parsimonious model remains challenging.
To address this issue, we propose an extension of the classical random forest algorithm, called Adaptive Random Forests (*AdaForest*). The basic concept of AdaForest is to repeatedly run the random forest algorithm while adapting the probability of each explanatory variable to be considered as a potential splitting variable (commonly assumed to be uniformly at random), taking into account the previously performed splits. The adaptation is inspired by the Adaptive Subspace (AdaSub) method (Staerk et al., 2021), focusing on those explanatory variables which turned out to be "important" (i.e., frequently selected as splitting variables) in previous iterations. The specific adaptation scheme in AdaForest can also be interpreted as sequential updating of a Dirichlet prior, which is related to a recent approach that utilizes Bayesian additive regression trees (BART; Linero, 2018). Our new extension of classical random forests (i) can yield benefits for prediction, particularly in sparse settings where the number of informative variables is much smaller than p, and (ii) enables the selection of a sparser model based on the adapted probabilities. The proposed AdaForest algorithm makes use of the R add-on package **ranger** (Wright and Ziegler, 2017), offering a fast implementation of random forests.

2 Adapative Random Forests

The classical random forest algorithm draws a set of bootstrap samples from the original sample $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, with n observations and fits a regression tree to each of the bootstrap samples. Then, for a (new) observation with explanatory variables $\tilde{\mathbf{x}}$, a prediction of the conditional mean $\mu \,|\, \tilde{\mathbf{x}}$ is obtained by averaging the predictions of the single trees. Each tree is built by recursively dividing the predictor space into disjoint subsets (nodes) using binary splits. Splitting is repeated in each newly created node until a stopping criterion is met. To mitigate the similarity of single trees, the set of explanatory variables that are candidates for splitting in each node is a randomly chosen subset of predefined size $\texttt{mtry} < p$.
We propose to fit the random forest not only once, but to apply the random forest algorithm repeatedly in an adaptive way. In each iteration of the proposed AdaForest algorithm the `mtry` variables are not selected uniformly at random from the p possible variables, but with specific probabilities (weights), depending on the selection frequencies in previous steps. More specifically, the AdaForest algorithm is initialized with equal probabilities

$$\boldsymbol{r}_0 = (r_{10}, \ldots, r_{p0}) = (\alpha_{10}/p, \ldots, \alpha_{p0}/p)\,, \tag{1}$$

where $\alpha_0 = (\alpha_{10}, \ldots, \alpha_{p0}) = (1, \ldots, 1)$ and $r_{j0} = 1/p$ is the initial weight for variable X_j. Then, after iteration k of AdaForest the weight of each

variable X_j is updated via

$$\alpha_{jk} = \alpha_{j,k-1} + \sum_{b=1}^{\texttt{ntree}_k} \sum_{s=1}^{s_b} I(S_{bs}^{(k)} = X_j)\,, \quad r_{jk} = \frac{\alpha_{jk}}{\sum_{l=1}^{p} \alpha_{lk}}\,, \tag{2}$$

where s_b denotes the number of splits in tree b, $S_{bs}^{(k)}$ is the selected variable in split s of tree b, and $I(\cdot)$ denotes the indicator function. According to (2), the weights are adapted based on the frequency the respective variables have been selected as splitting variables in the $\texttt{ntree}_k$ trees. For simplicity, in the following we only consider a single tree in each iteration k and directly adapt the weights (i.e., $\texttt{ntree}_k = 1$). The algorithm terminates when the stopping criterion $\|\boldsymbol{r}_k - \boldsymbol{r}_{k-1}\|_\infty < \varepsilon$, is satisfied, where ε is a convergence limit (e.g., $\varepsilon = 10^{-5}$). Finally, after convergence at iteration K one can fit the random forest using the adapted weights $\boldsymbol{r}_K$ with all available variables. Alternatively, one can first select a smaller number of s variables with the highest final weights $\boldsymbol{r}_K$ and then fit the random forest including only this subset of variables, accordingly. As also illustrated in our application, this step requires the specification of an appropriate threshold, which will generally depend on the specific application and modelling aims. Alternatively, data-driven approaches for the number of selected variables may be used (e.g., based on cross-validated prediction performance).

As the classical random forest algorithm, AdaForest also requires the specification of the following key parameters (see also Boulesteix et al., 2012): (i) An appropriate criterion for split selection. The most popular choice for regression is the mean squared error. (ii) Tuning parameters to control the size of the single trees. Typically one employs a small minimal node size (`mns`) criterion, which specifies the minimal number of observations required in any terminal node. (iii) The number of candidate explanatory variables (`mtry`) and the number of trees (`ntree`), which should be related to the number of (informative) explanatory variables. Here we use the default values of the R package **ranger**, i.e., $\texttt{mns} = 5$, $\texttt{mtry} = \lfloor\sqrt{p}\rfloor$ and $\texttt{ntree} = 500$ (for fitting the final forest after convergence).

3 Simulation Study

To assess the performance of our proposed AdaForest algorithm we performed a simulation study with 100 Monte Carlo replications. The aim of the study was to compare the AdaForest algorithm to the classical random forest with regard to the predictive performance. We simulated data with normally distributed outcome $y_i \sim N(\mu_i(\mathbf{x}_i), 1)$, $i = 1, \ldots, 500$, and considered scenarios with $p = 25$ (low-dimensional), $p = 100$ (moderate-dimensional) and $p = 500$ (high-dimensional) independent standard normally distributed covariates, where a fraction of 1%, 5% and 15% of the variables were informative, respectively. Using the informative explanatory variables only, the values of $\mu_i(\mathbf{x}_i)$, $i = 1, \ldots, 500$, were obtained by

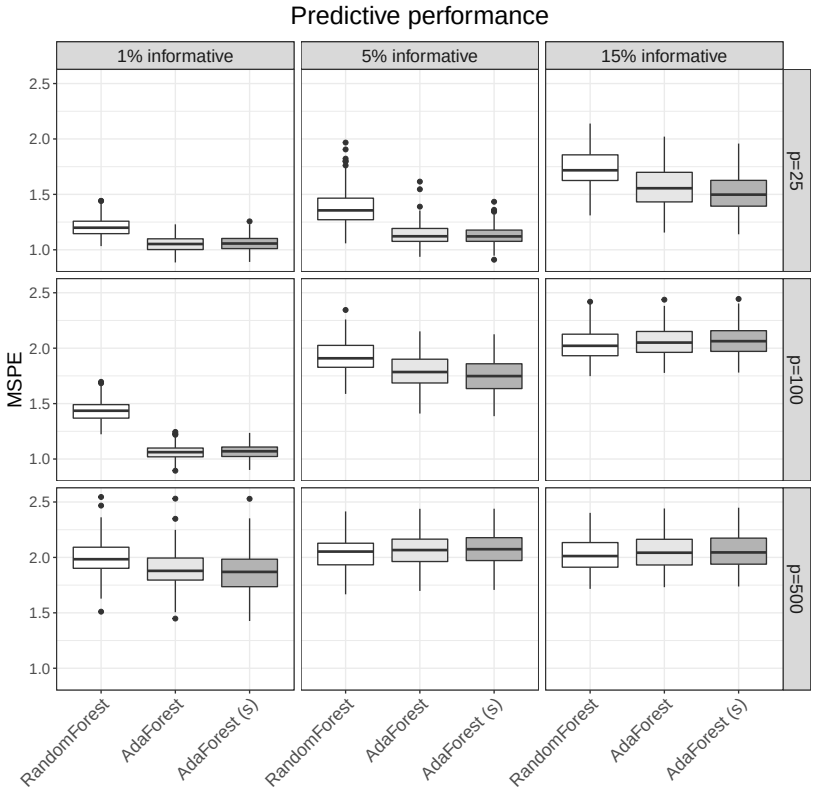


FIGURE 1. Results of the simulation study. For each of the nine scenarios, the boxplots show the mean squared prediction errors (MSPE) when fitting a classical random forest (first boxplots) and the proposed AdaForest, where all (second boxplots) or only a subset of s explanatory variables (third boxplots) were included in the final model (s chosen based on validation sample).

the standardized sum of all main effect terms $\beta_j x_{ij}$, all possible two-way interaction terms $\beta_{jk} x_{ij} x_{ik}$, and all possible three-way interaction terms $\beta_{jkl} x_{ij} x_{ik} x_{il}$, with $\beta_j, \beta_{jk}, \beta_{jkl} \sim U[-0.5, 0.5]$. In each Monte Carlo replication, we generated (i) a learning sample to fit the models, (ii) a validation sample to determine the optimal subset of s variables to be included in the final model of AdaForest, and (iii) a test sample to evaluate the mean squared prediction error (MSPE). We compared the classical random forest, the AdaForest algorithm, where all p variables were included, and the AdaForest algorithm, where only a subset of s variables with the highest weights were considered for the final forest (the number of variables s was selected based on minimizing the MSPE on the validation sample).
The results of the nine scenarios are shown in Figure 1. It is seen that

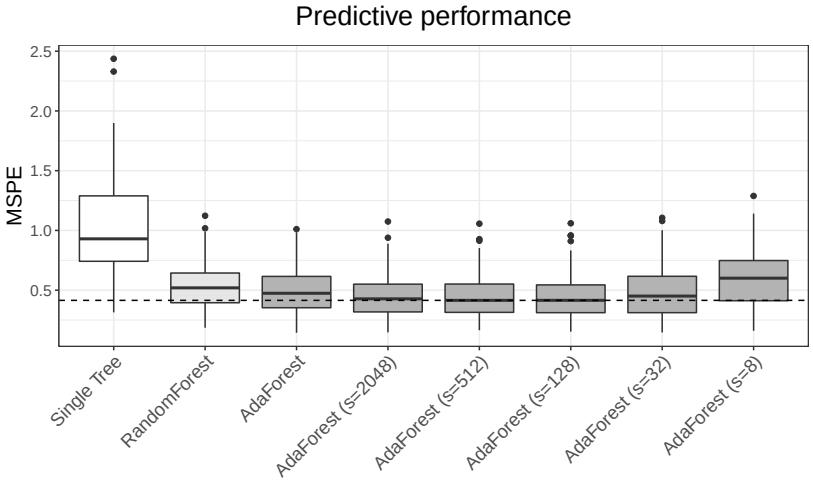


FIGURE 2. Analysis of the riboflavin data. The boxplots show the mean squared prediction errors (MSPE) when fitting a single tree, a classical random forest and the proposed AdaForest, where all (third boxplot) or only a subset of s genes (indicated in brackets) were included in the final model. The median value of the best-performing approach is marked by a dashed line.

the performance of the three approaches strongly depended on the number of informative variables. In the sparse scenarios (upper left triangle) the AdaForest algorithm clearly outperformed the classical random forest. The superiority of AdaForest, however, vanished with increasing p and increasing proportion of informative variables. Selecting a small subset of variables with the highest weights was beneficial in the three scenarios on the opposite diagonal of Figure 1.

4 Application to Gene Expression Data

As an application we consider a dataset with $n = 71$ observations on riboflavin production by Bacillus subtilis (Lee et al., 2001), where the outcome of interest is the log-transformed riboflavin production rate and the explanatory variables are given by logarithmic gene expression levels for $p = 4088$ genes. As in the simulation study, we compared various approaches with regard to their predictive performance. For this, we generated 100 subsamples, each of size $n = 47$ (containing two thirds of the data), and evaluated the MSPE on the remaining $n = 24$ observations.
Figure 2 shows the results obtained from fitting a single regression tree, a classical random forest and the AdaForest algorithm, where all $p = 4088$ genes (third boxplot) or only a subset of s genes with the highest weights (fourth to eighth boxplot) were included in the final fit. Results indicate

that AdaForest outperformed the classical random forest when incorporating all or a moderate number of genes s (for $128 \leq s \leq 2048$). The median MSPE was lowest for AdaForest with $s = 512$ included variables, while the performance of AdaForest became worse when the final model was very sparse. As expected, the single regression tree yielded by far the worst prediction accuracy. When fitting AdaForest to the entire data, the highest weights after convergence were obtained for the *yxlC* and *yxlD* genes, which were both more than 20 times the size of their starting values.

5 Conclusion

The proposed AdaForest algorithm adapts to sparsity by sequentially adjusting the weights of the candidate variables for splitting in random forests. Our simulation study and application to gene expression data show that AdaForest can outperform the classical random forest algorithm in terms of prediction accuracy, while also enabling the selection of a sparser model. Our simulations, however, also indicate that AdaForest may not always be preferred to classical random forests, particularly in less sparse settings.

References

Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I.R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **2**, 493 — 507.

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

Lee, J.M., Zhang, S., Saha, S., Santa, A.S., Jiang, C, and Perkins, J. (2001). RNA expression analysis using an antisense Bacillus subtilis genome array. *Journal of Bacteriology*, **183**, 7371 – 7380.

Linero, A.R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, **113**, 626 – 636.

Staerk, C., Kateri, M., and Ntzoufras, I. (2021). High-dimensional variable selection via low-dimensional adaptive learning. *Electronic Journal of Statistics*, **15**, 830 – 879.

Wright, M.N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, **77**, 1 – 17.

Evolutionary algorithm for the estimation of discrete latent variables models

Luca Brusa[1], Fulvia Pennoni[1], Francesco Bartolucci[2]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy
[2] Department of Economics, University of Perugia, Italy

E-mail for correspondence: luca.brusa@unimib.it

Abstract: The expectation-maximization (EM) algorithm is the most common iterative method employed for maximum likelihood estimation of discrete latent variable models. A common drawback of this estimation method, along with its variant named variational EM (VEM), is that it may be trapped into one of the multiple local maxima of the log-likelihood function. We propose a version of the algorithm based on the evolutionary approach, which allows us to explore the parameter space accurately. The proposal is validated through a Monte Carlo simulation study aimed at comparing its performance with the EM and VEM algorithms by estimating latent class, hidden Markov, and stochastic block models. Results show a significant increase in the chance of reaching a global maximum for the proposed evolutionary EM. The efficacy of the proposal is also validated by applications using longitudinal data on countries' energy production and interactions between karate club members.

Keywords: Expectation-maximization algorithm; Global optimization; Local maxima; Maximum likelihood estimation.

1 Introduction

Discrete latent variable (DLV) models have attracted much attention in statistical literature since they are formulated according to latent variables having a discrete distribution left unspecified. Among others, they ensure a high degree of flexibility in modelling complex dependence data structures (Bartolucci et al., 2022). Maximum likelihood estimation of DLV models is usually performed through the expectation-maximization (EM) algorithm (Dempster et al., 1977). When the latter approach is computationally unfeasible, a variational modification, namely the variational EM

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

(VEM) algorithm (Jordan et al., 1999), represents a popular alternative. A well-known drawback of both estimation methods is related to the multimodality of the likelihood function, resulting in a potential convergence of the algorithm to a local maximum. We propose an extension of the EM algorithm named evolutionary EM (EEM), defined according to the evolutionary algorithm (EA) approach (Ashlock, 2004). At each step of the EEM algorithm, multiple sets of parameters are evaluated according to a quality measure, while evolutionary operators, such as crossover and mutation, ensure an accurate parameter space exploration.

2 Discrete latent variable framework

The key idea of DLV models is to associate observed responses to latent variables according to a joint probability model. Denoting by $\mathbf{Y}$ and $\mathbf{U}$ the sets of observed responses and latent variables, respectively, a DLV model is characterized by the conditional distribution of the responses given the latent variables, and by the distribution of the latent variables.
The EM algorithm maximizes the observed-data log-likelihood function $\ell(\boldsymbol{\theta})$, expressed in terms of model parameters $\boldsymbol{\theta}$, relying on the complete-data log-likelihood function $\ell^*(\boldsymbol{\theta})$. Once the model parameters have been initialized, the algorithm alternates two steps until convergence: (i) an expectation step, where the conditional expected value of $\ell^*(\boldsymbol{\theta})$ is computed given the value of the parameters at the previous step and the observed data, and (ii) a maximization step, where the model parameters are updated by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$.
The VEM algorithm defines instead a lower bound $\mathcal{J}(\boldsymbol{\theta})$ for the observed-data log-likelihood function, to be maximized instead of $\ell(\boldsymbol{\theta})$. To explore the parameter space, the choice of multiple sets of starting values for the model parameters is crucial. The maximum is then taken as the solution corresponding to the largest likelihood value at convergence. Drawbacks of this strategy are the high computational time and the fact that the convergence may be to one of local maxima different from the global one.

3 Evolutionary expectation-maximization algorithm

Following the EA approach, the proposed EEM algorithm is inspired by the Darwinian theory of evolution principles. According to Pernkopf and Bouchaffra (2005), it takes into account an initial "population" P_0 of N_P potential solutions for the optimization problem at issue. Each element of P_0 is a different candidate array of posterior probabilities. The following steps are then alternated until convergence:

1. $P_1 \leftarrow \mathbf{Update}(P_0)$: population P_0 is updated by performing a small number of cycles of the standard EM algorithm with random initialization on each individual, resulting in a new population P_1.

2. $P_2 \leftarrow$ **Crossover**(P_1): pairs of individuals from population P_1 are randomly selected and recombined by swapping corresponding blocks of their arrays. We obtain the N_O offspring of the new population P_2.

3. $P_3 \leftarrow$ **Update**(P_2): population P_2 is updated by performing a small number of cycles of the standard EM algorithm with random initialization on each individual, resulting in the new population P_3.

4. $P_4 \leftarrow$ **Selection**$(P_1 \cup P_3)$: individuals from populations P_1 and P_3 are considered jointly, and the N_P with the highest value of the log-likelihood function are selected for the next generation P_4.

5. $P_5 \leftarrow$ **Mutation**(P_4): variation is introduced to each individual of population P_4 (apart from the best one): given a row of the corresponding array of posterior probabilities, mutation operator swaps the highest value with a random one.

Convergence of the EEM algorithm is measured focusing only on the best solution of population P_4 and analyzing both the relative difference of the log-likelihood of two consecutive steps and that between the corresponding parameter vectors.

4 Simulation studies

To evaluate the performance of the EEM algorithm, we rely on a Monte Carlo simulation study considering latent class (LC), hidden Markov (HMcat and HMcont for categorical and continuous response variables, respectively), and stochastic block (SB) models. This study is based on different scenarios for each model, depending on several features: sample size ($n = 500, 1000$), number of response variables ($r = 6, 12$), response categories ($c = 3, 6$), time occasions ($T = 5, 10$), and latent components ($k = 3, 6$). Concerning the SB model we also distinguish two different behaviors: one defined as assortative with high intra-group and low inter-group connection probabilities and the other as disassortative with low intra-group and high inter-group probabilities. For each scenario the corresponding model is applied 100 times to 50 samples using the EM and EEM algorithms. Both correctly specified and misspecified latent structures are estimated in order to compare the performance of the algorithms through the following criteria.

Global maximum achievement: considering the highest of the maximized log-likelihood values as the global maximum $\hat{\ell}_{\text{MAX}}$, we denote a generic log-likelihood value at convergence as $\hat{\ell}$ and compute the percentage of $\hat{\ell}$ such that $(\hat{\ell}_{MAX} - \hat{\ell})/|\hat{\ell}_{MAX}| < \tilde{\varepsilon}$, where $\tilde{\varepsilon}$ is a suitable threshold. The EEM algorithm performs better in each simulated scenario, significantly increasing the chance to reach the global maximum. Some results of 2 of

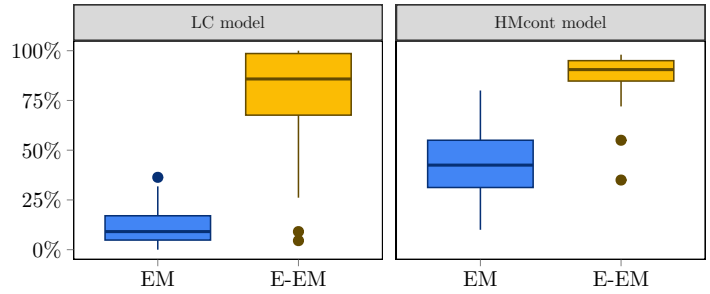


FIGURE 1. Percentages of global maxima reached using EM and EEM algorithms for (i) a correctly specified LC model with $n = 500$, $r = 6$, $c = 3$, and $k = 6$, and (ii) a misspecified HMcont model with $n = 500$, $r = 12$, $T = 5$, and $k = 3$.

the 22 simulated scenarios are depicted in Figure 1. In particular, regarding the estimation of models whose latent structure is correctly specified, the frequency of convergence to the global mode is usually very close to 100%, highlighting that it generally tends to avoid convergence to a local maximum of $\ell(\boldsymbol{\theta})$. The results of the extensive simulation study highlight that the proposal always outperforms the EM algorithm; the improvement is especially evident with many latent components and under scenarios related to the SB model. Its performance is even more remarkable considering models with misspecified latent structures. In this case, while the standard EM algorithm sometimes proves unable to locate the global maximum, the evolutionary approach is always able to correctly detect it, improving the value itself of the global mode, in addition to the chance to reach it.

Average distance from the global maximum: using the EEM algorithm, the distance between each maximum and the global one is quite low for all the examined scenarios. The average distance obtained through the EM algorithm is usually considerably higher. We mention for instance one scenario of the LC model in which the average distance decreases from $4.7 \cdot 10^{-7}$ using the EM algorithm to $2.5 \cdot 10^{-18}$ using the EEM algorithm. In scenarios related to the HMcat model the distance is still reduced by half with the EEM algorithm, dropping, for example, from $1.2 \cdot 10^{-3}$ to $6.8 \cdot 10^{-4}$.

Accurate parameters estimation: dealing with correctly specified models, we also provide the root mean square error (RMSE) between the true and estimated model parameters. Results show the RMSEs obtained with the EEM algorithm are very close to zero under all the simulated scenarios; on the contrary, values obtained with the EM algorithm are always larger, approaching one in some cases. This shows that the evolutionary approach entails a significantly greater accuracy. In particular, the improvement is especially evident when the HMcont and SB models are estimated.

5 Applications

The EEM algorithm is also evaluated to estimate LC, HMcat, HMcont, and SB models with cross-sectional, longitudinal, and network data.
In the following, as a first application, we use longitudinal data measuring the sources of electricity generation in 27 European Union countries (data are available at the link https://ourworldindata.org/energy). A multivariate time homogeneous HMcont model is considered for response variables collected yearly from 2011 to 2020 and referred to the share of electricity deriving from biofuel, coal, natural gas, hydroelectric, nuclear, oil, solar, and wind. Logit and Box-Cox transformations are applied to all the variables. The model is estimated for a number of states ranging from 1 to 12 with both the EM and EEM 100 times. A model with 8 latent states representing sub-populations of countries with similar energetic behaviour is selected according to the Bayesian information criterion. The EEM algorithm ensures convergence to the global maximum, corresponding to a value of the log-likelihood function equal to $-5,452$. The EM algorithm never detects such a maximum, providing $-5,574$ as the highest value for the log-likelihood function at convergence. The estimation with the EEM also provides a reasonable posterior dynamic classification of the countries into groups, while EM does not. Table 1 reports the estimated conditional means of the responses given the latent state. Groups are ordered from the lowest to the highest average value of wind power. Countries in the 1st group are using mainly nuclear power, in the 2nd are predominantly coal-dependent, in the 3rd heavily rely on oil, in the 4th they use a mix of coal, oil and gas, along with the highest average of solar energy. Countries in the 5th state are using mainly gas, in the 6th they use gas and a quota of biofuel over all the other groups, in the 7th they excels in hydroelectric power, and in the 8th they use mainly wind energy along with nuclear power.
As a second application, we estimate the SB model with network data on 34 karate club members (data are available in the `R` package `igraphdata`).

TABLE 1. Estimated means of the HMcont model with $k = 8$ latent states for the European Union countries electricity data.

	Latent states							
Source	1	2	3	4	5	6	7	8
Coal	5.58	41.91	4.52	38.52	14.86	0.00	20.57	15.10
Oil	3.62	2.85	51.80	10.77	6.62	3.97	3.55	3.04
Gas	5.34	12.24	7.20	24.53	57.43	44.52	21.06	19.87
Nuclear	50.21	19.96	0.00	0.00	1.42	0.00	13.20	31.94
Biofuel	7.68	4.55	4.77	0.53	3.17	12.35	2.92	9.65
Hydro	22.12	9.85	24.50	8.96	1.12	21.16	19.05	0.39
Solar	0.90	3.41	1.70	6.62	2.51	3.34	2.89	2.48
Wind	4.52	4.87	5.48	10.08	12.87	14.66	16.63	17.53

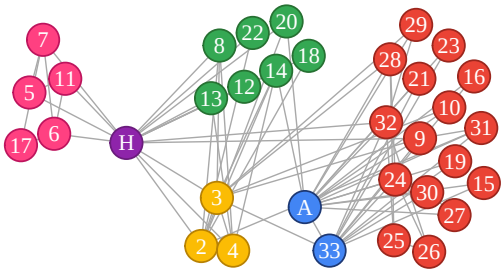


FIGURE 2. Graph visualization with nodes colored by estimated partition for the SB model with $k = 6$ latent blocks for the karate club data.

Relationships among members are measured by a 34×34 adjacency binary matrix. Using the EEM algorithm, an SB model with $k = 6$ latent blocks is selected according to the integrated classification likelihood criterion. The EEM algorithm consistently converges to a log-likelihood function value equal to -277.91; if the model is estimated with the EM algorithm its highest value is -316.46. Figure 2 shows the network with nodes colored by the estimated partition. The model correctly identifies positions taken for president (**A**, in blue) or instructor (**H**, in violet). The faction led by the president consists of a single additional latent block (in red), presenting a high connection probability with its leader (equal to 0.75). The remaining three latent blocks (depicted in pink, green, and yellow) constitute the faction led by the instructor; each of these blocks has a high connection probability with their leader (equal to 0.81, 1.00, and 1.00, respectively). Connection probabilities between blocks of different factions are very low (0.17 at most).

References

Ashlock, D. (2004). *Evolutionary Computation for Modeling and Optimization*. New York: Springer.

Bartolucci, F., Pandolfi, S. and Pennoni, F. (2022). Discrete Latent Variable Models. *Annual Review of Statistics and its Application*, **6**, 1 – 31.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, **37**, 183 – 233.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1 – 38.

Pernkopf, F. and Bouchaffra, D. (2005). Genetic-Based EM Algorithm for Learning Gaussian Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1344 – 1348.

Coherent cause-specific mortality forecasting via constrained penalized regression models

Carlo G. Camarda[1], María Durbán[2]

[1] Institut National d'Études Démographiques, Aubervillers, France
[2] Department of Statistics, Universidad Carlos III de Madrid, Spain

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

Abstract: Cause of death data provides additional insight on the future trends of mortality, as well as provide valuable information for governments and insurance companies. Models that fit and forecast by cause of death come across several methodological problems, one of them being the inconsistency between individual estimation and forecast of mortality per cause of death and an all-cause scenario. We propose a clear-cut and fast method to obtain coherent cause-specific mortality trajectories based on Lagrange multipliers. We apply the method proposed to fit and forecast mortality of males in USA for the most five leading causes of death.

Keywords: Cause of death; Constraints; Forecasting; Mortality; Penalized Likelihood.

1 Introduction

Overall mortality trends are the summation of cause-specific mortality experiences. Consequently modelling and forecasting changes in cause of death patterns allows us to recognize the drivers of all-cause mortality and identify emerging health challenges. On the one hand, early literature has argued that all-cause mortality projections based on cause-specific mortality present serious drawbacks (Wilmoth, 1995). On the other, some approaches for forecasting cause-specific mortality has been recently proposed, though either based on the Lee-Carter model and for specific cause (Kjærgaard et al., 2019) or on a Bayesian hierarchical model aiming to forecast cause-specific death rates for geographic subunits (Foreman et al., 2017).
When dealing with cause-specific mortality, we need to ensure that cause-specific deaths must sum to the total number of deaths. In the following,

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

we model log-mortality in a Poisson setting for each cause assuming only smoothness over age and time. The summation constraint thus become non linear with respect to the estimated coefficients. We embed the whole approach in a Generalized Linear Array Model framework (GLAM, Currie et al., 2006) in which Lagrangian multipliers are iteratively updated to enforce constraints.

2 The model

We have deaths, and exposures to the risk of death, arranged in two three-dimensional arrays, $\boldsymbol{Y} = (y_{ijk})$ and $\boldsymbol{E} = (e_{ijk})$, each $m \times n \times k$, whose rows, columns and layers are classified by age at death ($\boldsymbol{a}$), year of death ($\boldsymbol{t}$) and cause of death ($\boldsymbol{c}$). For ensuring coherence in the model, the final layer of $\boldsymbol{Y}$ contains total number of deaths (that is, k =number of causes of death+1). Note that each layer in $\boldsymbol{E}$ includes the same age-year matrix: we are in a competing risk setting. We assume that the number of deaths y_{ijk} is Poisson distributed with mean $\mu_{ijk}e_{ijk}$. The value of μ_{ijk}, commonly named force of mortality, is the object of all mortality models.
In the following we will illustrate the method for United States, males by age-groups 30-34, ..., 95-99, 100+, years 1978-2018 and the following five coherent groups of causes of deaths: Cardio-vascular diseases, Neoplasms, External causes, Diseases of the respiratory system and Other diseases (Human Cause-of-Death Database, 2023). We forecast total and cause-specific mortality up to 2040.
With $k-1$ causes of death, we deal with a three-dimensional setting. The vectorized linear predictor is given by $\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\alpha}$ where the design matrix is $\boldsymbol{B} = \boldsymbol{I}_k \otimes \boldsymbol{B}_t \otimes \boldsymbol{B}_a$. We use a rich basis of B-splines for age and year, and smooth surfaces are then obtained by marginal penalization. With no summation constraint the model simply reduces to a series of two-dimensional GLAMs.
To hinder singularity issues in the resulting scoring algorithm, we enforce our constraint for a large number of equally-spaced data-points. Smoothness will guarantee the remaining coherence between cause-specific and overall mortality. The penalized 3D GLAM is then subject to

$$\boldsymbol{C}\exp(\boldsymbol{\eta}) = \boldsymbol{0} \tag{1}$$

where $\boldsymbol{C}$ sums up, for the large selected number of age and years, cause-specific deaths and then subtract the associated total number of deaths. The matrix $\boldsymbol{C}$ can be written as a Kronecker product $\boldsymbol{C} = \boldsymbol{C}_c \otimes \boldsymbol{C}_t \otimes \boldsymbol{C}_a$ and therefore, as for the linear functions and the inner products within the scoring algorithm, it can be included as a sequence of nested matrix operations in a GLAM framework.
The use of Lagrange multipliers $\boldsymbol{\omega}$ for each age-year ensures that the constraint is enforced, yielding the following constrained penalized Poisson

log-likelihood:

$$\ell_P = \boldsymbol{y}'\boldsymbol{B}\boldsymbol{\alpha} - \boldsymbol{e}'\exp(\boldsymbol{B}\boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{P}\boldsymbol{\alpha} - \boldsymbol{\omega}'\boldsymbol{C}\boldsymbol{E}\exp(\boldsymbol{B}\boldsymbol{\alpha}). \tag{2}$$

We compute the derivatives of (2), and by means of Newton-Raphson we find the following scoring algorithm:

$$\begin{bmatrix} \boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P} + \boldsymbol{B}^{\mathrm{T}}\mathtt{diag}(\boldsymbol{C}^{\mathrm{T}}\tilde{\boldsymbol{\omega}})\tilde{\boldsymbol{V}}\boldsymbol{B} & \boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{V}}\boldsymbol{C}^{\mathrm{T}} \\ \boldsymbol{C}\tilde{\boldsymbol{V}}\boldsymbol{B} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\alpha}} \\ \tilde{\boldsymbol{\omega}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}^{\mathrm{T}}\tilde{\boldsymbol{W}}\boldsymbol{z} + \boldsymbol{B}^{\mathrm{T}}\mathtt{diag}(\boldsymbol{C}^{\mathrm{T}}\tilde{\boldsymbol{\omega}})\tilde{\boldsymbol{V}}\boldsymbol{B}\tilde{\boldsymbol{\alpha}} \\ \boldsymbol{C}\tilde{\boldsymbol{V}}\boldsymbol{B}\tilde{\boldsymbol{\alpha}} - \boldsymbol{C}\boldsymbol{\gamma} \end{bmatrix} \tag{3}$$

where $\boldsymbol{\gamma} = \exp\boldsymbol{\eta}$, $\boldsymbol{V} = \mathtt{diag}(\boldsymbol{\gamma})$. $\boldsymbol{W}$ and $\boldsymbol{z}$ are the Poisson regression weights and working response, respectively. The penalty $\boldsymbol{P}$ ensures smoothness over age and time for each cause and it has a block-diagonal structure. In order to handle $k-1$ causes of death across different age groups and years, we face the challenge of optimizing $2 \cdot k$ smoothing parameters. To avoid dealing with such a high-dimensional optimization problem, we decided to utilize the smoothing parameters optimized by BIC when estimating each cause-specific age-time matrix independently.
Confidence intervals for the estimated mortality surface are calculated by stepping into the Bayesian framework, therefore:

$$\mathrm{Var}(\boldsymbol{B}\hat{\boldsymbol{\alpha}}) = \boldsymbol{B}\boldsymbol{R}\boldsymbol{B}^{\mathrm{T}},$$

where $\boldsymbol{R}$ is the top left block of the inverse of the matrix on the left hand side of (3). Of course, if there are no constraints, i.e $\boldsymbol{\omega} = \boldsymbol{0}$, we obtain the usual expression of the variance.
As in Currie et al. (2004) we treat forecasting as a missing value problem and we add shape constraints to enforce future cause-specific mortality patterns to lie within a range of valid profiles computed from observed trends (Camarda, 2019).

3 Results

We fitted the proposed model to the US male mortality data described in the previous section. The left panel of Figure 1 shows actual, estimated and forecast log-mortality for a selected age (50) over years along with their 95% confidence intervals. The proposed model is able to well described historical cause-specific patterns as well as to reasonably extrapolate them into the future. The right panel of Figure 1 presents estimated death counts for a specific year (2000). Here one can easily acknowledge the equality between the sum of cause-specific deaths and the total number of deaths which is enforced by the non linear set of constraints in (1). Equally satisfactory outcomes are achieved for all ages, years and causes of deaths.

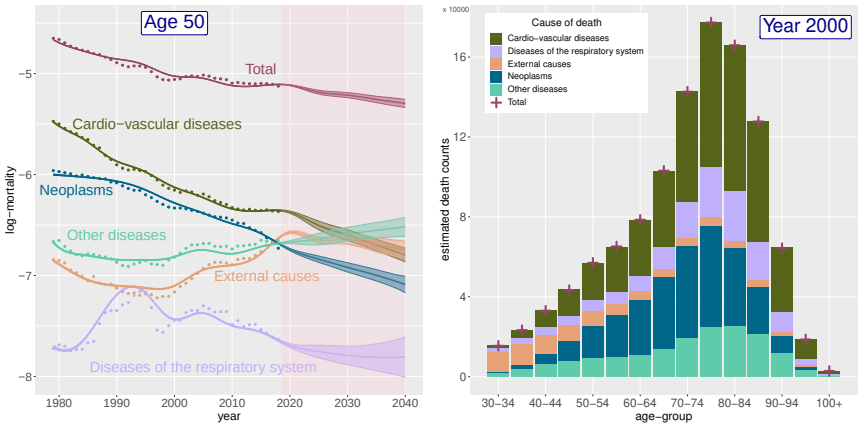


FIGURE 1. Left panel: observed, estimated and forecast cause-specific mortality rates over years for age 50 (log-scale) along with associated 95% confidence intervals. Right panel: Estimated cause-specific death counts over age-groups for year 2000. In both panels, total mortality and total number of deaths are plotted. USA, males, ages 30-100, years 1978-2018, forecast up to 2040.

In Figure 2, a commonly used summary indicator is presented: life expectancy, here at the starting age of 30. We compare our model with a estimates obtained on overall mortality using P-splines with shape constraints (Camarda, 2019). While the fitted values for the observed time-window are practically identical, noticeable differences emerge when making forecasts. When accounting for cause-specific patterns and ensuring that cause-specific deaths add up to the total number of deaths, the modeling of mortality yields slightly more pessimistic prospects: remaining life expectancy in 2040 at age 30 is forecast to 50.67 years in our model compared to 51.16 years in the alternative approach.
Of even greater significance is the notable reduction in the confidence intervals, indicating a significant decrease in future uncertainty around overall mortality when cause-specific trends are incorporated into the model. This outcome was expected, given the inclusion of substantial information through the incorporation of what we referred to as the summation constraint.

4 Conclusions

In the presented study, we propose a novel approach to model and forecast cause-specific mortality. By combining penalized likelihood and iteratively computed Lagrangian multipliers we obtain smooth cause-specific mortality surfaces over ages and time, and we simultaneously enforce the necessary constraints in this setting: cause-specific deaths sum to the total number

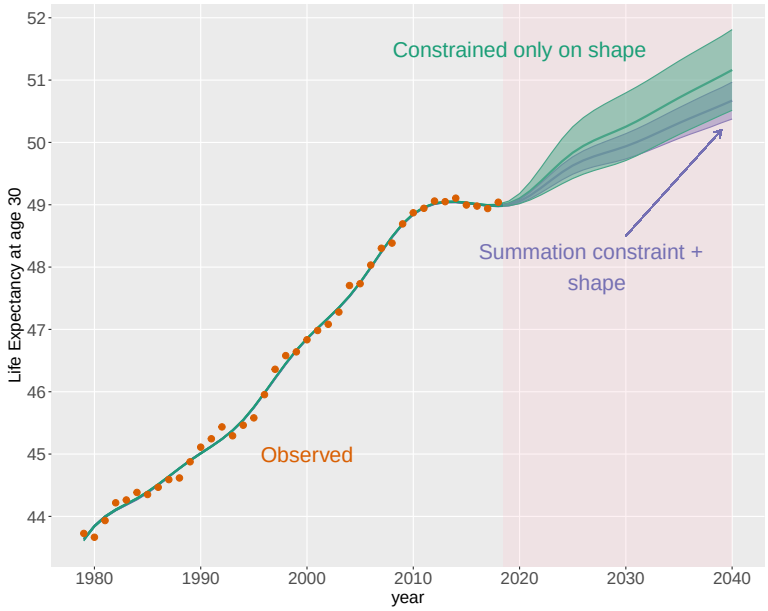


FIGURE 2. Observed, estimated and forecast life expectancy at age 30 along with associated 95% confidence intervals. Proposed model estimating cause-specific and overall patterns with summation and shape constraints is compared with a simpler approach without summation constraint. USA, males, ages 30-100, years 1978-2018, forecast up to 2040.

of deaths. Forecasting comes naturally in this setting and constraints are satisfied into future years, too. Additional shape are necessary to demographically informed projected patterns.

In the example analysis of US mortality, focusing on the five leading causes of death, the proposed model yields lower overall future life expectancy compared to similar models. Additionally, it exhibits remarkably narrower future uncertainty.

Moving forward, we intend to investigate other scenarios where coherence constraints may be required, such as mortality by region, sex, and other factors. Furthermore, we have plans for a comprehensive validation study to evaluate the accuracy of future point estimates and the coverage of associated confidence intervals.

Last, Covid-19 pandemic clearly taught us that forecast future mortality by extrapolating past trends can no longer be tacitly assumed. Accounting for the impact of such short-term shocks will be a clear challenge for further mortality forecasting research as well as for our model.

References

Camarda, C. G (2019). Smooth constrained mortality forecasting. *Demographic Research*, **41**, 1091 – 1130.

Currie, I.D., Durban, M. and Eilers, P.H.C. (2004). Smoothing and Forecasting Mortality Rates. *Statistical Modelling*, **4**, 279 – 298.

Currie, I.D., Durban, 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.

Foreman, K. J., Li, G., Best, N. and Ezzati, M. (2017). Small Area Forecasts of Cause-Specific Mortality: Application of a Bayesian Hierarchical Model to US Vital Registration Data. *Journal of the Royal Statistical Society: Series C*, **66**, 121 – 139.

Kjærgaard, S., Ergemen, Y.E., Kallestrup-Lamb, M. et al. (2019). Forecasting causes of death by using compositional data analysis: the case of cancer deaths. *Journal of the Royal Statistical Society: Series C*, **68**, 1351 – 1370.

Human Cause-of-Death Database (2023). French Institute for Demographic Studies (France) and Max Planck Institute for Demographic Research (Germany). Available at `www.causeofdeath.org`. Data downloaded on February 2023.

Wilmoth, J.R. (1995). Are mortality projections always more pessimistic when disaggregated by cause of death? *Mathematical Population Studies*, **5**, 293 – 319.

The influence of resolution on the predictive power of spatial heterogeneity measures as a biomarker of disease severity

Jari Claes[1], Annelies Agten[1], Alfonso Blázquez-Moreno[2], Marjolein Crabbe[3], Marianne Tuefferd[2], Hinrich Goehlmann[2], Helena Geys[3], Thomas Neyens[1,4], and Christel Faes[1]

[1] Data Science Institute, UHasselt - Hasselt University, Agoralaan 1, BE 3590 Diepenbeek, Belgium.
[2] Translational Biomarkers, Infectious Diseases, Janssen Research and Development, Turnhoutseweg 30, 2340 Beerse, Belgium.
[3] Discovery Statistics, Global Development, Janssen Research and Development, Turnhoutseweg 30, 2340 Beerse, Belgium.
[4] L-BioStat, KU Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium.

E-mail for correspondence: `jari.claes@uhasselt.be`

Abstract: Spatial heterogeneity of cells in liver biopsies can be used as biomarker for disease severity of patients. This heterogeneity can be quantified by non-parametric statistics of point pattern data, which make use of an aggregation of the point locations. The method of aggregation, however, is not standardized but generally chosen by the author, as dimensions, scale or other specifics of a problem tend to call for methods tailored to them. Increasing spatial resolution will not endlessly provide more accuracy, as limiting each grid cell to only a few data points might not yield as much information about heterogeneity between species in the sample. The question then becomes how changes in grid choice influence heterogeneity indicators derived from this grid aggregation, and subsequently how they influence predictive abilities of these indicators. The aim of this paper is to analyze this issue by evaluating different heterogeneity indicators on the predictive performance.

Keywords: point pattern; penalized regression; spatial resolution.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

1 Methods

1.1 Data

The data used in this analysis contained 110 liver biopsies that were collected from chronic hepatitis B patients. Immunofluorescent staining and subsequent cell segmentation was applied to the biopsies to identify each cell's two-dimensional (x and y) coordinates and the cell type. Cell types include HBsAg-positive, HBsAg-negative, or immune cells. HBsAg-negative cells will be referred to as type 1, immune cells as type 2 and HBsAg-positive cells as type 3. For each patient, a pathologist assessed the fibrosis stage of the liver, resulting in a a categorical score ranging from 0 to 4. Some scores were uncertain, and are available as interval censored observations.

1.2 Penalized ordinal regression model and used covariates

Spatial measures are used to quantify heterogeneity amongst biological cells, hereafter referred to as points. Commonly used measures of spatial heterogeneity of point pattern data are the Morisita-Horn index, the Shannon diversity index and the Getis-Ord hotspot analysis. These measures are based on a discretization of the spatial point pattern data by dividing the window into grid cells of the same size. In the following formulations, i denotes the index of a grid cell, c_i^k denotes the number of points of type k in grid cell i, p_i^k denotes the ratio of c_i^k to the total number of points in grid cell i, and n^k denotes the total number of points of type k. The Morisita-Horn index (MHI) by Horn (1966) was used as a *global* spatial summary statistic. The formula of the index is defined as

$$M = \frac{2\sum_i \frac{c_i^k}{n^k}\frac{c_i^l}{n^l}}{\sum_i (\frac{c_i^k}{n^k})^2 + \sum_i (\frac{c_i^l}{n^l})^2}.$$

The index is a measure calculating the heterogeneity of two point types k and l given their spatial distribution in the grid. This index ranges from 0 to 1, indicating complete homogeneity at value 0, and complete heterogeneity at value 1. Since this is a pairwise statistic, this was done for each pairwise combination of the three point types studied in this paper.

The Shannon diversity index (SDI), analogous to the Shannon (1948) entropy in information theory, was also used as a *global* spatial summary statistic and is an index measuring diversity per grid cell. The formula of the index is defined as

$$S_i = -\sum_l^3 p_i^l \log(p_i^l).$$

This index was calculated for each grid cell individually and thus requires to be summarized to give information about the entirety of the sample. Both the mean and variance of Shannon indices across all grid cells are used as a summary measure of heterogeneity.

The Getis-Ord hotspot analysis by Getis and Ord (1992) was used as a *local* spatial summary statistic. Formally, in grid cell i, the z-score is defined as:

$$z = \frac{\sum_{j=1}^{N} w_{i,j} c_j^l - \overline{c}^l \sum_{j=1}^{N} w_{i,j}}{SU},$$

with normalizing factors

$$S = \sqrt{\frac{\sum_{j=1}^{N} (c_j^l)^2}{N} - (\overline{c}^l)^2},$$

$$U = \sqrt{\frac{N \sum_{j=1}^{N} w_{i,j}^2 - (\sum_{j=1}^{N} w_{i,j})^2}{N-1}},$$

where $w_{i,j}$ is the binary indicator whether grid cell i and j are neighbors, $\overline{c}^l$ is the mean of the points in the grid cells and N is the total amount of non-empty grid cells. In its algorithm, a z-score is calculated comparing the count in a grid cell and its neighboring grid cells (depending on a chosen neighborhood structure) with the expected number of points. A significantly high and low z-score were defined respectively as a hotspot or coldspot. Both ratios of total hot- or coldspots of each type as well as total colocalized hot- or coldspots of different types were used a summary measure of the entire sample.

The choice of grid cell size influences the discretization process and therefore the summary statistics used in this study as specified below. The size of each grid cell was therefore investigated, and chosen for each biopt separately so that the mean amount of points per grid cell approaches the same value (pre-specified values are 5, 10, 15, 20, 25, 50, 100, 200, 500 and 1000).

The predictive performance of both the spatial (*local* and *global*) and non-spatial (i.e. percentage of points of a specific type within the whole sample) summary measures towards the fibrosis stage were subsequently investigated, using an ordinal regression model with L1 type penalization. The OrdinalNet package, made by Wurm et al. (2017), was used in R to construct a forward cumulative probability logit model, with formula

$$\text{logit}(P(Y_i \leq m | x_i)) = \alpha_m + x_i^T \beta,$$

where Y_i is the fibrosis score as ordinal category ranging from 0 to 4 for observation $i = 1, ..., N$ with N being the amount of observations, m is one of

said five ordinal categories, x_i is the vector of covariate values with length P, where P is the amount of covariates, α_m is the outcome specific intercept and β is the vector of weights corresponding to the covariates, with length P. A Least Absolute Shrinkage and Selection Operator (LASSO) type penalization was added by adding a term directly related to the magnitude of covariate weights to the likelihood function to be minimized, inhibiting covariate influence. The formula of the to be minimized function is then

$$M = -\frac{l}{N} + \lambda \sum_{j=1}^{P} |\beta_j|,$$

where l is the loglikelihood of the cumulative probability logit model, N is the number of samples, λ is a tuning parameter, P is the total number of covariates and β_j is the weight corresponding to covariate j. The OrdinalNet package uses a coordinate descent algorithm, providing an initial output of a selection of models ranging from low to high covariate usage, and the AIC minimized model was considered in the following analyses. Interval censoring of disease severity was accounted for in the model, using a weighted likelihood approach.
The performance of different model subtypes were investigated: a model including all covariates (M1), a model with only local covariates (M2), with only global and nonspatial covariates (M3) and a model with only local and global covariates (M4).

1.3 Simulated data

To validate the proposed methods and illustrate the influence that grid cell sizes have on these methods, a simulation study was performed. Similar to the data, point patterns were simulated corresponding to five categories of disease stage. For each category, 20 point patterns were simulated for each of the categories according to a clustering algorithm with 'parent points' for each point type, and with each category having some increasing or decreasing setting related to the diversity of point types.

2 Results & Discussion

The predictive ability of the fitted OrdinalNet models was measured by a leave-one-out-cross-validation (LOOCV) for each combination of model type (M1-M4) and grid choices with which the covariates were calculated. Figure 1 shows the results of the predictive power of the models.
In both real and simulated data, the model subtype where only local covariates were included generally displayed a downward trend in predictive ability as grid cell size increased. This trend was also reflected in the real data in the dwindling importance of the local covariates due to a lack of information at larger grid sizes. Even models M1, M3 and M4 had a drop-off in predictive quality in the largest choice(s) of grid size. While for local measures a smaller grid-size outperforms larger grid-sizes, global measures have a better performance with medium-sized grids. In conclusion, across all data and models, trends are not unambiguously apparent, and there is no 'one-size-fits-all' solution. We do recommend the use of small grid sizes (on average 5-10 points per grid cell) when using the local measure, as it quantifies locally the heterogeneity of point types as well as its interaction amongst point types. A medium grid size is recommended for global indicators (20-30 points per grid cells), which is required for a more stable estimation of the global measure. The use of both global and local measures of heterogeneity improve the predictive performance.

3 Acknowledgements

The authors extend their graditude towards Dr. Cheng-Yan Peng and the China Medical University Hospital, Taichung, Taiwan for the provision of the data used in this paper.

References

Getis, A., and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3), 189 – 206.

Horn, H. S. (1966). Measurement of "overlap" in comparative ecological studies. *The American Naturalist*, 100(914), 419 – 424

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379 – 423.

Wurm, M. J., Rathouz, P. J., Hanlon, B. M. (2017). Regularized ordinal regression and the ordinalNet R package. *arXiv preprint*, arXiv:1706.05003.

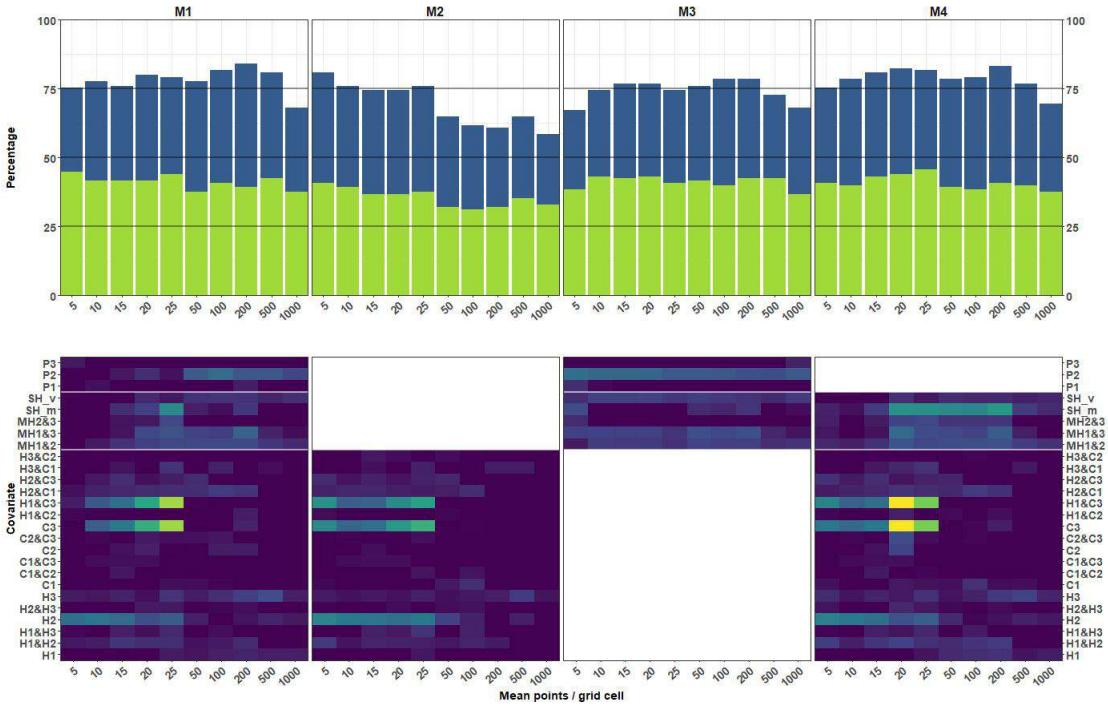


FIGURE 1. Performance and variable importance of actual data as a function of the grid size. Columns indicate the results from the model where respectively all (M1), only local (M2), global and nonspatial (M3), and local and global covariates (M4) were included. The top panel shows LOOCV results, with the perfect prediction % in green and one-off % cumulatively in green and blue with black lines indicating 25, 50 and 75%. The bottom panel shows variable importance, where from top to bottom respectively nonspatial, global and local covariates are shown and divided by white lines. Y axis label Pi refers to proportion of type i, SH_v and SH_m refer to variance and mean of the Shannon indices, MHi&j refers to the pairwise Morisita-Horn index of types i and j, Hi and Ci refer to hotspots and coldspots of type i, e.g. H1&C2 refers to the colocalization of type 1 hotspots and type 2 coldspots.

A multi-state 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] Department of Statistical Science, University College London, UK
[2] Department of Applied Health Research, University College London, UK

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

Abstract: The natural history of prostate cancer can be modelled using a continuous Markov model and can be used to evaluate screening strategies. The mean time between the onset of the preclinical screen-detectable cancer and the onset of the clinical state; the mean sojourn time, can be estimated. The data used are from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, and the methods develop will account for the use of interval-censored, left-truncation, and right-censored data. The model will include longitudinal PSA measures from the screened population.

Keywords: Survival analysis; Continuous Markov model; Cancer screening.

1 Introduction

Specifically for cancer settings and chronic disease in general, the sojourn time, the time spent in the detectable preclinical phase (PCDP); can be informative for the design of screening programs.
Prostate cancer is the second most commonly diagnosed cancer in men. The prostate-specific antigen (PSA) test measures the PSA levels in the blood. A PSA level above 3.0 or 4.0 ng/mL can be considered abnormal. However, different factors can cause the PSA level to fluctuate. The PSA levels can be used to assess if further clinical diagnosis is required; i.e. imaging or biopsy studies. Screening based on the PSA test can lead to overdiagnosis and false-positive screening results.
In this study, we consider panel data, a series of observations for each individual at a sequence of time points, where the sampling times vary across individuals. The data are from the (PLCO) Cancer Screening Trial. Bhatt et al (2021) developed a multi-state model for the natural history of prostate cancer, that accounts for interval-censored, left-truncated, and

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

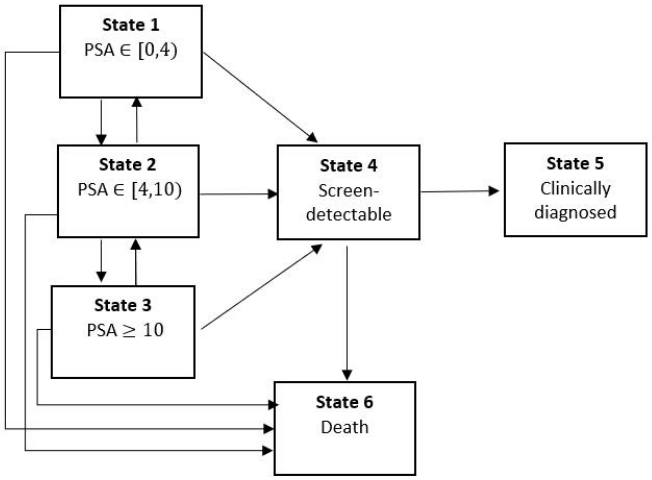


FIGURE 1. Six-state progressive model for prostate cancer. The first three states are defined by discretising the PSA scale.

right-censored data. It also incorporates age-varying hazards and misclassification, using data on nonscreened participants. The aim is to extend the model by Bhatt et al (2021) by including the PSA test level to inform screening frequency and to assess the outcome of screening in detecting cancer and preventing cancer death. The process is illustrated in Fig. 1.
As a first approach, we will impute the first PSA value to the control group by sampling from the PSA distribution from the screened group. Further in the project, we will explore different methods to take into account the first observed state for the control group. Furthermore, the methods will account for misclassification and mismeasured data.

2 Methods

2.1 Multi-state model

Let $\{Y_t \mid t \in (0,\infty)\}$ be a continuous Markov chain on the state space S, and let $P(t,u)$ be the $D \times D$ transition probability matrix with entries $p_{rs}(t,u) = P\left(Y_t = s \mid Y_u = r\right)$, for $0 \leq t \leq u$, and $r, s = 1 \ldots D$.
The transition probabilities can be derived from the transition hazards, $P(t,u) = \exp((u-t)Q(t))$, where Q is the $D \times D$ generator matrix with off-diagonal (r,s) entries q_{rs} and diagonal entries $q_{rr} = -\sum_{s \neq r} q_{rs}$.
The transition-specific hazards can be defined by combining a baseline hazard with a log-linear regression

$$q_{rs}(t \mid x) = q_{rs.0}(t) \exp\left(\beta_{rs}^{\top} x\right) , \tag{1}$$

where $\beta_{rs} = (\beta_{rs.1}, \ldots, \beta_{rs.p})^\top$ is a parameter vector, and $x = (x_1 \ldots x_p)^\top$ is the covariate vector. The baseline hazard $q_{rs.0}(t)$ describes the hazard's time dependency.

2.2 Estimation

Considering interval-censored observations, the maximum likelihood is constructed with the transition probabilities (Kalbfleisch and Lawless, 1985). Assuming the transition time into the death state D is known, and letting the living states be indexed by $1, 2 \ldots D-1$. For an individual i with observation times $t_1, \ldots, t_n$, and an observed trajectory of states $y_1, \ldots, y_n$, the likelihood contributions for interval the (t_{j-1}, t_j) is given by

$$= \left(\prod_{j=2}^{J-1} P\left(Y_j = y_j \mid Y_{j-1} = y_{j-1}, \theta, x\right) \right) C\left(y_J \mid y_{J-1}, \theta, x\right) ,$$

where θ is the vector of parameters, and x is the covariate vector. If t_j is a living state then $C\left(y_j \mid y_{j-1}, x\right) = P\left(Y_j = y_j \mid Y_{j-1} = y_{j-1}, \theta, x\right)$, and if death is observed at t_j then

$$C\left(y_j \mid y_{j-1}, x\right) = \sum_{s=1}^{D-1} P\left(Y_j = s \mid Y_{j-1} = y_{j-1}, \theta, x\right) q_{sD}\left(t_{j-1} \mid \theta, x\right) .$$

The probability transition matrix is computed using the eigenvalue-decomposition method, and the log-likelihood is maximised using the Nelder-Mead optimisation method from the general purpose optimiser `optim`, R-package.

2.3 Single imputation

Imputation is a versatile method for handling missing data. There are two generic approaches; explicit and implicit modelling. Explicit modelling assumes that the predictive distribution for the imputation values follows a known probability density function, e.g., a multinomial distribution, and samples the values explicitly from that distribution. Implicit modelling is based on methods which rely on underlying model assumptions. In this case, we use explicit methods, following a multinomial distribution (Little et. al, 2002).

3 Application

The PLCO trial included 76,685 men who were randomised in screening and control group. The screening group received an annual PSA test for 6 years and an annual digital rectal examination for 4 years. The men were aged 55 to 74 years and followed up for a median of 13 years.

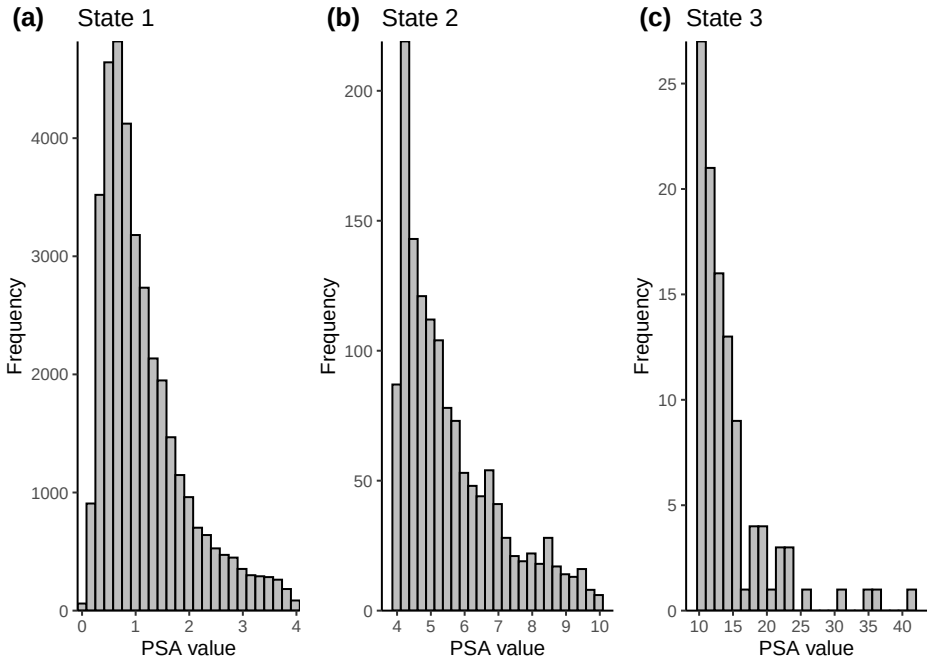


FIGURE 2. PSA distributions for the age interval $[50, 60]$ from the screened group from the PLCO data set.

		To state					
From state	Censored	1	2	3	4	5	6
1	50677	127057	3661	134	589	5487	11344
2	2874	2108	6107	363	765	656	712
3	244	106	197	380	120	72	79

TABLE 1. PLCO data state table.

As a first approach, we extend the healthy state (state 1) by discretising the PSA test value. The process is illustrated in Fig. 1. Given that the control group is not screened, there are no PSA test results for these individuals. However, we can impute these values using the screened group PSA distribution, by defining an age interval and assigning the first observed state to the control group. For instance, we can look at the PSA value distributions for the age interval $[50, 60]$. See, Fig. 2.
In Fig. 2 we can see that the three PSA distributions for the age interval $[50, 60]$ are positively skewed, and it seems to follow a bimodal distribution, which is concentrated between $[0, 2]$, and $[4, 6]$. Furthermore, we can see the variability within the PSA test results.

Parameter	Estimate	S.E.
$\beta_{12} = \beta_{23}$	-3.360	1.224e-04
$\beta_{21} = \beta_{32}$	-1.191	2.028e-04
$\beta_{14} = \beta_{24} = \beta_{34}$	-4.551	6.499e-05
β_{45}	-0.338	4.025e-07
$\beta_{16} = \beta_{26} = \beta_{36} = \beta_{46}$	-4.222	3.326e-05

TABLE 2. Estimated parameters for exponential model on prostate cancer screening data. S.E. stands for standard errors.

As a first approach, we imputed the first PSA value to the control group by sampling randomly from the PSA distributions from the screened group. Further in the project, we will explore different methodologies to deal with missing data that accounts for the variability of the estimates. The state table with the imputed PSA values for the control group is shown in Table 1.

Before modelling age-dependent hazards, an exponential model can be fitted; $h_{rs.0}(t) = \exp(\beta_{0.rs})$. Following the model assumptions for competing risks in survival defined by Bhatt et. al (2021), the parameters constraints for this model are such that $\beta_{16} = \beta_{26} = \beta_{36} = \beta_{46}$. Furthermore, the forward and backward hazards within the three healthy states are restricted such that $\beta_{12} = \beta_{23}$, and $\beta_{21} = \beta_{32}$. The results of the exponential model can be seen in Table 2. The model was fitted using the `msm` package in R with the Nelder-Mead optimisation method, allowing for $10,000$ iterations and using the crude estimates of the transition hazards as initial values.

Following an exponential model, the sojourn time in state 4; i.e. the expected time in the screen detectable state can be derived as $\frac{1}{\exp(\hat{\beta}_{45})+\exp(\hat{\beta}_{46})}$ which equals 1.373 years.

This model will be extended to consider age-varying hazards and misclassified and mismeasured data. Further methodologies to deal with the missing PSA values from the control group will be explored, accounting for the variability of the estimates given that imputed values are not actual observed values and the uncertainties associated with the imputations need to be addressed appropriately. These methods are particularly interesting given that the estimation of the transition to the clinical state will include both; screened and non-screened populations.

References

Bhatt R, van den Hout A, Pashayan N. (2021). A multistate survival model of the natural history of cancer using data from screened and unscreened population. *Statistics in Medicine*, **40**, 3791−−3807.

Jackson, C. (2011). Multi-state models for panel data: the msm package for R. *Journal of Statistical Software*, 38.

Kalbfleisch, J. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, **80**, 863−−871.

Little, R.J.A., Rubin, D.B. (2002). Statistical Analysis with Missing Data. Hoboken, NJ, USA: Wiley.

Bayesian smoothing for joint extremes

Miguel de Carvalho[1], Junho Lee[2]

[1] School of Mathematics, University of Edinburgh, UK
[2] Financial Supervisory Service, South Korea

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

Abstract: We develop a Bayesian smoothing model that learns about the dynamics governing joint extreme values over time. We resort to a suitable class of generalized linear models, conditioned on a large threshold to devise a class of dual measures of time-varying extremal dependence. An illustration of the proposed methods to some leading European stock markets reveals some intricate extremal dependence patterns over the past 30 years.

Keywords: Bayesian P-splines; Coefficient of tail dependence; Extreme value theory; Multivariate extremes; Time-varying extremal dependence.

1 Introduction

Statistics of extremes is tailored for extrapolating into the tail of a distribution beyond observed (Coles, 2001). In a multivariate framework, methods of statistics of extremes can be further used for assessing the dependence between the extreme values of a random vector.

This note adds to the current body of literature on multivariate nonstationary extremes, which can be applied to monitor the dynamics of extreme dependence over time and evaluate how covariates affect the structure of extremal dependence. Specifically, below we introduce a Bayesian time-varying model that infers about the dynamics that regulate the comovements of extreme values over time. We will utilize two time-varying versions of well-known measures of tail dependence as our starting point for modeling; see χ_t and $\bar{\chi}_t$ below. A suitable class of generalized linear models is then devised that can be used to infer about the dual measures of time-varying extremal dependence, conditional on a large threshold. To learn about the nonstationary patterns of the time-varying dual measures of extremal dependence from data, we then employ a Bayesian P-spline (Lang and Brezger, 2004) approach.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 A P-spline model for time-varying extremes

The parameters of interest and their specification: We start by laying the groundwork. Observations are assumed to be generated by a discrete-time bivariate stochastic process $\{(X_t, Y_t)\}_{t=1}^n$ with standard unit Fréchet marginal distributions, that is, $X_t \sim F_{X_t}$ and $Y_t \sim F_{Y_t}$ with $F_{X_t}(x) = F_{Y_t}(x) = \exp(-1/x)$, for $x > 0$ and $t \in \{1, \ldots, n\}$. To track the dynamics governing extremal dependence over time, define the time-varying coefficients of extremal dependence as

$$\chi_t = \lim_{u \to \infty} P(X_t > u \mid Y_t > u), \tag{1}$$

and

$$\bar{\chi}_t = \lim_{u \to \infty} \frac{2 \log P(X_t > u)}{\log P(X_t > u, Y_t > u)} - 1, \tag{2}$$

for $1 \leq t \leq n$. The goal below is to develop models for (1) and (2) and to learn about these quantities from data.

Our specification for χ_t is semiparametric and entails an inverse link function, $F : \mathbb{R} \to [0, 1]$ and a smooth function $g(t)$. Specifically, we set

$$\chi_t \approx P(X_t > u \mid Y_t > u) \equiv F\{g(t)\}, \tag{3}$$

as $u \to \infty$. The inverse link function F enforces the parametric constraint that the conditional survival probability is contained between 0 and 1, i.e. $0 \leq \chi_t \leq 1$, whereas the smooth function g reflects the effect of time on the tail dependence. In terms of $\bar{\chi}_t$, we follow a similar line of attack and specify

$$\bar{\chi}_t \approx 2H\{l(t)\} - 1, \tag{4}$$

as $u \to \infty$. Here, $H : \mathbb{R} \to [0, 1]$ is an inverse link function, and $l(t)$ is a smooth function. We complete the model specification by modelling the smooth functions using B-splines. Consider $m + 1$ equally-spaced knots, $t_0 < \cdots < t_m$. The smooth functions are then modelled as

$$g(t) = \sum_{k=1}^{K} \beta_k^{(g)} B_k^d(t), \qquad l(t) = \sum_{k=1}^{K} \beta_k^{(l)} B_k^d(t), \tag{5}$$

where $B_k^d(t)$ is a B-spline basis function of degree d evaluated at time t and $K = d + m$.

Learning from data: To learn about χ_t and $\bar{\chi}_t$ it is key to note that

$$I_t \dot{\sim} \text{Bern}(F\{g(t)\}), \qquad E_t \dot{\sim} \text{Exp}(H\{l(t)\}), \tag{6}$$

as $u \to \infty$, where $Z_t = \min(X_t, Y_t)$ and

$$\{I_t\} = \{1_{\{X_t > u\}} : Y_t > u\}, \qquad \{E_t\} = \{\log(Z_t/u) : Z_t > u\}.$$

The goal is hence to learn about χ_t and $\bar{\chi}_t$ from $k_I = |\{I_t\}|$ and $k_E = |\{E_t\}|$ pseudo-observations from $\{I_t\}$ and $\{E_t\}$, with $|\cdot|$ denoting cardinality. To learn about the coefficients of B-splines in (5) we use the Bayesian P-spline approach of Lang and Brezger (2004). To ease notation we focus on presenting the details for a single $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^{\mathrm{T}}$, rather than for both $\boldsymbol{\beta}^{(g)}$ and $\boldsymbol{\beta}^{(l)}$. We assign a first-order random walk prior to the coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^{\mathrm{T}}$ of each smooth function, which specifies a priori that the neighbouring components of $\boldsymbol{\beta}$ are related via an independent and identical Gaussian error ε_k with mean zero and variance τ^2; that is, we set

$$\beta_k = \beta_{k-1} + \varepsilon_k, \quad \varepsilon_k \sim N(0, \tau^2), \tag{7}$$

for $k = 2, \ldots, K$, and set a flat prior for the initial coefficient β_1. The first order random walk prior can be represented in a matrix form $\mathbf{D}\boldsymbol{\beta}$, where $\mathbf{D}$ is a difference matrix of dimension $(K-1) \times K$. The matrix $\mathbf{D}$ has 1's in diagonal elements $(i = j)$, -1's in the next elements from the diagonal $(i = j+1)$, and 0's otherwise for $i = 1, \ldots, K-1$ and $j = 1, \ldots, K$. The variance τ^2 controls the degree of smoothness of the smooth function (say, g or l); a small value of τ^2 results in a less wiggly curve, as each component of $\boldsymbol{\beta}$ tends to be close to the value of its neighbouring component. Accordingly, the conditional probability of the regression coefficients $\boldsymbol{\beta}$ given τ^2 is

$$\pi(\boldsymbol{\beta} \mid \tau^2) \propto \exp\left(-\frac{1}{2\tau^2} \boldsymbol{\beta}^{\mathrm{T}} \mathbf{K} \boldsymbol{\beta} \right), \tag{8}$$

where $\mathbf{K}$ is a penalty matrix, $\mathbf{K} = \mathbf{D}^{\mathrm{T}}\mathbf{D}$. In the full Bayesian setting, the smoothing parameter τ^2 is also estimated along with the regression coefficients by assigning an hyperprior distribution to it. Finally, we place a diffuse Inverse Gamma prior $\tau^2 \sim \mathrm{IG}(a_0, b_0)$ with $a_0 > 0$ and $b_0 > 0$.

3 A real data illustration

We now illustrate the proposed methodology on data from some leading European stock markets (CAC 40, France; DAX 30, Germany; FTSE 100, UK). The focus of the analysis will be on assessing the dynamics governing the dependence of extreme losses between FTSE 100 against CAC 40 and DAX 30. We obtained the closing daily stock index levels for the analyzed markets from Datastream, and the sample period ranges from March 5th, 1990 to May 4th, 2020. Since our focus is on (extreme) losses, we work with negative daily returns. Following standard practice in related literature (e.g., Castro *et al.*, 2018), returns are filtered using GARCH(1,1) model, and transformed into unit Fréchet margins. Finally, we threshold the data by using the 95% quantile to pairwise minima $Z_t = \min(X_t, Y_t)$, and obtain the pseudo-samples $\{I_t\}$ and $\{E_t\}$, per market. As can be seen from Fig. 1, for the first two decades extreme joint losses of UK–FRA and UK–GER

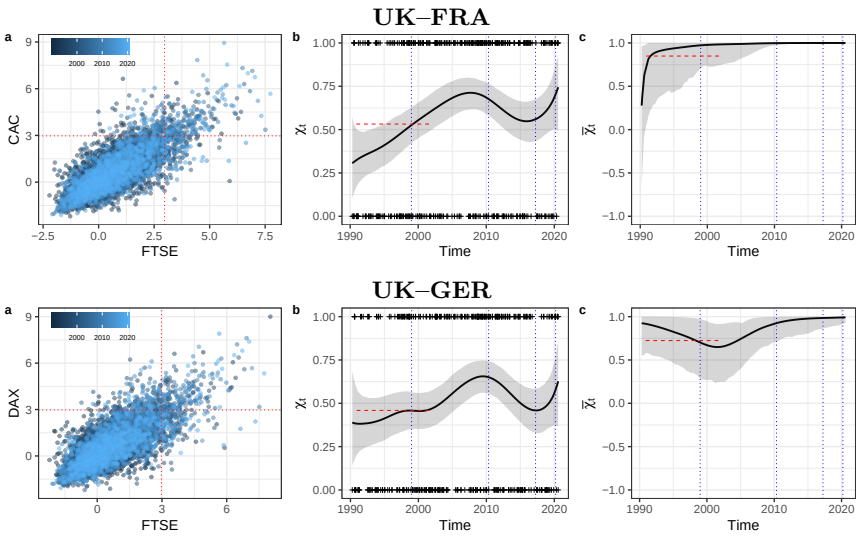


FIGURE 1. Left: Scatterplot of log transformed data. Middle and Right: Posterior mean time-varying χ_t and $\bar{\chi}_t$ (solid) along with credible bands; the rug in the middle panel corresponds to the points $\{(t, 1_{\{X_t > u\}}) : Y_t > u\}$ whereas the dashed red line corresponds to the available values from the subperiod analysis of Poon *et al.* (2003).

exhibit clear evidence for an increasing extremal dependence. Extremal dependence peaks right after the around the 2009 subprime crisis, it goes down, and it finally increases again around 2017, when the UK invocation of Article 50 of the Lisbon Treaty for Brexit took place.

Acknowledgments: Special thanks to CEAUL (*Centro de Estatística e Aplicações da Universidade de Lisboa*) for funding.

References

Castro, D., de Carvalho, M. and Wadsworth, J. L. (2018). Time-varying extreme value dependence with application to leading European stock markets. *The Annals of Applied Statistics*, **12**, 283–309.

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

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

Poon, S.-H., Rockinger, M. and Tawn, J. (2003). Modelling extreme-value dependence in international stock markets. *Statistica Sinica*, **13**, 929–953.

Semi-parametric estimation of growth curves

Chiara Di Maria[1], Vito M. R. Muggeo[1]

[1] Department of Economics, Business and Statistics, University of Palermo, Italy

E-mail for correspondence: chiara.dimaria@unipa.it

Abstract: Sigmoidal curves, very common in epidemiology and biology, have traditionally been fitted using parametric models or fully non-parametric approaches like splines. In this paper, we propose a semi-parametric approach which is flexible enough to capture several sigmoidal shapes. The estimation procedure is iterative and relies on a first-order Taylor expansion around the inflection point. The performance of our approach is compared to some parametric models through a simulation study and an application to data. Results of simulations show that our approach performs well in terms of mean integrated squared errors in a variety of scenarios.

Keywords: Growth curves; Inflection point; Semi-parametric estimation

1 Introduction

Growth curves are a very common tool to model both observational and experimental data in epidemiology and biology. Indeed, many phenomena, modelled as functions of some predictors, assume a sigmoidal shape, i.e. they show an increasing trend up to a certain point, from which the growth decelerates and stabilises to an asymptotic level. Examples include bacterial growth and neurological disorders like Huntington's disease. In the literature, many parametric models have been proposed to estimate this kind of curves, such as logistic, Gompertz and Weibull, most of which were proved to belong to a unified family (Tjørve and Tjørve, 2017). As an alternative to a fully parametric specification, other authors used non-parametric methods like splines (Aggrey, 1991), which can easily adapt to observed data, guaranteeing good fits, but have no interpretable parameters.
In this article, we propose a compromise between the two above: a semi-parametric approach. The model is based on two pieces of curves joining at the inflection point of the sigmoidal curve, and it is flexible enough to

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

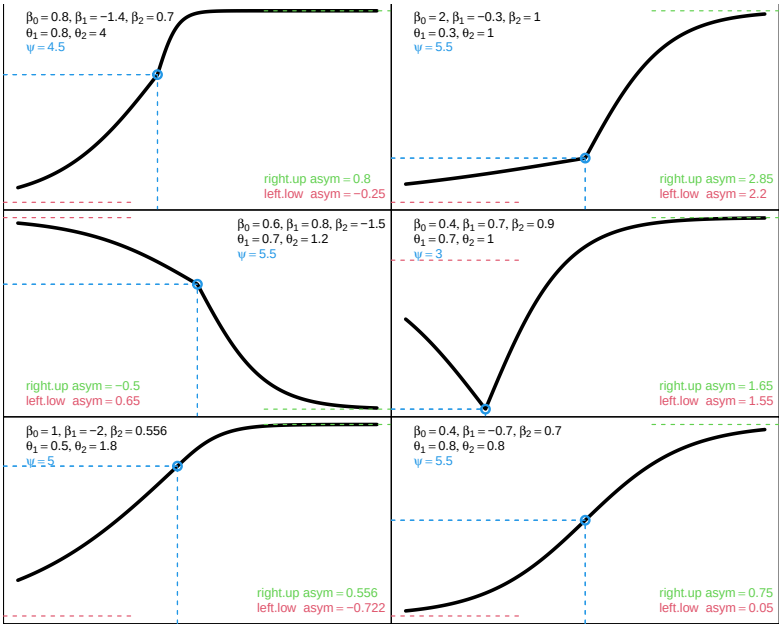


FIGURE 1. Some examples of growth curves according to equation (1). Dashed red and green lines are the left and right asymptotes, respectively, and the blue lines indicate the coordinates of the change point (not necessarily an inflection point).

fit non-standard curves with very different growth rates before and after the inflection point. Simulation results in Section 3 bear this out.

2 Modelling

The proposal can be considered as an extension of the segmented regression presented in Muggeo (2003). Let Y be the response variable of interest and X a continuous explanatory variable in $[m, M]$ where $m = min(X)$ and $M = max(X)$. A growth curve with an inflection point ψ can be represented by two branches, defined in $[m, \psi)$ and $[\psi, M]$, respectively. Namely

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 \frac{\exp[\theta_1(\psi - x)_+]}{1 + \exp[\theta_1(\psi - x)_+]} + \beta_2 \frac{\exp[\theta_2(x - \psi)_+]}{1 + \exp[\theta_2(x - \psi)_+]} \qquad (1)$$

where $(\psi - x)_+ = (\psi - x)I(\psi > x)$ and $(x - \psi)_+ = (x - \psi)I(x > \psi)$, with $I(\cdot)$ being the indicator function. θ_1 and θ_2 are scale parameters regulating the concavity or steepness of the two branches, while the β's characterise the asymptotes: the left one is given by $\beta_0 + \beta_1 + \beta_2/2$, the right one by $\beta_0 + \beta_1/2 + \beta_2$. Some examples are portrayed in Figure 1.

Model (1) has 6 parameters and the curve is continuous but not differentiable at $x = \psi$. It is differentiable at the inflection point if the derivatives of the two branches evaluated at ψ coincide, i.e. $\beta_1\theta_1 + \beta_2\theta_2 = 0$, which reduces the number of parameters to 5. The curve is asymmetric around ψ, but symmetry is ensured if $\theta_1 = \theta_2$ and $\beta_1 = -\beta_2$, making the number of parameters equal to 4.
Parameter estimation can be based on the first-order Taylor expansion of each branch around some specified starting values for $(\psi, \theta_1, \theta_2)$, which results in a quite efficient algorithm consisting of fitting simple linear models iteratively.

3 Simulation study

To evaluate the performance of our proposal, we contrast it with some fits obtained via the `drc` R package (Ritz et al., 2015), which includes a large number of parametric models traditionally used in the context of growth curve analyses. We simulate data from four different models, with 3 sample sizes ($n = 25, 50, 100$) and three levels of noise (low, medium and high). Formally, for $i = 1, \ldots, n$: $\mu_i = f(x_i; \boldsymbol{\xi})$, $y_i = \mu_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, $\sigma \in \{0.05, 0.2, 0.5\}$, where x_i is an equispaced sequence from 1 to 10 of length n. The true $f(\cdot; \boldsymbol{\xi})$ functions are 1) four-parameter logistic model, 2) Gompertz model, 3) arctan function and 4) Weibull model, and $\boldsymbol{\xi}$ is the vector of parameters characterising each model; in all models, the inflection point is fixed at $\psi = 6$. For each generating model, we simulate 500 replicates and we fit six models: i) five-parameter logistic (L5); ii) four-parameter logistic (L4); iii) Gompertz (G, 4 parameters); iv) Weibull (W, 4 parameters); v) model (1) (S, 6 parameters); vi) model (1) with symmetry constraints (Ss, 4 parameters). We compare the performance of each model estimated on the simulated data sets via the MISE (Mean Integrated Squared Error). Figure 2 reports the ratios (on the log scale) between the MISE of each adapted model and the MISE corresponding to the true model. When the true generating model is arctan, since none of the estimated models is the correct one, we use the median MISE as a baseline.
We note that the proposed approach exhibits values of MISE equal or, surprisingly, even lower than those coming from the true fitted models in 33 cases out of 36. Only in three cases, for Gompertz, Arctan and Weibull curves for $n = 100$ and σ low, our semi-parametric model returns MISEs which are higher than those corresponding to the true fitted model.

4 Data analysis

We analyse the Heartrate data set in the `drc` package, which provides measurements of mean arterial pressure and heart rate collected on $n = 18$

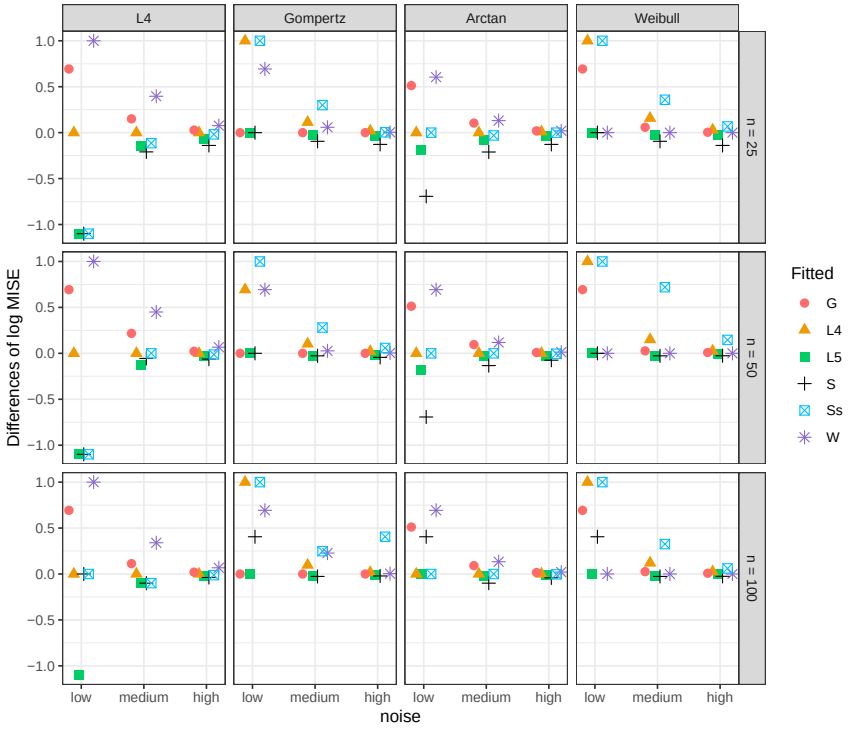


FIGURE 2. Simulation results: Ratios of MISEs (on the log scale), comparing the MISE of each fitted model to that corresponding to the true generating model.

subjects. Table 1 reports the estimated inflection points and the AIC values of our models and some competitors, including the *Baroreflex* model (Ricketts and Head, 1999), commonly used in studies about arterial pressure and recommended by the authors of `drc` for these data. It can be noticed that our approach, in the six-parameter version, performs the best. The symmetric curve is clearly not appropriate, as the data show an asymmetric pattern. Moreover, it is worth noticing that the 5-parameter logistic model, which was a good competitor of the segmented model in the simulation study, performs poorly on the analysed data. The estimated value of the asymmetry parameter is 16.72, reflecting the clear-cut asymmetric pattern of the data and, as a consequence, the resulting estimated $\hat{\psi} = 89.47$ cannot be interpreted as an inflection point of the fitted curve. Figure 3 shows a graphical representation of the fitted models, where we draw in colour our segmented models and Baroreflex, in black the other functions considered, which behave quite similarly to each other.

TABLE 1. Estimated ψ and AIC of different models fitted on Heartrate data.

Model	$\hat{\psi}$	AIC	number of parameters
Baroreflex	75.59	128.14	5
L5	-	142.00	5
L4	75.42	144.40	4
Gompertz	76.80	138.19	4
Weibull	76.82	139.64	4
S	77.67	115.54	6
Ss	75.37	143.22	4

Notes: In the logistic model with five parameters (L5), when the symmetry parameter is different from 1, i.e. the curve is asymmetric as in our case, the estimated value of ψ cannot be interpreted as the inflection point.

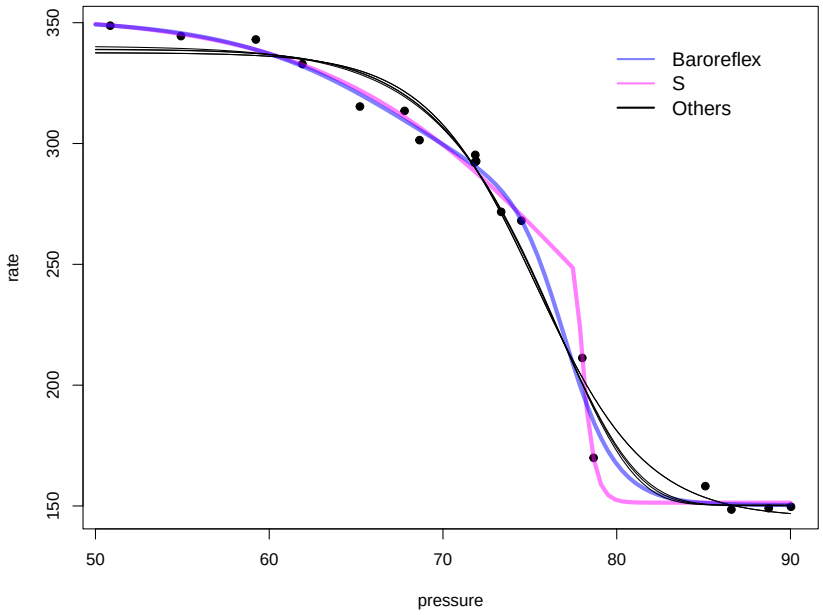


FIGURE 3. Fitted models for Heartrate data. The models with lowest AIC values, Baroreflex and S, are represented in blue and magenta, respectively. The thinner lines in black refer to the other fitted models, which have a very similar pattern.

5 Conclusion

We have proposed a novel semi-parametric approach to fit sigmoidal curves, which is very flexible in catching different growth shapes and it appears to perform quite well in several scenarios. Several extensions can be considered and will be discussed in the next version of the abstract. The methods discussed in this paper will be implemented in the R package `segmented`.

References

Aggrey, S. E. (1991). Comparison of Three Nonlinear and Spline Regression Models for Describing Chicken Growth Curves. *Poultry Science*, **81**, 1782 – 1788.

Muggeo, V. M. R. (2003). Estimating regression models with unknown breakpoints. *Statistics in Medicine*, **22**, 3055 – 3071.

Ricketts, J. H. and Head, G. A. (1999). A five-parameter logistic equation for investigating asymmetry of curvature in baroreflex studies. *The American Journal of Physiology*, **277**(2), 441 – 454.

Ritz, C., Baty, F., Streibig, J. C. and Gerhard, D. (2015). Dose-response analysis using R. *PLoS One*, **10**(12), e0146021

Tjørve, K. M., and Tjørve, E. (2017). A proposed family of Unified models for sigmoidal growth. *Ecological Modelling*, **359**, 117 – 127.

Modelling time-of-day variation in hidden Markov models using cyclic P-splines

Carlina C. Feldmann[1], Sina Mews[1], Roland Langrock[1]

[1] Bielefeld University, Germany

E-mail for correspondence: carlina.feldmann@uni-bielefeld.de

Abstract: Within the class of hidden Markov models (HMMs), which is a popular tool for modelling time series driven by underlying states, periodic variation in the state-switching dynamics is routinely modelled using trigonometric functions. This parametric modelling can be too inflexible to capture complex periodic patterns, e.g. featuring multiple activity peaks per day. We explore an alternative approach using penalised splines to model periodic variation within HMMs. The practicality and potential usefulness of our approach is demonstrated in a real-data application modelling the movements of an African elephant.

Keywords: EM algorithm; Nonparametric modelling; Periodicity; Time series

1 Introduction

Ecological time series data is often characterised by periodicities such as diel variation, i.e. recurrent patterns over a 24-hour period. Adequately modelling such periodic variation is crucial to comprehensively understand behavioural dynamics. One popular tool for modelling ecological time series and the periodicities therein is given by the class of hidden Markov models (HMMs), which links the observed ecological data (e.g. step lengths and turning angles in animal movement) to underlying non-observable states (e.g. resting, foraging, travelling; cf. McClintock et al., 2020).
In principle, relatively basic HMMs can be used to infer an animal's behavioural sequence (state decoding), based on which diel variation can be investigated using simple visualisations (see e.g. Schwarz et al., 2021). From the statistical perspective however, such a two-stage approach will often not be ideal: the uncertainty in the state allocation is not propagated, statistical inference on the periodic effects is not straightforward, and the dimension of the state space may be overestimated. Alternatively, periodic variation

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

can be directly incorporated in HMMs using trigonometric modelling, for instance relating the state transition probabilities to say the hour of the day using sine and cosine basis functions (see e.g. Leos-Barajas 2017). While this will often be sufficient, such a parametric approach may lack flexibility to capture complex periodic variation.
Here we explore a more flexible, nonparametric estimation of periodicities in the state-switching dynamics of an HMM using cyclic P-splines. For inference, we devise an expectation-maximisation (EM) algorithm, thereby isolating the estimation of the nonparametric periodic effect. This allows us to exploit the powerful machinery available for nonparametric (regression) modelling, specifically P-splines or other smoothing methods implemented in existing software packages such as `mgcv` (Wood, 2017).

2 Methods

HMMs are used to model time series data $x_1, \ldots, x_T$ (e.g. step lengths of an animal) driven by underlying states $s_1, \ldots, s_T$ (e.g. the behavioural modes). In a basic HMM, the state process is assumed to be a Markov chain with N states, characterised by the initial state distribution and the state transition probabilities $\gamma_{ij}^{(t)} = \Pr(S_{t+1} = j \mid S_t = i)$. The state active at time t selects which of N possible state-dependent distributions $f_1, \ldots, f_N$ generates the observation x_t. Interest often lies in the drivers of the state process, in which case the transition probabilities can be modelled as a function — the inverse multinomial logit link — of a linear predictor:

$$\gamma_{ij}^{(t)} = \frac{e^{\tau_{ij}^{(t)}}}{\sum_{k=1}^{N} e^{\tau_{ik}^{(t)}}},$$

with $\tau_{ii}^{(t)} = 0$ (reference category). When the aim is to model periodic patterns in the state-switching dynamics, the linear predictor $\tau_{ij}^{(t)}$ is typically constructed using trigonometric basis functions with the desired periodicity. For example, for modelling diel variation in a time series with hourly data, a possible general form of the linear predictor is

$$\tau_{ij}^{(t)} = \mathbf{z}_t' \boldsymbol{\beta}^{(ij)} + \sum_{k=1}^{K} \omega_k^{(ij)} \sin\left(\frac{2\pi k t}{24}\right) + \sum_{k=1}^{K} \psi_k^{(ij)} \cos\left(\frac{2\pi k t}{24}\right). \qquad (1)$$

By increasing K, arbitrary (smooth) modelling of the periodic effect can be achieved. However, when complex periodic patterns are to be modelled, it can be more straightforward to avoid making any assumptions on the functional shape of the periodic effect. To this end, we suggest replacing the sum of trigonometric basis functions in (1) by a linear combination of Q basis functions,

$$\tau_{ij}^{(t)} = \mathbf{z}_t' \boldsymbol{\beta}^{(ij)} + \sum_{q=1}^{Q} a_q^{(ij)} B_q(t \bmod 24), \qquad (2)$$

with the scaling coefficients $a_1^{(ij)}, \ldots, a_Q^{(ij)}$ to be estimated. We use cubic B-spline basis functions $B_1, ..., B_Q$, which are easy to compute and yield visually smooth functions. To enforce the desired periodicity, these are wrapped at the boundaries of the support. In practice, a large Q (e.g. 20) is typically used to guarantee sufficient flexibility. Overfitting is avoided by including a penalty on the sums of squared differences between the coefficients $a_q^{(ij)}$ associated with adjacent B-splines (an approach commonly referred to as P-spline modelling, cf. Eilers and Marx, 1996).

This model formulation effectively corresponds to a nonparametric regression within HMMs. For example, in case of $N = 2$ states, the model features one nonparametric logistic regression for each of the state-switching probabilities $\gamma_{12}^{(t)}$ and $\gamma_{21}^{(t)}$. For such nonparametric regression modelling, the inferential machinery is well-established. Therefore, we apply the expectation-maximisation algorithm (EM) to isolate the estimation of the logistic regression component from the estimation of the other parameters of the HMM, in particular those associated with the state-dependent process.

In the E-step of the EM algorithm, we replace all functions of the unobserved states in the complete-data log-likelihood (CDLL) by their conditional expectations, given the data and the current guess of the parameter values. In the M-step, we optimise the resulting CDLL with respect to the model parameters. The updated estimates of the initial state distribution as well as the state-dependent distributions are routinely obtained. For updating the parameters affecting the state transition probabilities, we exploit that each row of the t.p.m. implies a categorical regression for the transition to the next state, such that the associated parameters of these regressions can conveniently be estimated separately, for example using `mgcv`.

3 Case study

We consider hourly GPS data collected for an African elephant from October 2008 to August 2010. From the positional data, we calculate the Euclidean step lengths as well as the turning angles between consecutive compass directions, based on which we aim to investigate diel patterns in the elephant's behaviour. We model the data using 2-state HMMs with gamma and von Mises distributions for the step lengths and turning angles, respectively. For modelling diel variation in the state-switching dynamics, we consider the cyclic P-spline approach and, as a benchmark, the trigonometric approach (1) with $K = 1, 2$ and 3. All fitted models feature an "encamped" state with relatively short step lengths and frequent reversals in direction (state 1) and an "exploratory" state with longer steps and higher persistence in direction (state 2).

Figure 1 displays the state-switching probabilities estimated under the nonparametric as well as the parametric approach. All models detect a reduction in exploratory activity during the night. However, the flexible P-spline

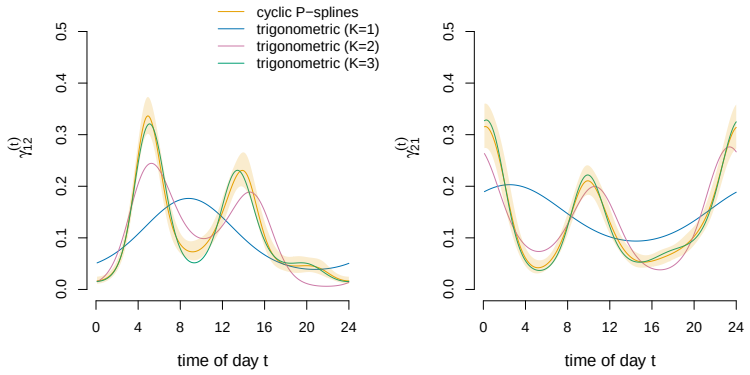


FIGURE 1. Estimated transition probabilities as a function of time of day, for the different HMMs considered. For the P-spline model, the pointwise 95% confidence intervals as provided by `mgcv` are shown.

approach additionally captures a bimodal diel variation, with more frequent switching to the exploratory mode in the early morning and in the early afternoon. In contrast, the commonly used trigonometric effect modelling with $K = 1$ (i.e. one sine and one cosine basis function) is not sufficiently flexible to identify this bimodality. When increasing the order to $K = 2$, the bimodality can be identified, but only with $K = 3$ the parametric approach produces results similar to those obtained using splines. Modelling periodic variation nonparametrically thus allows us to investigate temporal patterns without making any (a priori) restrictive assumptions and can uncover relevant patterns that may otherwise go unnoticed.

References

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

Leos-Barajas, V. et al. (2017). Analysis of animal accelerometer data using hidden Markov models. *Methods in Ecology and Evolution*, **8**, 161 – 173.

McClintock, B.T. et al. (2020). Uncovering ecological state dynamics with hidden Markov models. *Ecology Letters*, **23**, 1878 – 1903.

Schwarz, J. et al. (2021). Individuality counts: A new comprehensive approach to foraging strategies of a tropical marine predator. *Oecologia*, **195**, 313 – 325.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

Bayesian inference of dynamic models emulated with a time series Gaussian process

Yuzhang Ge[1], Arash Rabbani[2], Hao Gao[1], Dirk Husmeier[1]

[1] University of Glasgow, UK
[2] University of Leeds, UK

E-mail for correspondence: 2670035g@student.gla.ac.uk; Dirk.Husmeier@glasgow.ac.uk; Hao.Gao@glasgow.ac.uk;

Abstract: This work is motivated by cardiophysiological disease diagnosis, where the objective is to infer certain biophysical parameters of a cardiac mechanics model from a time series of quantities extracted from magnetic resonance images (the volume of the left ventricle of the heart during different time points). Due to the computational complexity of the cardiac model, we use a Gaussian process as a statistical surrogate model for emulation. This requires us to build the emulator in the Cartesian product space of time and biophysical parameters. In this paper, we explore an approach based on a decomposition of the full covariance matrix as a Kronecker product of two separate covariance matrices in biophysical parameter space and time. We first evaluate the accuracy and computational efficiency of this approach on a simple toy problem before applying it to a cardiac mechanics model of the passive filling process during diastole.

Keywords: Gaussian process; Time series kernel; Inverse estimate

1 Introduction

There have recently been impressive advancements in the mathematical modelling of cardiophysiological processes (Mangion et al., 2017). However, getting these models into the clinic for improved decision support is still challenging. The main difficulty is related to the fact that these models depend on various biophysical parameters, which differ from patient to patient. These parameters cannot be measured directly and need to be inferred based on a comparison between model predictions and data; the latter is typically related to quantities of interest (QoIs) extracted from magnetic resonance image (MRI) scans. A commonly used approach for inferring these parameters involves iterative optimization. This entails defining a loss function that compares the model predictions and measured

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

quantities of interest, which are then optimized to obtain the set of parameter estimates. Since the mathematical models do not have closed-form solutions, and repeated numerical simulations are computationally expensive, new methods have to be developed to achieve patient-specific model calibration in real-time. A promising approach is emulation (Davies, V. et al., 2019), where we build a statistical surrogate model of the computationally expensive original mathematical model. Previous work has explored the option of emulating the QoIs extracted from an MRI scan at a fixed time point (Lazarus et al., 2022). The motivation of the current study is to extend this work by emulating a time series of QoIs. To this end, we will explore the application of a Gaussian process (GP) (Roberts et al., 2013) in the cartesian product space of time and biophysical parameters. We will first evaluate the performance of the method on a simple toy problem before applying it to the cardiac passive filling process during diastole.

2 Time-series Gaussian process

Consider $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2 .... \mathbf{x}_n]$, where $\mathbf{x}_i$ is the input variable for the ith case, and the output space

$$\mathbf{Y} = [\mathbf{f}(\mathbf{x}_1) .... \mathbf{f}(\mathbf{x}_n)] \tag{1}$$

with

$$\mathbf{f}(\mathbf{x}_i) = [f(\mathbf{x}_i, t_1), \ldots, f(\mathbf{x}_i, t_e)] \tag{2}$$

a time series of data as our QoI, where $\mathbf{f}$ has a joint Gaussian distribution (Roberts et al. (2013)): $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$, where $\boldsymbol{\mu}$ is the mean of $\mathbf{Y}$, and $\mathbf{K}$ is the covariance matrix with $i, j \in [1, ..., n]$. For m unobserved cases $\mathbf{X}^* = [\mathbf{x}_1^*, ..., \mathbf{x}_m^*]$, we also assume $\mathbf{f}(\mathbf{X}^*) \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\sigma^2}^*)$, then we have

$$\boldsymbol{\mu}^* = \mathbf{K}(\mathbf{X}, \mathbf{X}^*)^T \mathbf{K}(\mathbf{X}, \mathbf{X})^{-1} \mathbf{Y} \tag{3}$$

$$\boldsymbol{\sigma}^{*2} = \mathbf{K}(\mathbf{X}^*, \mathbf{X}^*) - \mathbf{K}(\mathbf{X}, \mathbf{X}^*)^T \, \mathbf{K}(\mathbf{X}, \mathbf{X})^{-1} \, \mathbf{K}(\mathbf{X}, \mathbf{X}^*), \tag{4}$$

in which, $\boldsymbol{\sigma}^{*2}$ is the covariance of $\mathbf{f}(\mathbf{X}^*)$. Now we consider the time-series GP, each case $\mathbf{x}_i$ has k different parameters. Furthermore, we consider the temporal space $\mathbf{T}$ is divided into e divisions, such that $\mathbf{T} = [t_1, .., t_e]$. Accordingly, for each time point t_i, there are corresponding outputs Y_{t_i}. To describe the time-series outputs of $\mathbf{Y}$, Roberts et al. (2013) introduced a covariance matrix using the Kronecker product as $cov(\mathbf{Y}, \mathbf{Y}) = \mathbf{K}_1(\mathbf{X}, \mathbf{X}) \otimes \mathbf{K}_2(\mathbf{T}, \mathbf{T})$ where $\mathbf{K}_1(\mathbf{X}, \mathbf{X})$ and $\mathbf{K}_2(\mathbf{T}, \mathbf{T})$ can be calculated from different kernel functions. The advantage of using the Kronecker product is to separate the parameter axis and the time axis, allowing us to use alternative kernels to compute the mean function and split the large-size covariance matrix into the combination of two smaller matrices, which will substantially reduce computational costs.

3 Results

3.1 The toy model

We first introduce the toy model without added noise, that is

$$Y = A\sin(2\pi B \cdot t + C) + D, \tag{5}$$

in which the input space is $\mathbf{X} = [A, B, C, D]$ with $A \in [1.0, 5.0]$, $B \in [0.5, 2.0]$, $C \in [0.0, \pi]$, $D \in [1.0, 5.0]$ and time $t \in [0.0, 1.0]$. For a space-filling design, we use the Sobol sequence to generate 200 cases with 180 cases for training and 20 for testing. The time-series GP for the toy model is

$$\mathbf{Y} \sim \mathcal{GP}(0, \mathbf{K}(\mathbf{X}, \mathbf{X})) \quad \text{with} \quad \mathbf{K} = \mathbf{K}_1(\mathbf{X}, \mathbf{X}) \otimes \mathbf{K}_2(\mathbf{T}, \mathbf{T}),$$

where the kernel for parameters is the ARD Gaussian kernel, and the kernel for time is from Matern family because of their flexibility in controlling the smoothness to find the best description of the time axis. We further compare different Matern kernels (5/2, 3/2) and the ARD Gaussian kernel for the time axis using the 20 test cases. We found that the Matern 3/2 kernel gave the highest adjusted R^2 score and the least mean squared error (MSE) on the test set from table 1 below. Therefore the Matern 3/2 kernel is used in the following analysis.

TABLE 1. The choice of k_2 the kernel function when applied to the 20 test cases.

Kernel function	R^2	**MSE**
Matern 5/2	0.993	0.04
Matern 3/2	0.996	0.03
the ARD Gaussian kernel	0.962	1.10

After training the time-series GP, we now use it to infer unknown parameters of the toy model with noisy observed data $\tilde{Y}$, which is synthetically generated with $\sigma_m^2 = 0.2$ as

$$\tilde{Y} = A\sin(2\pi B \cdot t + C) + D + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma_m^2). \tag{6}$$

For each test case, we can predict the mean function $\mu(\cdot)$ of Y with unknown parameters $\boldsymbol{\theta} = [\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}]$. The prior of $\boldsymbol{\theta}$ is the uniform distribution ($\tilde{A} \sim U(1.0, 5.0)$, $\tilde{B} \sim U(0.5, 2.0)$, $\tilde{C} \sim U(0.0, \pi)$ and $\tilde{D} \sim U(1.0, 5.0)$). The likelihood function can be defined as $\tilde{Y} \sim \mathcal{N}(\mu(\boldsymbol{\theta}), \sigma_m^2)$

$$\log(p(\tilde{\mathbf{y}}|\boldsymbol{\mu}(\boldsymbol{\theta}))) = -\frac{1}{2}\log(2\pi\sigma_m^2) - \frac{1}{2\sigma_m^2}\sum_{i=1}^{e}(\tilde{y}_i - \mu(\boldsymbol{\theta}, t_i))^2. \tag{7}$$

By using Bayesian inference, we can sample $\boldsymbol{\theta}$ from the posterior $p(\boldsymbol{\theta}|\mu, \tilde{Y})$ by applying the Hamiltonian Monte Carlo Method (Casella and Robert, 2008).

One test case is selected to evaluate the effects of time points on parameter inference, we then quantify how much more peaked the posterior distribution becomes as we increase the sample size from 10 to 50, shown in

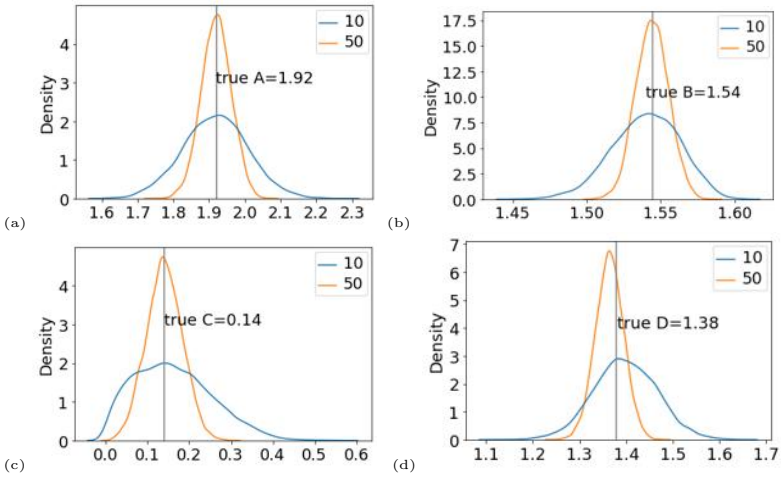


FIGURE 1. Posterior distributions for A, B, C and D, respectively.

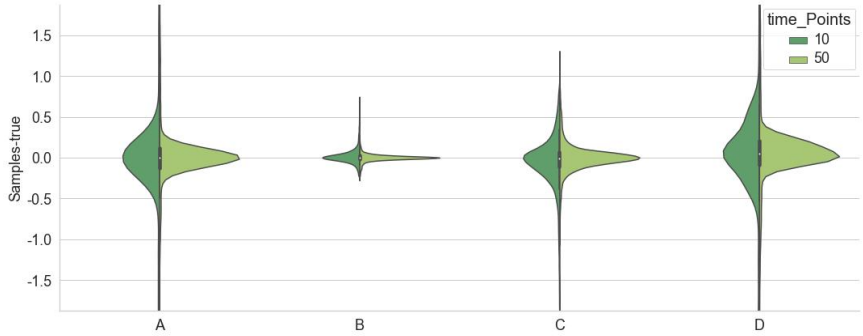


FIGURE 2. Residual of 20 test cases for $e = 10$ or $e = 50$ time points.

figure 1, where samples of $[\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}]$ from the posterior with a kernel density estimator are drawn, respectively.

For each test of parameter inference, we would get 10000 samples as results and calculate residuals defined as the difference between samples and true values. All residuals are then combined for the 20 test cases to quantify the reduction in the posterior uncertainty as a consequence of increasing the length of the time series by a factor of 5, shown in figure 2.

3.2 The cardiac model

We now consider a nonlinear biomechanical cardiac model in diastole, as shown in figure 3(a). The myocardial material property is described by a strain energy function, the so-called H-O model (Holzapfel, G. and Ogden, R., 2009), which has 8 parameters $(a, b, a_{\mathrm{f}}, b_{\mathrm{f}}, a_{\mathrm{s}}, b_{\mathrm{s}}, a_{\mathrm{fs}}, b_{\mathrm{fs}})$. In this study, we only vary a and b while fixing $(a_{\mathrm{f}}, b_{\mathrm{f}}, a_{\mathrm{s}}, b_{\mathrm{s}}, a_{\mathrm{fs}}, b_{\mathrm{fs}})$. The input space consists of $[a, b]$ and 10-time points to describe LV cavity volumes in

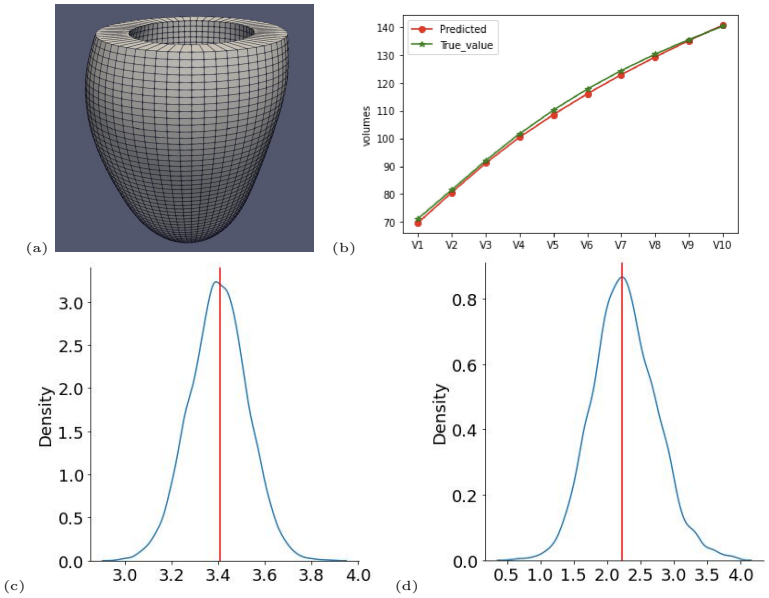


FIGURE 3. The LV geometry (a); the predicted LV volume time series (b); Posterior distributions of a (c) and b (d).

diastole for each case. To optimise kernel parameters of the time-series GP, 300 samples were generated using the Sobol sequence with 20 reserved for testing. We finally inversely estimated a and b for one test case using the trained time-series GP. Our preliminary results show that the time-series GP can emulate the LV dynamics in diastole well (figures 3(b)), and can be used for parameter inference without running computationally expensive mathematical models, see the posterior distributions of a and b in figures 3(c) and (d).

4 Conclusions

We have constructed an emulator to accurately emulate time series data generated from a toy model and a cardiac model. The toy model has four parameters (frequency, amplitude, phase and offset) to describe a sine-like curve that makes it challenging to emulate. Our experiments demonstrate the time-series GP can well emulate the toy model. We quantify the reduction in posterior uncertainty as the time series length increases. Finally, we demonstrate that the time-series GP can also accurately emulate the LV cavity volume time series in diastole, and unknown material parameters can be inferred using this emulator. Our next step is to infer material parameters of the human heart using this time-series GP with automatically measured LV cavity volume data from in vivo MRI, which is based on the work from papers (Lazarus, A. and et al.,2022) and (Arash, R. and et al., 2023).

Acknowledgements

This work has been supported by EPSRC (grant reference numbers EP/T017899/1, EP/S020950/1, EP/R511705/1) and the British Heart Foundation (PG/22/10930). We would like to thank Dr Alan Lazarus for help with code.

References

Roberts, S. and Osborne, M. and Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. **371**, 20110550

Lazarus, Al. and Dalton, D. and et al. (2022). Sensitivity analysis and inverse uncertainty quantification for the left ventricular passive mechanics. *Biomechanics and Modeling in Mechanobiology*. **21**, 953 – 982.

Casella, G. and Robert, C. (2008). Monte Carlo statistical methods. University of Florida.

Mangion, K. and Gao, H. and et al. (2017). Advances in computational modelling for personalised medicine after myocardial infarction. *Heart*. **104**, 550 – 557.

Davies, V. and Noè, U.and Lazarus, A. and et al. (2017). Fast parameter inference in a biomechanical model of the left ventricle by using statistical emulation. *Journal of the Royal Statistical Society. Series C, Applied Statistics*. **68**, 1555.

Lazarus, A. and Gao, H. and Luo, X. and Husmeier, D. (2022). Improving cardio-mechanic inference by combining in vivo strain data with ex vivo volume-pressure data. *Journal of the Royal Statistical Society*. **74**, 906 – 931.

Arash, R. and Hao, G. and Lazarus, A. and et al. (2023). Image-Based Estimation of the Left Ventricular Cavity Volume Using Deep Learning and Gaussian Process with Cardio-Mechanical Applications. *Computerized Medical Imaging and Graphics*. **106**, 102203.

Holzapfel, G.and Ogden, R. (2009). Constitutive modelling of passive myocardium: a structurally based framework for material characterization. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. **367**, 3445 – 3475.

Gradient boosting for parsimonious additive covariance matrix modelling

Vincenzo Gioia[1], Matteo Fasiolo[2], Ruggero Bellio[3]

[1] Department of Economics, Business, Mathematics, and Statistics, University of Trieste, Italy
[2] School of Mathematics, University of Bristol, UK
[3] Department of Economics and Statistics, University of Udine, Italy

E-mail for correspondence: vincenzo.gioia@units.it

Abstract: Gradient boosting algorithms are attractive for effect selection in multi-parameter generalized additive models. Due to the high-dimensionality of the problem, a parsimonious covariance matrix model is required for modelling multivariate Gaussian data. Here, we address covariance matrix model specification using gradient boosting. In particular, the aim is ranking the effects used to model the elements of the modified Cholesky decomposition of the precision matrix. The performance of the proposal is illustrated on electricity demand data.

Keywords: Covariance matrix modelling; Generalized additive models; Model selection; Modified Cholesky decomposition; Multivariate electricity load modelling.

1 Introduction

Additive covariance matrix models for multivariate Gaussian data account for the potentially varying nature of the covariance matrix, $\mathbf{\Sigma}$. However, the number of distributional parameters increases quadratically with the dimension of the outcome vector, which affects the scalability of model fitting and poses a barrier to manual model selection. For multi-parameter generalized additive models (GAMs), non-cyclical component-wise gradient boosting is an effective tool for automatic effect selection (Thomas et al., 2018). In this work, we select a covariance matrix model, based on the modified Cholesky decomposition (MCD) parametrisation (Pourahmadi, 1999), by leveraging gradient boosting. We evaluate model performance on multivariate electricity demand data.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Parsimonious covariance matrix models

Let $\mathbf{y}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, \ldots, n$, be independent d-dimensional response vectors. Denote with $\boldsymbol{\eta}$ the $n \times q$ matrix of the linear predictors, where $q = d + d(d+1)/2$ is the number of distributional parameters involved in both mean and covariance matrix modelling. The covariates $\mathbf{x}_i$ enter the model through the linear predictor vector $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iq})$, which specifies the mean vector ($\mu_{ik} = \eta_{ik}$ for $k = 1, \ldots, d$) and the unconstrained elements of the MCD parametrisation, $\boldsymbol{\Sigma}_i^{-1} = \mathbf{T}_i^{\mathrm{T}} \mathbf{D}_i^{-2} \mathbf{T}_i$. That is, η_{ik}, $k = d+1, \ldots, q$, control the non-trivial entries of $\log \mathbf{D}_i^2$ and $\mathbf{T}_i$, where $\mathbf{T}_i = \mathbf{D}_i \mathbf{C}_i^{-1}$, with $\mathbf{D}_i$ the diagonal matrix containing the diagonal elements of $\mathbf{C}_i$, the lower-triangular Cholesky factor of $\boldsymbol{\Sigma}_i$. The elements of $\boldsymbol{\eta}_i$ are

$$\eta_{ik} = \sum_h f_{hk}(x_i^h)\,, \tag{1}$$

where the $f_{hk}(\cdot)$'s can be linear or smooth effects of covariate x^j. Linear and the smooth effects, the latter being built using spline bases, are parametrised by the regression coefficients vector $\boldsymbol{\beta}$. Their complexity is controlled via quadratic penalties scaled by smoothing parameter vector $\boldsymbol{\lambda}$. Vector $\boldsymbol{\lambda}$ is selected via an outer generalised Fellner-Schall iteration (Wood and Fasiolo, 2017) while, for fixed $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$ is estimated by maximising the log-posterior density via Newton's algorithm. See also Wood (2017).
We use non-cyclical component-wise gradient boosting to select the effects $f_{hk}(\cdot)$ to include in the covariance matrix model, that is for $k = d+1, \ldots, q$. Briefly, having initialised the linear predictors, the gradient boosting algorithm fits by least squares (but note that penalised least squares could be used instead) the gradient of the log-likelihood with respect to the linear predictors involved in covariance matrix modelling, spanning a list of candidate effects. The effect-linear predictor pair that leads to the largest log-likelihood increase is used to update the model, where the step length (learning rate) of the update is usually fixed to a sufficiently small value. The algorithm is run for a certain number of iterations and the output of the procedure is a list of effects, which are ordered by decreasing cumulative log-likelihood gains. Then, the number of effects included in the final model is selected by maximising the out-of-sample predictive performance. Having selected the mean model manually and the covariance matrix model as just described, we fit the corresponding multivariate Gaussian additive model, using the Fellner-Schall iteration. See Gioia et al. (2022) for details and Strömer et al. (2022) for related work.

3 Illustration

We consider data from the electricity load forecasting track of the GEFCom2014 challenge (Hong et al., 2016). The outcome vector elements are

the hourly loads (load_j), from 12 p.m. ($j = 1$) to 21 p.m. ($j = 10$), covering the period 2005/01/02 to 2011/11/30. Hence, $d = 10$ and $n = 2520$. The covariates include day of the year (doy), day of the week (dow), exponentially smoothed temperature at the j-th hour (temp_j^S), and the hourly loads of the previous day (load_j^{24}).

To select the covariance matrix model, we first fit univariate Gaussian GAMs, $y_{ik} \sim \mathcal{N}(\mu_{ik}, \sigma_k^2)$, on 2005-2010. For the 2010 validation data we adopt a 1-month block rolling origin forecasting procedure starting from 2010/01/01. The mean vector components are

$$\mu_{ik} = f_{1k}^{10}(\text{doy}_i) + f_{2k}^{10}(\text{temp}_{ik}^S) + f_{3k}(\text{dow}_i) + f_{4k}(\text{load}_{ik}^{24}), \ k = 1, \ldots, d, \quad (2)$$

where f_{1k} and f_{2k} are smooth effects, with the superscript denoting the spline bases dimensions, while f_{3k} and f_{4k} are parametric linear effects. Gradient boosting is run for 10^4 iterations over 2005 - 2009 data, with a learning rate equal to 0.01, and spans the candidate effects of

$$\eta_{ik} = f_{1k}^{10}(\text{doy}_i) + f_{2k}^{5}(\text{temp}_{il_k}^S) + f_{3k}(\text{dow}_i), \ k = d+1, \ldots, q, \quad (3)$$

where f_{1k} and f_{3k} are as in (2), while $q = 65$ here. Denoting with l_k the column of $\log \mathbf{D}^2$ or $\mathbf{T}$ where the k-th linear predictor acts, we exploit the interpretation of the MCD parametrisation (see Pourahmadi, 1999) to specify the candidate temperature effects, involved in f_{2k}. E.g., the temperature at 12 p.m. is a candidate for modelling $\log \mathbf{D}_{11}^2$ and the first column of $\mathbf{T}$. Then, the cumulative log-likelihood gains of the effects selected for modelling the entries of the MCD parametrisation are computed and the effects are ordered in terms of decreasing importance.

Given the ordered list of effects obtained from the gradient boosting, we fit the MCD-based multivariate Gaussian additive models on a grid of the ordered list. Figure 1a shows the out-of-sample log-likelihood, which suggests including 40 effects in the model. Figure 1b shows the position of such effects on the MCD and summarises the results of the gradient boosting. While the day of the week effects are discarded by the selection procedure, the day of year effects are selected for modelling all the elements of $\log \mathbf{D}^2$ and mainly the first three sub-diagonals of $\mathbf{T}$. The temperature effects are mostly selected to model $\log \mathbf{D}^2$. Note that the entries of $\log \mathbf{D}^2$ and $\mathbf{T}$ are non-linearly related to one or more elements of the covariance matrix, hence it is not possible to interpret the effects acting on the MCD entries directly as variances and covariances.

The covariance matrix model obtained leveraging gradient boosting with learning rate of 0.01, henceforth `Parsimonious` (lr = 0.01), is compared with the `Static` covariance matrix model, with only intercepts in (3), and the `Full` model, where each MCD element is modelled using all the effects in (3). The resulting multivariate Gaussian additive models, with mean vector specified by (2), are evaluated on 2011 data, using a 1-month rolling origin forecasting procedure. The comparisons are in terms of the logarithmic

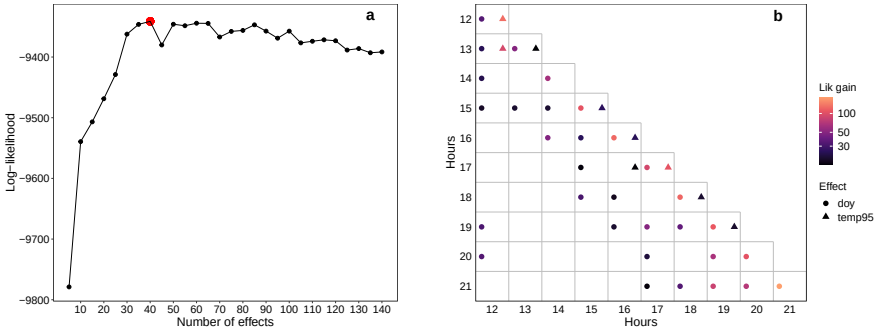


FIGURE 1. Gradient boosting results with learning rate equal to 0.01. a) out--of-sample log-likelihood; b) effects acting on $\log \mathbf{D}^2$ (diagonal) and $\mathbf{T}$ (off-diagonal). Colours are proportional to the cumulative log-likelihood gain.

(Log) and p-variogram score (Var $- p$, Scheuerer and Hamill, 2016), with $p = 0.5$ and $p = 1$. Table 1 shows that the `Parsimonious` (lr $= 0.01$) model is preferable to the `Full` one, and much superior to the `Static` model.

TABLE 1. Evaluation metrics. Underline indicates the best model.

Model	lr	Log	Var - 0.5	Var - 1
`Static`		11213.95	86.54	5334.76
`Parsimonious`	0.01	<u>11062.17</u>	<u>82.03</u>	<u>4982.96</u>
	0.10	11104.13	82.55	5042.40
`Full`		11064.26	82.16	5037.27

A sensitivity analysis is carried out for evaluating whether the model selection procedure is affected by the step length of the update. In particular, we increase the learning rate to 0.1, which is the default choice in many applications. Figure 2a shows the out-of-sample log-likelihood, which suggests including 15 effects in the covariance matrix model, i.e. much less than with a learning rate of 0.01. Figure 2b shows that all the effects are selected for modelling the $\log \mathbf{D}^2$ entries. It appears that a larger learning rate leads to more aggressive initial updates of the $\log \mathbf{D}^2$, thus overfitting occurs before the elements of $\mathbf{T}$ start to be modelled. The resulting MCD-based multivariate Gaussian additive model is then referred to as `Parsimonious` (lr $=$ 0.1). While the evaluation metrics on 2011 data for `Parsimonious` (lr $=$ 0.1), which are reported in Table 1, are slightly worse compared to both `Full` and `Parsimonious` (lr $= 0.01$) models, such a model might represent a sensible alternative when looking for a trade-off between parsimony and predictive performance.

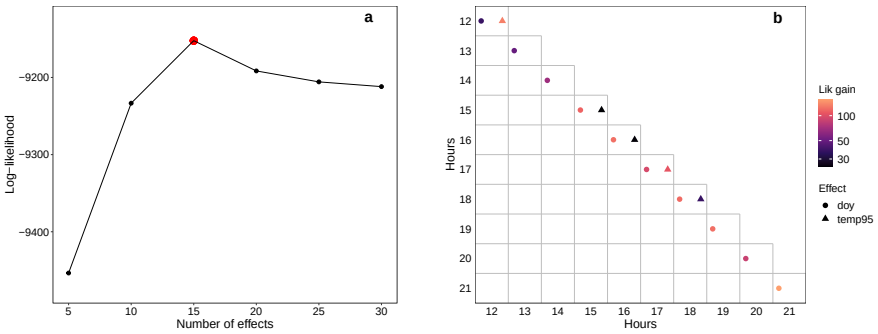


FIGURE 2. Gradient boosting results with learning rate equal to 0.1. a) out--of-sample log-likelihood; b) effects acting on $\log \mathbf{D}^2$ (diagonal) and $\mathbf{T}$ (off-diagonal). Colours are proportional to the cumulative log-likelihood gain.

4 Conclusions

The results so far show that gradient boosting is effective for ranking the effects to be used within highly parametrised additive covariance matrix models. The resulting automatic model selection approach is particularly useful for this class of models, where a manual parsimonious model specification is challenging. However, the sensitivity of the model selection procedure to the learning rate highlights the need for further studies on how to tune this parameter. This is the focus of current research. The **SCM** R package for fitting additive covariance matrix models is available at https://github.com/VinGioia90/SCM.

References

Gioia, V., Fasiolo, M., Browell, J. and Bellio, R. (2022). Additive covariance matrix models: Modelling regional electricity net-demand in Great Britain. arXiv preprint arXiv:2211.07451.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, **32**, 896 – 913.

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

Scheuerer, M. and Hamill, T. M. (2016). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, **143**, 1321 – 1334.

Strömer, A., Staerk, C., Klein, N., Weinhold, L., Titze, S., and Mayr, A. (2022). Deselection of base-learners for statistical boosting—with an application to distributional regression. *Statistical Methods in Medical Research*, **31**, 207 – 224.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., and Hofner, B. (2018). Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, **28**, 673 – 687.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. 2nd ed. Boca Raton, FL, USA: Chapman & Hall/CRC.

Wood, S. N., and Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*, **73**, 1071 – 1081.

Functional multilevel modelling of the influence of the menstrual cycle on the performance of female cyclists

Steven Golovkine[1], Tom Chassard[2], Alice Meignié[2], Emmanuel Brunet[3], Jean-Francois Toussaint[2], Juliana Antero[2]

[1] Department of Mathematics and Statistics, University of Limerick, Ireland
[2] Institut de Recherche BioMédicale et d'Épidémiologie du Sport, Paris, France
[3] Fédération Française de Cyclisme, Paris, France

E-mail for correspondence: steven.golovkine@ul.ie

Abstract: The relation between hormonal fluctuations along menstrual cycles and physical performance is of particular interest in sport science research. With the development of sensors technologies, the recording of large scale, high frequency performance data sets is now available. For cycling, performance can be measured using Mean Maximal Power curve. A functional linear mixed model is proposed to assess whether performance differs between phases of the menstrual cycle and how performance varies over the cycle based on the athletes, training intensities and types of the bike. Our methodology captures the continuous dynamic change characteristic of the data. The results indicate no difference in average performance between the phases. The performance variability is also similar for each phase. Most of the performance variability is induced by the differences between the athletes.

Keywords: Cycling; Functional Data Analysis; Menstrual Cycle; Mixed-effects Model; Performance Analysis

1 Introduction

Menstrual cycles affect women's health and wellness. Female sex hormones, and especially, estradiol and progesterone, fluctuate along the menstrual cycle (see Figure 1). These hormones affect multiple parameters on women ranging from adverse symptoms, such as fatigue, sleep disturbance or mood disorders along menstrual cycle phases (Pierson et al., 2021), to many beneficial cardiovascular, muscular and metabolic pa- rameters (Meignié et al.,

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2021). Performance-based research in women sport science is still scarce in regards to the influence of menstrual cycle phases (Meignié et al, 2021). Cycling is interesting to analyze the influence of hormonal fluctuation onto female performance. Mobile power meters are fitted to bicycles to measure the power delivered by cyclists during training. These data can be used to monitor and evaluate training performance. Mean Maximal Power (MMP) curves have been introduced to analyze power output profile at the individual level (Pinot and Grappe, 2010). MMP curves are defined as the maximal amount of power a cyclist can produce in a given period of time. We analyse whether performance, in terms of MMP, is influenced by the menstrual phases. We study performance variability with respect to menstrual cycle phases, athletes, rating of perceived exertion (RPE) using the Borg-CR10 scale (Borg, 1982) and types of the bike. We developed a functional linear mixed model to answer these questions.

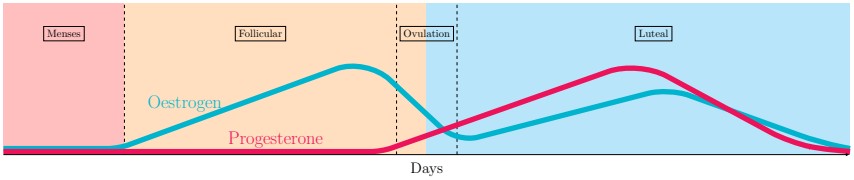


FIGURE 1. Schematic representation of the phases' division and hormonal fluctuations for naturally cycling women.

Power output data are recorded at 1Hz by personal powermeter. An MMP curve is derived from every individual training. Consider an exercise which last T seconds and $Z = \{z_t\}_{1 \leq t \leq T}$ a sequence of observation of the power output. Let $t_1, t_2 \in [\![1, T]\!]$, such that $t_2 - t_1$ is constant, an MMP curve is

$$X(t) = \max_{t_2 - t_1 = t} \frac{z_{t_1} + \cdots + z_{t_2}}{t_2 - t_1}, \quad t = 1, \ldots, T. \tag{1}$$

The data collection lasted from February 2021 to November 2022. Eight high-level female cyclists, with natural cycles, volunteered to participate in the study. To investigate how the menstrual cycle affects the performance of female cyclists, we estimated the different phase of the cycle for each athlete. We asked the cyclists to inform us of the start and end of their period, and we used a robust linear regression model (Soumpasis et al., 2020) to estimate the day of ovulation for each cycle. Their menstrual cycles are then divided into three phases: the menstruation phase, the follicular phase (between the end of the bleeding period and the estimated ovulation day), and the luteal phase (from the estimated ovulation day until the start of the next period). Prior to participation, all the athletes were informed about the purpose of the study. All investigations conformed to the code of ethics of the World Medical Association and were approved by the Institutional

Ethics Committee. Data collection was compliant with the General Data Protection Regulation (2016/679) applied in the European Union.

2 Model

The model is a hierarchical model that takes into account that observations depend on bike types, RPE and athletes and for each athlete we have repeated measurements for each of the three phases (menstrual, follicular and luteal). We assume that the RPE factors and bike type factors are crossed between athletes. This assumption is reasonnable since the factors are independent of the considered athlete. The factors are only partially crossed because we did not observed all the combinations of training intensity and bike type for all athletes. We consider the following model

$$X_{jklmn}(t) = \mu_k(t) + B_{jk}(t) + C_{lk}(t) + D_{mk}(t) + E_{jklmn}(t),\ t \in [\![1, T]\!], \quad (2)$$

where $j = 1, \ldots, 8$ (athletes), $k = 1, \ldots, 3$ (phases), $l = 0, \ldots, 10$ (RPE, Borg-CR10 scale), $m = 1, \ldots, 4$ (bike types), $n = 1, \ldots, N_{jklm}$ (observations). $X_{jklmn}(t)$ represents the MMP output of the observation n for athlete j during phase k, training intensity l and bike type m for a period of t seconds. $\mu_k(t)$ is the fixed effect for the phase of the menstrual cycle. $B_{jk}(t)$, $C_{lk}(t)$ and $D_{mk}(t)$ are a phase-specific functional random intercept for athletes, for RPE and for bike type respectively. $E_{jklmn}(t)$ is a smooth error term accounting for observation-specific variability. $B_{jk}(t)$, $C_{lk}(t)$, $D_{mk}(t)$ and $E_{jklmn}(t)$ are assumed to be centered and mutually uncorrelated. We allows the covariances of the functional random intercepts to be different for each phase. This assumption is motivated by the intra-phase variation (Figure 2) and by our aim to characterize this variability. The

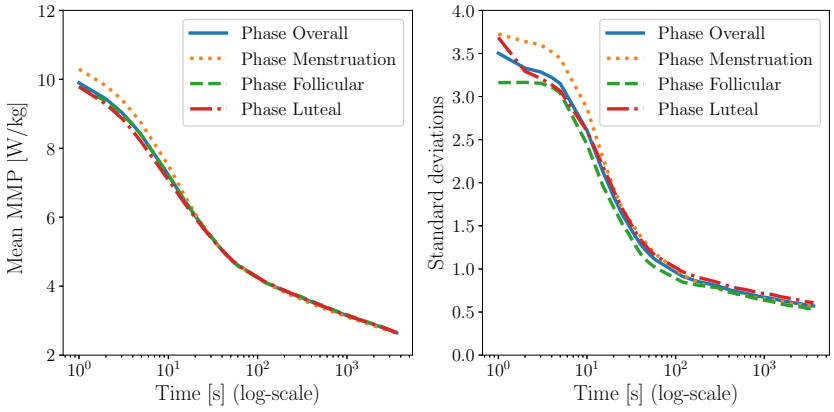


FIGURE 2. Point-wise mean curves (left) and standard deviation curves (right) per phase on a log-scale.

comparison of the fixed effects is performed using bootstrap estimation of the statistics

$$S_N = \frac{N_k N_{k'}}{NT} \|\mu_k - \mu_{k'}\|^2 = \frac{N_k N_{k'}}{NT} \int_{\mathcal{T}} (\mu_k(t) - \mu_{k'}(t))^2 \, \mathrm{d}t, \qquad (3)$$

where N, N_k and $N_{k'}$ are the total number of observations, the number of observation for phase k and k' respectively. The sampled bootstrap statistics, under the assumption of equality of the mean curves, are compared to S_N computed on the observed data. The estimation of the composents of the model is performed following Cederbaum (2017).

3 Results

For the comparison of the fixed effects, we generated 5000 bootstrap samples such that there is no difference between phases from the observed data to compare the mean MMP curves of the different phase. For each bootstrap sample, we computed the test statistic (3) for each combination of the cycle phases. Histograms of the resulted test statistics are plotted in Figure 3 with S_N computed on the observed data (plain line) and the 95%-quantile of the distribution of the test statistics computed on the bootstrap samples (dashed line). The test statistic computed on the observed data is smaller than the 95%-quantile of the distribution of the test statistics computed on the bootstrap samples for all phases comparison (Figure 3). There is thus no evidence of a difference between the phases considering their mean MMP curves. We fit the model (2) to all data with μ_k re-

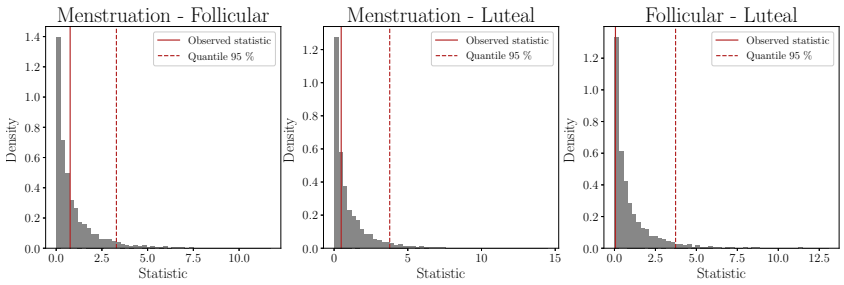


FIGURE 3. Histogram of the test statistic S_N computed on 5000 bootstrap samples. Difference between menstruation and follicular phases (left), menstruation and luteal phases (middle), follicular and luteal phases (right).

placed by a functional random intercept for the phases to obtain the full variance decomposition (Table 1). The curves are standardized and we set the percentage of variance explained to 99.999%. This decomposition highlights the importance of accounting for the different sources of variability as most of the overall variability is induced by the different observations.

TABLE 1. Full variance decomposition for a model with a functional random intercept for phase with pre-specified variance explained of 99.999%.

Variability source	Phase	Athlete	RPE
Variance explained (in %)	2.41×10^{-3}	22.0	11.5
Variability source	Bike type	Observation	Error variance
Variance explained (in %)	16.6	49.8	6.60×10^{-11}

The second most important source of variability is induced by the athletes. It appears that the different phases induce zero variation of power output. Part of the variability in the MMP curves is due to the training intensity (11.5%) and the bike type (16.6%). We have however not proven that there is no variation between phases, we have failed to find evidence of variation between phases. The athletes are thus likely to achieve their peak performance in each phase. These results may be helpful for coaches who use these curves for training planing or the comprehension of their athletes.

Acknowledgments: This study received funding from ANS and from INSEP. The authors thank all the athletes who participated in this study and the French Cycling Federation. S. Golovkine was partially supported by SFI under Grant No. 19/FFP/7002 and co-funded under the ERDF.

References

Borg, G. a. V. (1982). Psychophysical bases of perceived exertion. *Medicine & Science in Sports & Exercise*, **14**, 377 – 381.

Cederbaum, J. (2017). *Functional linear mixed models for complex correlation structures and general sampling grids.* Ph.D. thesis.

Meignié, A., Duclos, M., Carling, C., et al. (2021). The Effects of Menstrual Cycle Phase on Elite Athlete Performance: A Critical and Systematic Review. *Frontiers in Physiology*, **12**.

Pierson, E., Althoff, T., Thomas, D., et al. (2021) Daily, weekly, seasonal and menstrual cycles in women's mood, behaviour and vital signs. *Nature Human Behaviour*, **5**, 716 – 725.

Pinot, J. and Grappe, F. (2010). The 'Power Profile' for determining the physical capacities of a cyclist. *Computer Methods in Biomechanics and Biomedical Engineering*, **13**, 103 – 104.

Soumpasis, I., Grace, B. and Johnson, S. (2020) Real-life insights on menstrual cycles and ovulation using big data. *Human Reproduction Open.*

Confidence intervals for finite mixture regression based on resampling techniques

Colin Griesbach[1], Tobias Hepp[1,2]

[1] Georg-August-Universität Göttingen, Germany
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

E-mail for correspondence: colin.griesbach@uni-goettingen.de

Abstract: In this work we propose a resampling technique for finite mixture regression models to construct confidence intervals for the regression coefficients in order to hold the type-I error threshold. The routine relies on bootstrapping as intervals derived from standard regression theory tend to have insufficient coverage rates.

Keywords: Mixture Models; Resampling; Uncertainty.

1 Overview

Mixture Regression Models (McLachlan and Peel, 2000) are widely used to quantify associations between outcomes and various covariates in scenarios with unobserved heterogeneity. However, uncertainty estimates are not immediately available as regular statistical inference neglects any variance regarding class assignments yielding biased results (Grün and Leisch, 2008). This issue has been addressed for ordinary mixture models in Basford *et al.* (1997) or O'Hagan *et al.* (2019) by employing resampling techniques like various bootstrapping routines or the jackknife. In the case of mixture regression models, Grün and Leisch (2004) already used bootstrapping to detect identifiability issues of fitted mixture regression models. In this work, we propose a resampling approach for uncertainty estimates of regression parameters in finite mixture regression models. The method applies empirical bootstrapping and in addition uses a matching mechanism based on correlations of posterior class probabilities to aggregate estimates across all bootstrapping iterations and prevent label switching.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Resampling Method

Model Specification. For $\psi = (\pi_1, \ldots, \pi_K, \theta_1^T, \ldots, \theta_K^T)$ we consider the finite mixture models

$$h(y|x, \psi) = \sum_{k=1}^{K} \pi_k f(y|x, \theta_k)$$

with K components and prior class probabilities $\pi_k \geq 0$. Maximum likelihood estimation, usually achieved using an expectation–maximization (EM) algorithm, then yields estimates $\hat{\theta}_k$ and $\hat{\pi}_k$ with posterior class probabilities $\hat{\mathbf{P}} \in [0,1]^{n \times K}$ where n denotes the number of total observations.
Resampling Algorithm. The following algorithm highlights the proposed resampling routine.

Algorithm 1 Bootstrapped confidence intervals for finite mixture regression.

- **Initialize** a regular model fit to obtain estimates $\hat{\theta}_k$, $k = 1, \ldots, K$, and posteriors $\hat{\mathbf{P}}$. Choose number of bootstrap samples B.
- **for** $b = 1$ to B **do**

  Draw a new empirical bootstrap sample $(\boldsymbol{x}^b, y^b)$ and compute estimates $\hat{\theta}_k^b$, $k = 1, \ldots, K$, with posteriors $\hat{\mathbf{P}}^b$.

  Reorder estimates $\hat{\theta}_k^b \to \hat{\theta}_{\varrho(k)}^b$ where $\varrho(k)$ denotes the cluster of the initial model fit with highest correlation of posterior probabilities.

  end for

- **Compute** confidence intervals for each estimate $\hat{\theta}_k$ based on the quantiles of $(\hat{\theta}_k^1, \ldots, \hat{\theta}_k^B)$.

Computational Details. For the correlation matrix $\mathbf{C} = \mathrm{cor}(\hat{\mathbf{P}}\big|_b, \hat{\mathbf{P}}^b) \in [-1,1]^{K \times K}$, the reordering permutation ϱ can be formally expressed as

$$\begin{aligned} \varrho \colon \{1, \ldots, K\} &\to \{1, \ldots, K\}, \\ k &\mapsto \arg\max_l \ (c_{kl})_{l=1,\ldots,K}\,, \end{aligned}$$

where $\hat{\mathbf{P}}\big|_b$ denotes the initial posterior probabilities $\hat{\mathbf{P}}$ with (possibly duplicated) entries corresponding to the bth bootstrap sample and c_{kl} stands for the entry in the kth row and lth column of $\mathbf{C}$.
Furthermore, the starting values regarding class assignments in each bootstrap iteration b are the configurations of the initial model fit evaluated for the specific bootstrap sample $(\boldsymbol{x}^b, y^b)$.

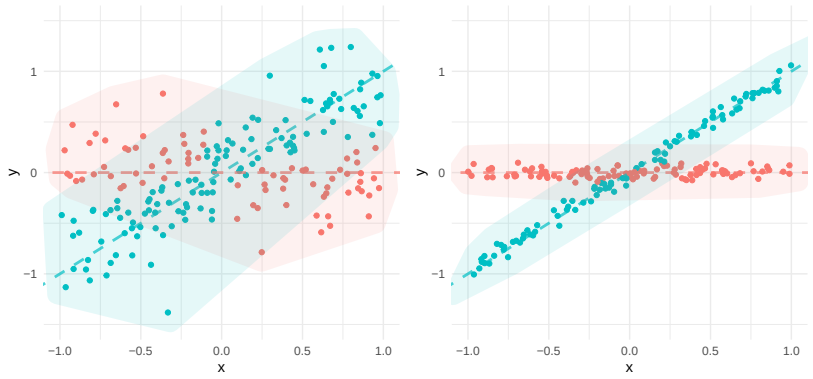


FIGURE 1. Depiction of the mixture data for average ($\sigma = 0.3$) and rather high ($\sigma = 0.05$) separability.

3 Evaluation

We evaluate the approach by considering $K = 2$ univariate normal densities $f(y|\beta_k x, \sigma^2)$ with equal prior probabilities and coefficients $\beta_1 = 0$ and $\beta_2 = 1$. Figure 1 depicts exemplary data for $\sigma \in \{0.05, 0.3\}$. Overall, we independently simulated 1000 datasets and applied Algorithm 1 with $B = 1000$ accordingly. Coverage rates for the so obtained confidence intervals with various confidence levels are displayed in Table 1.

While `flexmix` and Algorithm 1 reveal fairly desirable coverage rates in a scenario with high separability, `flexmix` clearly falls behind the corresponding thresholds for $\sigma = 0.3$. The confidence intervals obtained via

TABLE 1. Different coverage rates by `flexmix` and Algorithm 1 based on 1000 independent simulation runs.

		$\sigma = 0.3$		$\sigma = 0.05$	
$1 - \alpha$		$\beta_1 = 0$	$\beta_2 = 1$	$\beta_1 = 0$	$\beta_2 = 1$
80%	`flexmix`	0.691	0.693	0.760	0.795
	Algorithm 1	0.795	0.796	0.769	0.792
90%	`flexmix`	0.803	0.792	0.882	0.910
	Algorithm 1	0.919	0.908	0.879	0.912
95%	`flexmix`	0.866	0.863	0.935	0.953
	Algorithm 1	0.969	0.955	0.934	0.953
99%	`flexmix`	0.936	0.925	0.985	0.985
	Algorithm 1	0.995	0.990	0.982	0.982

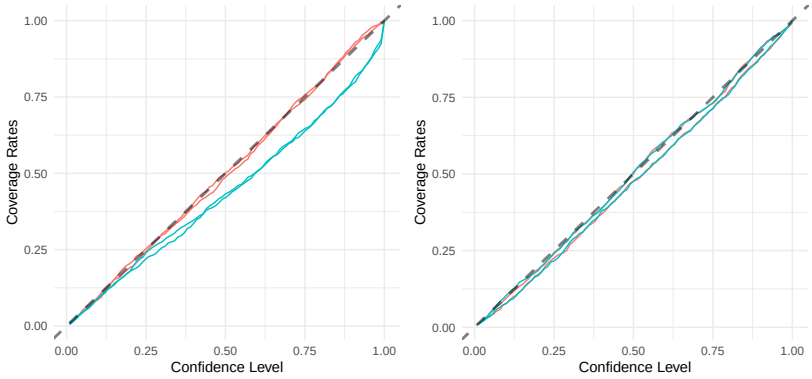


FIGURE 2. Depiction of the coverage rates as highlighted in Table 1 for average ($\sigma = 0.3$) and rather high ($\sigma = 0.05$) separability. Bootstrapped coverage rates are plotted in red, rates obtained by `flexmix` in blue.

bootstrapping on the other hand stay arguably closer to the given thresholds, which becomes even more clear in Figure 2. While the confidence intervals derived from standard theory clearly lack behind in the $\sigma = 0.3$ case, the bootstrapped confidence intervals match the desired optimum quite close.

4 Application

The *Seizure* data featured in the `flexmix` package consists of 140 daily observations of a single patient with number of epileptic seizures as outcome and a dummy for intravenous gamma-globulin as treatment covariate. A Poisson mixture regression with $K = 2$ and the covariates `treatment` and

TABLE 2. Effect estimates with corresponding 95% confidence intervals for the seizure data.

			(95%-CI)	
		$\hat{\beta}$	`flexmix`	Algorithm 1
Component 1	`treat`	−0.49	(−0.93; −0.05)	(−1.37; 0.07)
	`day`	−0.03	(−0.04; −0.02)	(−0.04; −0.01)
	`treat:day`	−0.01	(−0.02; −0.00)	(−0.03; 0.01)
Component 2	`treat`	−0.18	(−0.53; 0.16)	(−0.67; 0.20)
	`day`	−0.03	(−0.05; −0.01)	(−0.06; −0.00)
	`treat:day`	−0.02	(−0.01; 0.04)	(−0.01; 0.05)

`day` (of observation) yields the parameter estimates with corresponding 95% confidence intervals depicted in Table 2. Similar to the simulation study, the confidence intervals obtained via resampling are wider as the ones by standard inference and occasionally even cross the border of significance on the $\alpha = 0.05$ level, which indicates possibly false positive findings.

5 Summary and Outlook

While resampling techniques have already been applied for mixture regression models regarding detection identifiability problems, they had yet to be used regarding the estimation of confidence intervals. The proposed algorithm in Section 2 relies on well established bootstrapping routines and is, due to the tweaks with respect to relabeling and starting values, capable of producing reliable confidence intervals which very accurately hold the type-I error threshold as revealed by simulations and the *Seizure* application. Further investigations could, among other aspects, focus on more challenging and flexible model setups or an extension of the resampling concept to other model classes like latent class analysis.

Acknowledgments: The work on this article was supported by the DFG (Number 426493614) and the Volkswagen Foundation (Freigeist Fellowship).

References

Basford, K., Greenway, D., McLachlan G. and Peel, D. (1997). Standard Errors of Fitted Component Means of Normal Mixtures. *Computational Statistics*, **12(1)**, 1 – 17.

Grün, B. and Leisch, F. (2004). Bootstrapping Finite Mixture Models. In: *COMPSTAT 2004 – Proceedings in Computational Statistics*. Heidelberg: Physica Verlag, 1115 – 1122.

Grün, B. and Leisch, F. (2008). FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, **28(4)**, 1 – 35.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* New York: John Wiley and Sons Inc.

O'Hagan, A., Murphy, T., Scrucca, L. and Gormley, I. (2019). Investigation of parameter uncertainty in clustering using a Gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Computational Statistics*, **34**, 1779 – 1813.

Component-wise boosting for mixture distributional regression models

Tobias Hepp[1,2], Jakob Zierk[3], Elisabeth Bergherr[2]

[1] Institut für Medizininformatik, Biometrie und Epidemilogie, Universität Erlangen-Nürnberg, Germany
[2] Professur für Raumbezogene Datenanalyse und Statistische Lernverfahren, Universität Göttingen, Germany
[3] Kinder- und Jugendklinik, Universitätsklinikum Erlangen, Germany

E-mail for correspondence: `tbs.hepp@fau.de`

Abstract: In this work, we present a gradient boosting algorithm for the estimation of finite mixture regression models. A first version of the algorithm is demonstrated on a laboratory dataset for hemoglobin values and performs on par with alternative strategies. In addition, an outlook on variable selection performance is given using a small simulation study.

Keywords: Gradient boosting; Mixture models; Distributional regression

1 Introduction

Applying statistical models to a dataset comprised of $i = 1, \dots, n$ observations usually requires the assumption of a probability density function to describe the (conditional) distribution of the variable of interest. However, in the case of unobserved heterogeneity, e.g. if the data consists of two or more unlabeled sub-populations, a single density function is not sufficient. Then, given the number of latent components M, a weighted sum of $m = 1, \dots, M$ probability density functions $f_m(x^{(i)}, \boldsymbol{\theta}_m)$ with parameter vectors $\boldsymbol{\theta}_m = (\theta_1, \dots, \theta_{K_m})$ of distribution-dependent size K_m can be used to construct a finite mixture distribution as

$$f(y^{(i)}) = \sum_{m=1}^{M} \alpha_m f_m(y^{(i)}, \boldsymbol{\theta}_m). \tag{1}$$

With $\alpha_m > 0$ and $\sum_{m=1}^{M} \alpha_m = 1$, $f(x^{(i)})$ is a convex combination of all

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

$f_m(y^{(i)}, \boldsymbol{\theta}_m)$ and a probability density function itself. The overall number of distribution parameters across all components is then $K = \sum_{m=1}^{M} K_m$. Mixture regression models (DeSarbo *et al.*, 1988) extend the basic mixture formula (Eq. 1) by allowing one or more of the distribution parameters to be functions of the observed covariate vector $\boldsymbol{x}^{(i)}$:

$$f\left(y^{(i)}\middle|\boldsymbol{x}^{(i)}\right) = \sum_{m=1}^{M} \alpha_m\left(\boldsymbol{x}^{(i)}\right) f_m\left(y^{(i)}, \boldsymbol{\theta}_m\left(\boldsymbol{x}^{(i)}\right)\right) \tag{2}$$

Component-wise boosting algorithms (Bühlmann and Yu, 2003) provide a versatile estimation approach suitable for a large variety of statistical models in potentially high-dimensional data settings. In the following, we provide a first working algorithm applied for the indirect estimation of diagnostic reference distributions and compare it to results obtained from expectation-maximization and mixture density networks (Bishop, 1994). Moreover, a small simulation is conducted to investigate variable selection performance via early stopping.

2 Methods

The core concept of boosting algorithms is to repeatedly fit simple "base-learners" to the data with each iteration emphasizing areas of the response that are still insufficiently predicted. In the component-wise approach, these base-learners are usually regression-type functions (e.g. linear effects or splines) fit to the negative gradient(s) of a specified loss-function. In each iteration, the algorithm updates only the best performing base-learner, successively expanding the model in the direction where the loss is reduced most. For the general mixture regression model presented in Equation (2), the loss is the negative loglikelihood

$$-\ln \mathcal{L} = -\ln\left(\prod_{i=1}^{n}\sum_{m=1}^{M} \hat{\alpha}_m\left(\boldsymbol{x}^{(i)}\right) f_m\left(y^{(i)}, \hat{\boldsymbol{\theta}}_m\left(\boldsymbol{x}^{(i)}\right)\right)\right)$$

with hats indicating the parameters to be estimated from the data. This is performed via $j = 1, \ldots, (K + M)$ additive predictors

$$g_j(\eta_j^{(i)}) = \beta_{0j} + \sum_{l \in B(j)} h_{lj}(x_l^{(i)})$$

with $B(j)$ comprising the indices of the covariates used for the base-learners $h_{lj}()$ and a link function $g_j()$ that transforms the input to the domain of the corresponding parameter.
In the following, we consider a Gaussian mixture regression setup with all $f_m\left(y^{(i)}, \hat{\boldsymbol{\theta}}_m\left(\boldsymbol{x}^{(i)}\right)\right) := \mathcal{N}\left(y^{(i)}, \hat{\mu}_m\left(\boldsymbol{x}^{(i)}\right), \hat{\sigma}_m\left(\boldsymbol{x}^{(i)}\right)^2\right)$, where $\mathcal{N}()$ denotes

the Gaussian density function with mean and variance depending on the covariates. As a consequence, $K_m = 2$ for all m, resulting in $(2+1)M$ parameters to be estimated via the additive predictors $\hat{\eta}_{\mu_m}^{(i)}$, $\hat{\eta}_{\sigma_m}^{(i)}$ and $\hat{\eta}_{\alpha_m}^{(i)}$. In the t-th iteration, the algorithm first computes the current mixture weights $\alpha_m^{(i)[t]}$ from all $\hat{\eta}_{\alpha_m}^{(i)[t-1]}$ using the softmax function

$$\alpha_m^{(i)[t]} = \frac{\exp\left(\hat{\eta}_{\alpha_m}^{(i)[t-1]}\right)}{\sum_{l=1}^{M} \exp\left(\hat{\eta}_{\alpha_m}^{(i)[t-1]}\right)}$$

and the current posterior probabilities

$$\pi_m^{(i)[t]} = \frac{\alpha_m^{(i)[t]} \mathcal{N}\left(y^{(i)}, \hat{\eta}_{\mu_m}^{(i)[t-1]}, \exp\left(\hat{\eta}_{\sigma_m}^{(i)[t-1]}\right)^2\right)}{\sum_{l=1}^{M} \alpha_m^{(i)[t]} \mathcal{N}\left(y^{(i)}, \hat{\eta}_{\mu_m}^{(i)[t-1]}, \exp\left(\hat{\eta}_{\sigma_m}^{(i)[t-1]}\right)^2\right)}.$$

This allows the calculation of the gradients to be fitted by the base-learners from

$$u_{\alpha_m}^{(i)[t]} = \alpha_m^{(i)[t]} - \pi_m^{(i)[t]}$$

$$u_{\mu_m}^{(i)[t]} = \pi_m^{(i)[t]} \left(\frac{\hat{\eta}_{\mu_m}^{(i)[t-1]} - y^{(i)}}{\exp\left(\hat{\eta}_{\sigma_m}^{(i)[t-1]}\right)^2} \right)$$

$$u_{\sigma_m}^{(i)[t]} = -\pi_m^{(i)[t]} \left(\frac{\left(\hat{\eta}_{\mu_m}^{(i)[t-1]} - y^{(i)}\right)^2}{\exp\left(\hat{\eta}_{\sigma_m}^{(i)[t-1]}\right)^3} - \frac{1}{\exp\left(\hat{\eta}_{\sigma_m}^{(i)[t-1]}\right)} \right)$$

Updating component-wise boosting algorithms with a single gradient is relatively straightforward, but adaptive learning rates may be advantageous in settings with multiple additive predictors (Zhang *et al.*, 2022). While these considerations most definitely play an important role for mixture distributional regression models, this first implementation is based on the original strategy for boosted distributional regression models (Mayr *et al.*, 2012). As a consequence, each iteration of the algorithm updates all predictors with the base-learner performing best for the corresponding gradient using a fixed learning rate of 0.1. For this approach to work, different offsets are required for the location parameters of the latent components, as the algorithm would otherwise not be able to differentiate between them. Therefore, we currently use different quantiles of the outcome variable based on the number of components K (e.g. 33% and 66% for $K = 2$).

3 Hemoglobin data and variable selection

In order to asses the suitability of our algorithm, we demonstrate its application to a laboratory dataset of the hemoglobin concentration in blood

samples from girls. Figure 1 provides an illustration of the results together with an comparison to solutions from earlier implementations based on expectation-maximization and neural networks (Hepp *et al.*, 2022).

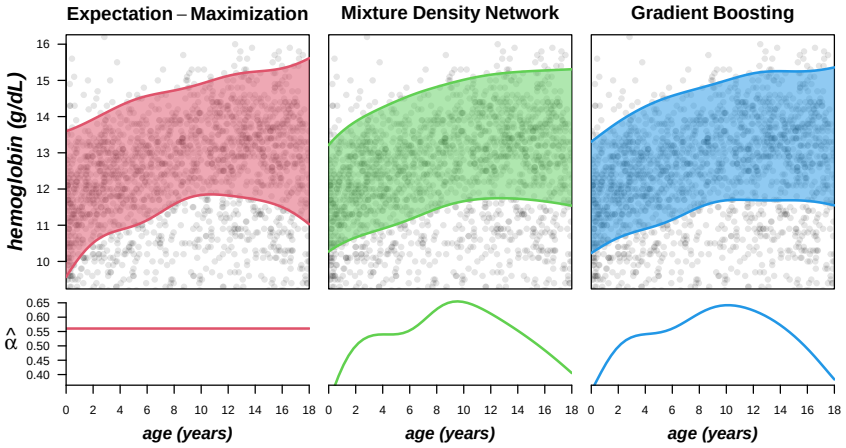


FIGURE 1. Top row: Age-dependent reference limits (i.e. 2.5% and 97.5% quantiles) for the latent distribution of healthy hemoglobin concentration values of girls estimated via three different methods. Bottom row: Corresponding estimates of the mixture component weights.

The implemented models use $M = 2$ components to separate the unlabeled healthy and pathological samples in order to determine a diagnostic reference distribution for the healthy component. Ultimately, all solutions cover rather similar areas describing the age-dependent range of the healthy measurements. Moreover, the implemented gradient boosting algorithm results in very similar estimates for the non-linear dependency of the component weights (i.e. the ratio of healthy samples) compared to the mixture density network.

While this presents a promising outlook on the general suitability of boosting algorithms in settings with unobserved heterogeneity, the advantage of boosting over other methods is of course the ability to perform variable selection simultaneous to parameter estimation. In the first steps to evaluate this feature, we conducted a simulation study with data generated from a two-component Gaussian mixture with focus on the location parameters. For this purpose, each repetition samples $l = 1, \ldots, p$ covariates $\boldsymbol{x}_l$ from independent standard normal distributions and subsequently uses two different models

$$\boldsymbol{y}_m \sim \mathcal{N}\left(\beta_{0m} + Z\boldsymbol{\beta}_m, \sigma_m\right)$$

to generate the outcome variable, where the matrix Z comprises the first five covariates $Z = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5]$, $\boldsymbol{\beta}_1 = (-1, 2, 3, 0, 0)$ and $\boldsymbol{\beta}_2 = (0, 0, -1, 2, 3)$.

All additional covariates are pure noise variables and not correlated to either latent component. Further, $\beta_{01} = -1$, $\beta_{02} = 1$ and the standard deviations are defined as $\sigma_1 = 1.5$ and $\sigma_2 = 3$ and completely independent from the covariates. In all runs, the total number of observations is $n = 250$ with $n_1 = 150$ and $n_2 = 100$, resulting in fixed mixture weights $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$. For our proposed boosting algorithm, we used five-fold cross-validation to find the best stopping iteration with respect to the average predictive performance on the test folds. Then, we calculated the mean squared error of the estimated regression coefficients with respect to both $\boldsymbol{\beta}_m$ to investigate if the solutions computed at the final iteration can be distintively assigned to one of the true latent components, which was in fact possible in all runs and settings. A summary of the results averaged over all 100 simulation runs is provided in Table 1.

TABLE 1. Variable selection performance averaged over 100 simulation runs. TPR: True positive rate, FPR(a): False positive rate for covariates correlated to the other component, FPR(b): False positive rate, Size: Number of non-zero coefficients

p	m	TPR	FPR(a)	FPR(b)	FDR	Size
10	1	1	0.745	0.756	0.615	8.27
	2	0.957	0.45	0.306	0.383	5.3
100	1	0.973	0.145	0.054	0.557	8.32
	2	0.803	0.22	0.034	0.544	6.04

Looking at the true positive rates, i.e. the proportion of informative variables selected, the proposed boosting algorithm reliably identifies the relevant variables for each component, but better so for the first latent model $m = 1$. This is not very surprising considering that this component is based on the larger part of the sampled data (60%) and σ_1 is only half the size of σ_2. However, this may be a disadvantage when it comes to the selection of false positives, where we differentiate between the false positive rate of the variables that are informative for the other component in FDR(a) and those not related to both in FDR(b). Here, the relatively higher uncertainty seems to prevent the inclusion of too many variables, as can be noted from the smaller average model size. As a consequence, this also results in less false positive selections in all but one comparison. While for the setting with only $p = 10$ covariates the difference between the two FPR's is not as obvious, increasing p to 100 reveals a clear tendency of the current algorithm to falsely select variables that are only relevant for the other component compared to the completely non-informative variables, with the difference more pronounced in the second component. Finally, differences in the false discovery rate, i.e. the number of false positives in the set of selected variables, can be noted for $p = 10$, but the average rates are about equal for

$p = 100$. This can most likely be traced back to the stronger decrease in the TPR for the second component together with its minor increase in overall number of selected variables.

4 Conclusion

In this short presentation of our proposed algorithm, we demonstrate that component-wise gradient boosting algorithms are generally capaple of estimating the latent structure of mixture distributional regression models. This provides a promising outlook to a more thorough analysis regarding different initialization and updating strategies necessary to evaluate the variable selection performance of boosting also with respect to the estimation of dependency patterns between covariates, scale parameters and mixture weights in potentially high-dimensional settings.

Acknowledgments: The work on this article was supported by the DFG (Project HE 9468/1-1) and Volkswagen Foundation (Freigeist Fellowship).

References

Bishop, C.M. (1994): *Mixture density networks*. NCRG/94/004, Aston University (Technical report).

Bühlmann, P. and Yu, B. (2003) Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association*, **98**, 324 – 339.

DeSarbo, W.S. and Cron, W.L. (1988): A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, **5**(2), 249 – 282.

Hepp, T., Zierk, J., Rauh, M., Metzler, M. and Seitz, S. (2022): Mixture density networks for the indirect estimation of reference intervals. *BMC Bioinformatics*, **23**, 524.

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 – 27.

Zhang, B., Hepp, T., Greven, S. and Bergherr, E. (2022): Adaptive steplength selection in gradient boosting for Gaussian location and scale models. *Computational Statistics*, **37**, 2295 – 2332.

Fusion, smoothing and model selection for item-on-item regression

Aisouda Hoshiyar[1], Jan Gertheiss[1]

[1] Department of Mathematics and Statistics, School of Economics and Social Sciences, Helmut Schmidt University, Hamburg, Germany

E-mail for correspondence: aisouda.hoshiyar@hsu-hh.de

Abstract: Motivated by a survey on the willingness to pay for luxury food products, which consists of various Likert-type items, we present a penalized version of ordinal-on-ordinal regression in the framework of cumulative logit models. By use of difference penalties on neighboring dummy coefficients, thus taking the predictors' ordinal structure into account, we provide a group lasso-type penalty for smoothing and selection of ordinal predictors and a fused lasso penalty for fusion and selection.

Keywords: Cumulative Logit; Fused Lasso; Group Lasso; Proportional Odds Model; Likert-Scale

1 Introduction

We consider a study concerning the segmentation of German consumers based on the perceived dimensions of "luxury food" (Hartmann et al., 2016). The aim of the study was to investigate the perceived dimensions of luxury food and the shift of consumer consumption motives toward indulgence, quality, and sustainability. The part of the dataset we examine consists of 821 observations of 44 Likert-type items on eating and shopping habits, diet styles, price, and luxury statements in general. One such items is, for example, "I particularly associate high quality with luxury", with coding scheme -2 = 'not true at all', ..., 2 = 'absolutely true'. Our response of interest is the willingness to pay for luxury food products, i.e., whether participants would be willing to pay a higher price for a food product that they associate with luxury, measured again on the ordinal -2 to 2 scale. High-dimensional surveys like this, with ordinal data both on the left and right-hand side of the regression equation, highlight the importance of

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

penalization in ordinal regression such as the cumulative logit model, because the full model is often hard to fit by ordinary/unpenalized maximum likelihood, and not all of the items available will be relevant for the response. We aim to investigate the ordinal covariates' effect on the ordinal response while integrating strategies for variable selection and smoothing across covariate levels. At IWSM 2022, we presented a poster on a selection approach for item-on-item regression (Hoshiyar and Gertheiss, 2022).
In short, the logistic group lasso estimator $\boldsymbol{\beta}_\lambda$ for variable selection (Meier et al., 2008; Yuan and Lin, 2006) was extended to the class of cumulative logit models, with our estimator being the minimizer of the function $l_\lambda(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} J_j(\boldsymbol{\beta}_j)$, where $l(\boldsymbol{\beta})$ is the log-likelihood of the unknown parameters in the proportional odds model. In order to take into account the ordinal structure of the covariates, as proposed by Tutz and Gertheiss (2016), we modified the usual L_2-norm by the first-order difference penalty

$$J_j(\boldsymbol{\beta}_j) = \sqrt{\left\{ \sum_{l=2}^{k_j} \mathrm{df}_j (\beta_{jl} - \beta_{j,l-1})^2 \right\}}, \tag{1}$$

with β_{jl} being the dummy coefficient of level l of covariate x_j, k_j the number of corresponding levels, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jk_j})^\top$, and $\mathrm{df}_j = k_j - 1$ being the respective degrees of freedom. For identifiability, we use the constraint $\sum_l \beta_{jl} = 0 \, \forall j$.
In some applications, however, we may be rather interested in collapsing certain categories instead of (quadratic) smoothing (1). This type of clustering can be done by a fused lasso penalty using the L_1-norm on adjacent categories. We hence extend the fused lasso (Tibshirani et al., 2005; Tutz and Gertheiss, 2016)

$$J_j(\beta_j) = \sum_{l=2}^{k_j} |\beta_{jl} - \beta_{j,l-1}| \tag{2}$$

to the framework of the cumulative logit model. Penalty (2) has the effect that neighboring categories may be fused; namely, they may have exactly the same β-values as a result of penalized maximum likelihood fitting. Moreover, the fused lasso also enforces variable selection, as a covariate is excluded if all its categories are combined into one cluster. The proposed method will soon be made available in the `R` package `ordPens`, which already offers fusion and selection for other (generalized) linear models (Gertheiss and Hoshiyar, 2021; Hoshiyar, 2021).

2 Numerical Experiment

We carried out a simulation study to investigate the properties of the proposed ordinal-on-ordinal selection approach using penalties (1) and (2).

We assumed to have $p = 50$ ordinally scaled covariates, with effects of $x_1, \ldots, x_4$ being non-monotone, effects of $x_5, \ldots, x_8$ being monotone but non-linear, and the effects of $x_9, \ldots, x_{12}$ being linear across categories. The remaining 38 covariates were irrelevant, i.e., with effects being zero. Factor levels were randomly drawn from $\{1, \ldots, 5\}$, meaning that each covariate had the same number of levels. The (true) effects for some covariates are shown in Figure 1(a). Using the 12 relevant predictors, we constructed the ordinal response with 5 levels through a cumulative logit model. We considered three different sample sizes $n = 200, 500, 1000$. We then assigned ranks to the predictor variables according to the order of non-zero ordinal group lasso coefficients in the coefficient path, and call this ordinal rank selection (ORS). We proceeded in a similar manner with the fused lasso estimates, and call this ordinal rank fusion (ORF). For comparison, we also fit a proportional odds model using `polr()` from R package `MASS` (Venables and Ripley, 2002; R Core Team, 2021) with forward stepwise selection. To evaluate the methods' performance, we constructed the Receiver Operating Characteristic (ROC) by varying selection thresholds and calculated the Area Under the Curve (AUC) in each of 100 iterations of our simulation, which is illustrated in Figure 1(b)–(d) for sample size $n = 500$. It is seen that results for ORF are very similar to those of ORS. `polr` estimation failed in 37 of the datasets (zero cases with $n = 1000$, but all cases with $n = 200$). Taking only the 63 successful runs into account, mean and median AUC were 0.908 and 0.919, respectively. In summary, the ordinal penalties proposed here (ORS and ORF) worked very well and both accounted for the ordinal-on-ordinal structure. Further simulation studies showed that `polr` only worked well when the sample size was large enough and failed otherwise. Ordinal selection and ordinal fusion worked equally well and outperformed `polr` by far in most scenarios considered (details not shown here).

3 Case Study: Spending on Luxury Food

Figure 2 illustrates the estimated coefficients of selected covariates and different values of tuning parameter λ when applying the ordinal lasso to the luxury food data (top: cumulative fused lasso; bottom: cumulative group lasso). For smaller λ (light gray), the estimates are more wiggly and become more and more smoothed out/shrunk as λ increases. Willingness to pay tends to increase among people who eat out frequently in expensive restaurants (a), prefer vegetarian food (b), and associate high quality with luxury (c). If using the fused lasso and choosing the amount of penalty by (5-fold) cross-validation, we obtain an optimal λ around $18.5/n$. For the variable "Vegetarian", for example, it is seen that most categories are fused. If using the ordinal group lasso instead, the optimal/cross-validated λ is around $14.5/n$.

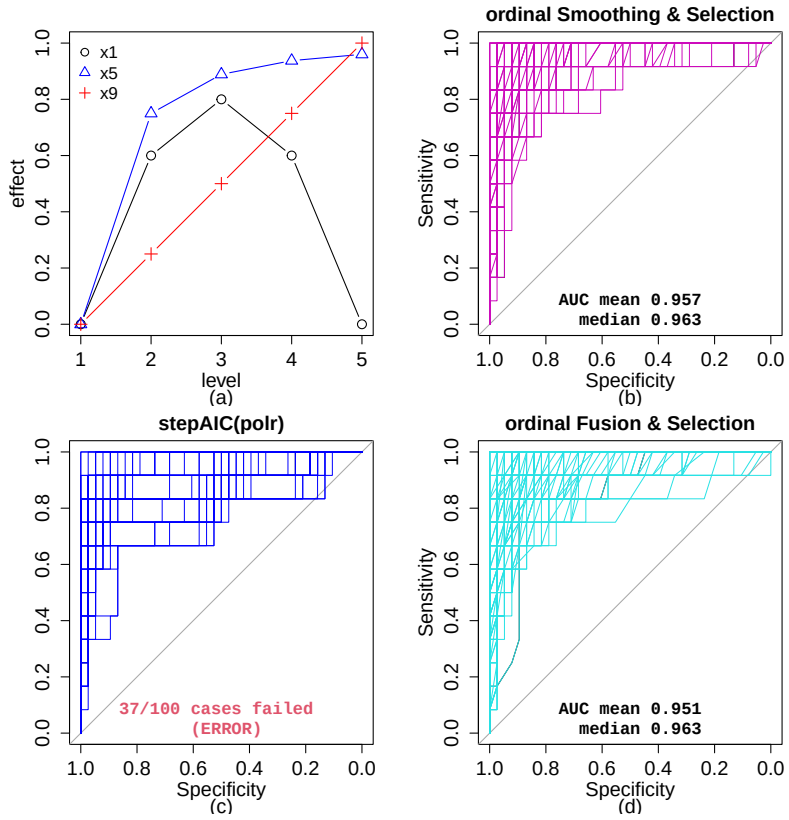


FIGURE 1. (a) True effects, exemplarily for influential predictors x_1, x_5, x_9. (b) ROC curves when using ORS with $n = 500$; (c) ROC curves when using `polr` and stepwise selection; (d) ROC curves when using ORF; results based on 100 simulated data sets.

References

Gertheiss, J., Hoshiyar, A. (2021). *ordPens: Selection, Fusion, Smoothing and Principal Components Analysis for Ordinal Variables*. R package version 1.0.0. `https://CRAN.R-project.org/package=ordPens`

Hartmann, L.H., Nitzko, S., Spiller, A. (2016). The significance of definitional dimensions of luxury food. *British Food Journal*, **118**, 1976–1998.

Hoshiyar, A. (2021). ordPens: An R package for Selection, Smoothing and Principal Components Analysis for Ordinal Variables. *Journal of Open Source Software*, **6**(68), 3828

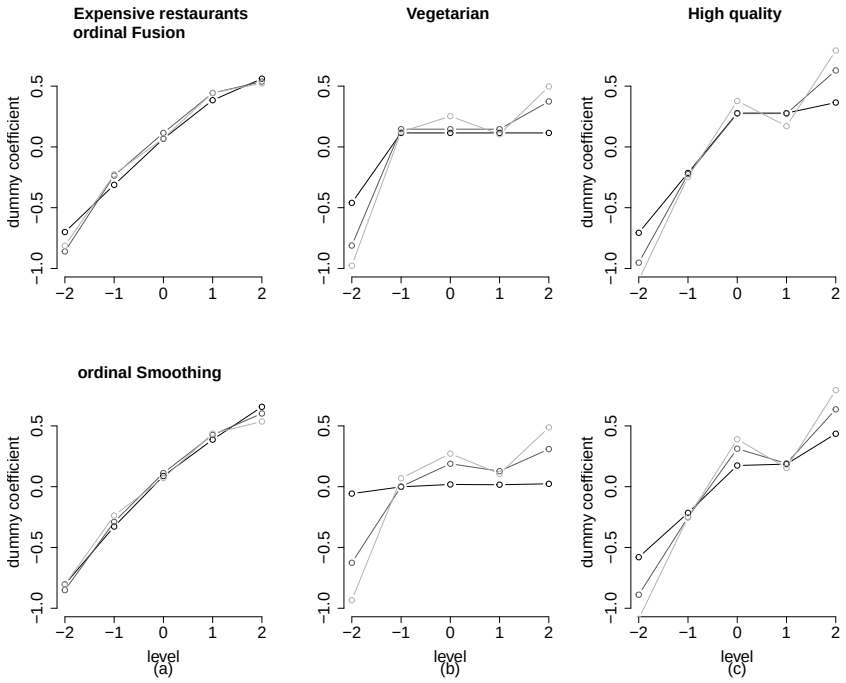


FIGURE 2. Top: Cumulative fused lasso estimates of dummy coefficients as functions of class labels ($\lambda \in \{15, 5, 1\}/n$). Bottom: Cumulative group lasso estimates. Light gray indicates smaller λ.

Hoshiyar, A., Gertheiss, J. (2022). Regularization and Model Selection for Item-on-Item Regression. In: *Proceedings of the 36th International Workshop on Statistical Modelling: July 18-22, 2022 Trieste, Italy*, 467 – 471

Meier, L., Van De Geer, S., Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society B*, **70**, 53 – 71.

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Kneight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B*, **67**, 91 – 108.

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

Venables, W.N., Ripley, B.D. (2002). *Modern Applied Statistics with S*. 4th Edition. New York: Springer.

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

Induced nonparametric ROC surface regression

Vanda Inácio[1], María Xosé Rodríguez-Álvarez[2]

[1] School of Mathematics, University of Edinburgh, UK
[2] CINBIO, Universidade de Vigo, Department of Statistics and Operations Research, Spain

E-mail for correspondence: vanda.inacio@ed.ac.uk

Abstract: We develop semiparametric inference for the covariate-specific receiver operating characteristic (ROC) surface, a popular tool for evaluating the accuracy of diagnostic tests measured on a continuous scale when there exist three ordered disease groups. In our application we seek to assess if and how the accuracy of a potential biomarker of Alzheimer's disease (AD) to distinguish between individuals with normal cognition, mild cognitive impairment, and dementia, changes with age and gender.

Keywords: Location-scale regression model; Optimal thresholds; Penalised-splines, Receiver operating characteristic surface; Volume under the surface.

1 Introduction

Before a test is routinely used in practice, its ability to distinguish between different disease stages must be rigorously evaluated. As a direct generalisation of ROC curves, ROC surfaces have been developed to evaluate the accuracy in ordered three-class diagnostic problems. It is well recognised that the performance of a test may be impacted by covariates (e.g., age and gender) and that ignoring covariate information may lead to incorrect conclusions about a test's discriminatory ability. Although there is now a quite extensive literature of methods for accommodating covariates in ROC curves, approaches for ROC surface regression are scarse. In this work, we develop a flexible approach to estimate the covariate-specific ROC surface that relies on modelling the relationship between test outcomes and covariates, in each of the three groups, through a location-scale regression model where the mean and variance functions are estimated with penalised-splines. In addition, estimation of the distribution function of the regression

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

errors, needed to compute the induced covariate-specific ROC surface, is done via a smoothed version of the distribution function of the standardised residuals. As a result, estimates of the covariate-specific ROC surface are smooth without the need to specify a parametric distribution. This is appealing as in practice we do not expect the accuracy of the test to change abruptly for close threshold values.

2 Induced nonparametric inference for the covariate-specific ROC surface and its functionals

2.1 Background

Let Y_1, Y_2, and Y_3 be continuous random variables denoting the outcomes of the test in each of the three groups (e.g., in our AD application, index 1 stands for normal cognition, index 2 for mild cognitive impairment, and index 3 for dementia) and let $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ be covariate vectors (the same in the three groups). For a given covariate value $\mathbf{x}$, the covariate-specific true classification fraction (TCF), for a given pair of threshold values (c_1, c_2), $c_1 < c_2$, for each of the three groups, can be written as

$$\begin{aligned}
\mathrm{TCF}_1(c_1, c_2 \mid \mathbf{x}) &= \Pr(Y_1 < c_1 \mid \mathbf{X}_1 = \mathbf{x}) = F_1(c_1 \mid \mathbf{X}_1 = \mathbf{x}),\\
\mathrm{TCF}_2(c_1, c_2 \mid \mathbf{x}) &= \Pr(c_1 \le Y_2 < c_2 \mid \mathbf{X}_2 = \mathbf{x}) = F_2(c_2 \mid \mathbf{X}_2 = \mathbf{x}) -\\
&\qquad\qquad F_2(c_1 \mid \mathbf{X}_2 = \mathbf{x}),\\
\mathrm{TCF}_3(c_1, c_2 \mid \mathbf{x}) &= \Pr(Y_3 \ge c_2 \mid \mathbf{X}_3 = \mathbf{x}) = 1 - F_3(c_2 \mid \mathbf{X}_3 = \mathbf{x}),
\end{aligned}$$

where $F_d(y \mid \mathbf{x}) = \Pr(Y_d < y \mid \mathbf{X}_d = \mathbf{x})$, for $d \in \{1, 2, 3\}$. In this setting, for each possible $\mathbf{x}$, the covariate-specific ROC surface measures the amount of separation of the conditional distribution of test outcomes in the three groups. In particular, the covariate-specific ROC surface is the plot in the unit cube depicting the covariate-specific TCFs in each group as the thresholds c_1 and c_2 vary. By specifying the TCF in group 1, say p_1, it is possible to define the covariate-specific threshold values $c_{1,\mathbf{x}}^{p_1}$ which give rise to a TCF of p_1 in each of the subpopulations defined by the covariates, i.e., $c_{1,\mathbf{x}}^{p_1} = F_1^{-1}(p_1 \mid \mathbf{x})$. In the same vein, if we let the TCF in group 3 to be p_3, then $c_{2,\mathbf{x}}^{p_3} = F_3^{-1}(1 - p_3 \mid \mathbf{x})$ is the covariate-specific threshold that leads to a TCF of p_3 in each of the subpopulations defined by the covariates. The covariate-specific ROC surface is therefore defined as

$$\begin{aligned}
\mathrm{ROCS}(p_1, p_3 \mid \mathbf{x}) &= \Pr(c_{1,\mathbf{x}}^{p_1} \le Y_2 < c_{2,\mathbf{x}}^{p_3} \mid \mathbf{X}_2 = \mathbf{x})\\
&= F_2\left\{F_3^{-1}(1 - p_3 \mid \mathbf{x}) \mid \mathbf{x}\right\} - F_2\left\{F_1^{-1}(p_1 \mid \mathbf{x}) \mid \mathbf{x}\right\},
\end{aligned}$$

if $F_1^{-1}(p_1 \mid \mathbf{x}) < F_3^{-1}(1 - p_3 \mid \mathbf{x})$.
A popular index of the overall diagnostic accuracy is the volume under the (ROC) surface (VUS). The covariate-specific VUS is given by $\mathrm{VUS}(\mathbf{x}) =$

$\int_0^1 \int_0^1 \text{ROCS}(p_1, p_3 \mid \mathbf{x}) \mathrm{d}p_1 \mathrm{d}p_3$. Conditional on $\mathbf{x}$, when the distributions of test outcomes in the three populations completely overlap, and thus the test has no discriminatory ability, the VUS takes the value 1/6, while a VUS of 1 corresponds to the case of no overlap between any of the three distributions, i.e., the test discriminates perfectly between the three groups. In practice, once the accuracy of the test to distinguish simultaneously between the three classes has been evaluated, the next step is to find the pair of thresholds that should be used in practice to diagnose individuals. A possible criterion to find such a pair of thresholds is to minimise the euclidean distance of the surface to the corner that corresponds to perfect accuracy (where all three TCFs equal one)

$$(c_1^*(\mathbf{x}), c_2^*(\mathbf{x})) = \arg\min_{c_1, c_2 : c_1 < c_2} \sqrt{(\text{TCF}_1(c_1, c_2 \mid \mathbf{x}) - 1)^2 + (\text{TCF}_2(c_1, c_2 \mid \mathbf{x}) - 1)^2 + (\text{TCF}_3(c_1, c_2 \mid \mathbf{x}) - 1)^2}.$$

2.2 Estimation

We assume that the relationship between covariates and test outcomes in each population is given by a location-scale regression model, i.e.,

$$Y_1 = \mu_1(\mathbf{x}) + \sigma_1(\mathbf{x})\varepsilon_1, \quad Y_2 = \mu_2(\mathbf{x}) + \sigma_2(\mathbf{x})\varepsilon_2, \quad Y_3 = \mu_3(\mathbf{x}) + \sigma_3(\mathbf{x})\varepsilon_3,$$

where $\mu_d(\mathbf{x})$ and $\sigma_d^2(\mathbf{x})$ are the conditional mean and variance functions, for $d \in \{1, 2, 3\}$. The errors ε_1, ε_2, and ε_3 are independent of each other and independent of the covariates and are further assumed to have mean zero and unit variance. The corresponding cumulative distribution functions are denoted by F_{ε_1}, F_{ε_2}, and F_{ε_3}. The induced form of the covariate-specific ROC surface can therefore be expressed as

$$\begin{aligned} \text{ROCS}(p_1, p_3 \mid \mathbf{x}) = F_{\varepsilon_2} \left\{ \frac{\mu_3(\mathbf{x}) - \mu_2(\mathbf{x})}{\sigma_2(\mathbf{x})} + \frac{\sigma_3(\mathbf{x})}{\sigma_2(\mathbf{x})} F_{\varepsilon_3}^{-1}(1 - p_3) \right\} - \\ F_{\varepsilon_2} \left\{ \frac{\mu_1(\mathbf{x}) - \mu_2(\mathbf{x})}{\sigma_2(\mathbf{x})} + \frac{\sigma_1(\mathbf{x})}{\sigma_2(\mathbf{x})} F_{\varepsilon_1}^{-1}(p_1) \right\}, \end{aligned} \tag{1}$$

for p_1 and p_3 such that $\text{ROCS}(p_1, p_3 \mid \mathbf{x}) \geq 0$ and 0 otherwise. The expression for the pair $(c_1^*(\mathbf{x}), c_2^*(\mathbf{x}))$ is also rewritten is a similar fashion.
As Equation (1) makes clear, estimating the covariate-specific ROC surface and associated VUS and optimal pair of thresholds, under the assumption of a location-scale regression model in each group, is a matter of estimating the mean and variance functions as well as the distribution function of the regression errors. We propose to estimate, in each group, the mean and variance functions, in a sequential manner, using penalised splines (Eilers and Marx, 1996) and to estimate F_{ε_d} by a smoothed version of the empirical distribution function of the standardised residuals (Pya and Wood, 2015). With regard to inference, a bootstrap of the residuals, in each group, is employed.

3 Application

We apply our methods to data derived from the Alzheimer's Disease Neuroimaging Initiative. Here we study how the hypometabolic convergence index (HCI) simultaneously distinguishes between individuals with normal cognition, with mild cognitive impairment, and with AD and how this discriminatory ability may change with age and gender. Of the 1032 individuals in our study, 138 have been diagnosed with AD, 581 have mild cognitive impairment, and 313 are cognitively normal. In Figure 1 we can see that the discriminatory ability of the HCI to distinguish between the three groups across all ages is quite good and that it is slightly better for ages between 70 and 80 years old (by opposition to ages between 65 ad 70 and between 80 and 85). Gender does not seem to greatly affect the discriminatory ability of the HCI. Further, while for c_1 there does not seem to exist a marked variation with age, c_2 does vary with age, although the bootsrap confidence bands are wider.

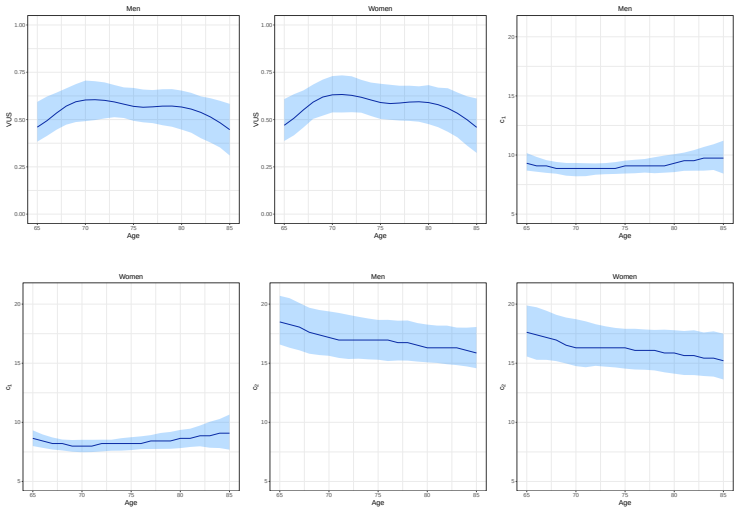


FIGURE 1. VUS , c_1^*, and c_2^*. Solid lines are the point estimates and shaded bands are the 95% pointwise bootstrap confidence bands (based on 500 resamples).

References

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

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

Assessing spatial trends in health outcomes using primary care registry data

Arne Janssens[1], Pieter Libin[2,3,4], Gijs Van Pottelbergh[1], Jonas Crèvecoeur[3], Bert Vaes[1], Thomas Neyens[3,5]

[1] KU Leuven, Academic Centre for General Practice, Leuven, Belgium
[2] Vrije Universiteit Brussel, Artificial Intelligence Lab, Brussels, Belgium
[3] Hasselt University, Interuniversity Institute of Biostatistics and statistical Bioinformatics, Hasselt, Belgium
[4] KU Leuven, Rega Institute for Medical Research, Leuven, Belgium
[5] KU Leuven, Interuniversity Institute of Biostatistics and statistical Bioinformatics, Leuven, Belgium

E-mail for correspondence: `arne.janssens@kuleuven.be`

Abstract: There is a great interest to spatially analyse data from primary care registries, as they can provide timely information about the spatial risk of diseases. Disease mapping methods aim to interpret geographical variation in disease risk in order to identify regions at higher risk for a certain disease. However, it is unclear to which extent characteristics of primary care registry data, such as a spatially disproportionate data sample or variation in reporting efforts, may affect spatial inference. This paper investigates these issues using a spatially discrete geostatistical model on a case from the Flemish (Belgium) Intego primary care registry. By means of a simulation study, we describe several important considerations concerning the registry design as well as the use of spatial models on such data.

Keywords: Spatial analysis; primary care registry; R-INLA.

1 Introduction

Primary care registries hold individual patient information that is routinely collected during daily general practice (e.g., disease diagnoses, demographic information, medications, etc.). Spatial analyses of such data, e.g., by means of disease mapping methods (Lawson et al. (2000)), can provide timely information about the geographical variation in disease risks. However, it is unclear to which extent often encountered characteristics of

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

primary care registry data, such as spatially disproportionate data collection or variation in reporting efforts, may invalidate inference about the spatial mechanisms. We evaluate these issues in the context of the Intego registry (Truyers et al. (2014)) by means of a Bayesian spatial analysis of a health outcome and a subsequent simulation study, after which we give recommendations to set up similar studies.

2 Data

We investigated the spatial distribution of lower respiratory tract infections (LRTIs) in Flanders, i.e., the Dutch-speaking northern region of Belgium, in the year 2019 using data from the Intego registry. The study population consisted of 260.017 individuals who visited, at least once in 2019, one of the 105 general practices that were part of the Intego network in 2019. The outcome of interest concerns an LRTI diagnosis which was defined as the diagnosis of either acute bronchitis or pneumonia. The Intego database provided information about an individual's age, sex, increased reimbursement for healthcare status, municipality of residence, and visited practice. These data represent a case of unbalanced sampling as there were strong regional differences in the sample's representation of the population at risk. For the simulation, we constructed an additional data set that mimics a balanced sampling scheme, i.e., where all municipalities provide an equal number of participating practices and where all inhabitants of a municipality visit one of the practices in that municipality.

3 Model

We modelled the binary LRTI state of a patient, Y_i, with $i = 1, ..., N$, living in municipality x_i, and visiting practice z_i, via a generalized linear mixed model,

$$Y_i \sim Bernoulli(\pi_i), \tag{1}$$

$$logit(\pi_i) = \mu_i + U(x_i) + V(z_i), \tag{2}$$

with μ_i containing the intercept and covariate effects,

$$\begin{aligned} \mu_i = \beta_0 &+ \beta_1 * \text{sex}_i + \\ &+ \beta_2 * \text{agegroup}_{1i} + \beta_3 * \text{agegroup}_{2i} + \\ &+ \beta_4 * \text{agegroup}_{4i} + \beta_5 * \text{agegroup}_{5i} + \beta_6 * \text{agegroup}_{6i} + \\ &+ \beta_7 * \text{reimb}_i, \end{aligned} \tag{3}$$

where $agegroup_3$ was used as the reference age group. We apply a random-effects parametrization similar to the BYM convolution (Besag, York, and

Mollié (1991)), where $U(x)$ is a spatially structured conditional autoregressive (CAR) random effect at the level of the residential location (Besag (1974)),

$$U(x)|U(x^-) \sim \mathcal{N}(\bar{u}_{\delta_x}, \frac{\sigma_u^2}{n_{\delta_x}}), \tag{4}$$

with $\bar{u}_{\delta_x} = \frac{\sum_{l \in \delta_x} U(l)}{n_{\delta_x}}$, where x^- represents the set of all municipalities except location x; δ_x and n_{δ_x} represent, respectively, the set and the number of neighbours of municipality x, using a first-order neighborhood structure. V(z) is a spatially unstructured Gaussian random effect at the practice level,

$$V(z) \sim \mathcal{N}(0, \sigma_v^2). \tag{5}$$

Integrated Nested Laplace Approximation (INLA) was used for model estimation in R (R-INLA, Rue (2009)).

4 Simulation configuration

As a sensitivity analysis of the model, applied to the case from Intego, indicated that the estimation of the spatial process ($U(x)$) was sensitive to the set of practices providing the data, a simulation study was performed. The simulations considered two general scenarios, one with an unbalanced sampling scheme and one with a balanced sampling scheme. Table 1 shows the values and explanation of the three parameters that were adjusted in different simulation settings. In total, there were 36 parameter combinations (scenarios), 9 for the unbalanced (without c) and 27 for the balanced sampling scheme. Per scenario, we simulated values for the random effects and together with the covariate estimates obtained from the case, the probability π_i was computed and the number of LRTI cases were simulated from a binomial distribution. This process was repeated 150 times per scenario. The resulting data sets were analysed with the same model. The estimation of the mechanism of interest, i.e., the spatial effect was assessed.

TABLE 1. Simulation parameters. Parameter c only applied to the balanced sampling scheme scenario.

parameter	interpretation	values
σ_u	variation in the random spatial effect	0.1, 0.3, 0.6
σ_v	variation in the random practice effect	0.1, 0.3, 0.6
c	number of practices per municipality	1, 2, 3

5 Results

Figure 1 shows the boxplots of the mean absolute errors (MAEs) between the estimated and simulated spatial pattern after standardizing them to a process with standard deviation $\sigma_u = 0.1$. The MAE for one simulation run (for one scenario) s was computed as $MAE_s = \frac{\sum_{j=1}^{m} |\hat{U}(x_j)_s - U(x_j)_s|}{m}$ with $m = 300$, i.e., the number of municipalities in Flanders. The MAEs of the spatial pattern are closer to zero for higher σ_u, lower σ_v and higher number of practices c. Interestingly, lowering the practice variability leads to a substantially better estimation of the spatial trend in the balanced sampling scheme scenario.

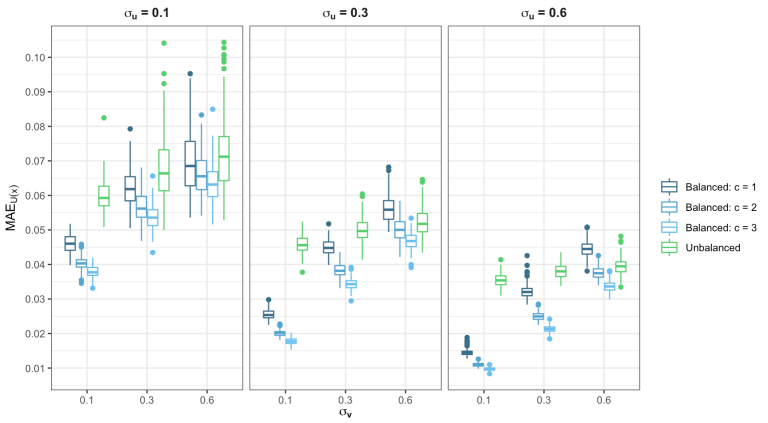


FIGURE 1. Mean absolute errors of the spatial patterns in different scenarios after standardizing.

References

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B (Methodological)*, **36**, 192 – 236

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

Lawson, A.B., Biggeri, A.B., Böhning, D., Lesaffre, E., Viel, JF, Clark, A., Schlattmann, P., and Divino, F. (2000). Disease mapping models: an empirical evaluation. *Statistics in medicine*, **10**, 2217 – 2241

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace

Approximations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **71**, 319 – 392

Truyers, C. et al. (2014). The Intego Database: Background, Methods and Basic Results of a Flemish General Practice-Based Continuous Morbidity Registration Project. *BMC Medical Informatics and Decision Making*, **14**, 48

Statistical inference for high-dimensional logistic regression: Variable selection and levels fusion for 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: High-dimensional problems occur in a huge variety of applications, increasing the need of suitable models that are interpretable. Especially in the presence of (dummy-coded) categorical covariates, the dimension of the parameter vector to be estimated grows rapidly. Assuming that the true underlying structure is sparse, variable selection procedures eliminate the non-influential covariates from a regression-type model. We target at further reducing the dimension through categorical covariates levels fusion. That is, if a subset of levels of a categorical covariate have the same influence on the response, they will be fused. After introducing L_0-fused group lasso (L_0-FGL) for logistic regression, performing variable selection and levels fusion simultaneously, we ensure the quality of our new approach by investigating its theoretical properties. Developing and implementing two computational approaches for L_0-FGL, we further point out the performance of the resulting estimates in simulation studies. To obtain post selection inference for L_0-FGL, we obtain a (multiple) sample splitting approach including a likelihood ratio test framework. For the resulting two-stage L_0-FGL, we investigate theoretical properties and examine the asymptotic behavior.

Keywords: High-Dimensional Statistics, Group Lasso, L_0 Norm, Likelihood Ratio Test, Sample Splitting

1 The L_0-Fused Group Lasso for Logistic Regression

Under a logistic regression setting, we assume the response variable Y to be binary. Further, we assume to have $J \in \mathbb{N}$ categorical covariates $\boldsymbol{X}_j$, $j \in \{1, ..., J\}$, each having $p_j + 1$ levels, coded by 0, ..., p_j. In particular, zero is chosen to be the reference category. With $p := \sum_{j=1}^{J} p_j$, the resulting parameter vector is $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J)^T \in \mathbb{R}^{p+1}$, where β_0 denotes the

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

intercept. The subvector $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_j})$, $j \in \{1, ..., J\}$ is the parameter subvector corresponding to the j-th factor $\boldsymbol{X}_j$. For a sample size $n \in \mathbb{N}$, we assume a fixed $n \times (p+1)$ design matrix $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X}_1^T, ..., \boldsymbol{X}_J^T)$. Performing regularized or penalized regression, our goal is to minimize the sum of the log-likelihood function (denoted by $L_n(\boldsymbol{\beta})$) and a penalty function (denoted by $P_\lambda(\boldsymbol{\beta})$), hence

$$M_{pen}(\boldsymbol{\beta}) = -L_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta})$$

is minimized, where we call $M_{pen}(\boldsymbol{\beta})$ the objective function. Consequently, the resulting penalized estimator is given by

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} M_{pen}(\boldsymbol{\beta}). \tag{1}$$

The well known group lasso (see Kim et al., 2006; Meier et al., 2008) is designed for factor selection, meaning that a factor is either completely excluded from the model, or it is entirely included in the model. Besides factor selection, performing levels fusion will further reduce the dimension of the problem. To do so, penalty functions are applied on the differences of the covariates levels within a factor, considering all pairwise differences or only those of adjacent levels, depending on whether the factor is nominal or ordinal, respectively. Applying the L_1 penalty on the differences of the levels, such fusions were considered by Bondell and Reich (2009) for ANOVA and by Gertheiss and Tutz (2010) for linear regression. To overcome the known issue of biased estimates using L_1, the L_0 norm $||\cdot||_0$ applied on the differences was considered by Oelker et. al (2014), where for some $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ the L_0 norm is defined as $||\boldsymbol{\gamma}||_0 := |\{j \,|\, \gamma_j \neq 0\}|$, counting the number of nonzero entries in $\boldsymbol{\gamma}$.
For the purpose of simultaneously performing factor selection and levels fusion, we introduce the L_0-FGL penalty function

$$P_\lambda(\boldsymbol{\beta}) := \lambda_1 \sum_{j=1}^{J} ||\boldsymbol{\beta}_j||_{K_j} + \lambda_0 \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0,$$

where $\boldsymbol{K}_j \in \mathbb{R}^{p_j \times p_j}$ is some positive definite and symmetric matrix (see Kaufmann and Kateri, 2022). The first summand of the penalty function above

$$P_\lambda^1(\boldsymbol{\beta}) := \lambda_1 \sum_{j=1}^{J} ||\boldsymbol{\beta}_j||_{K_j}$$

is the Group Lasso penalty, while the second one

$$P_\lambda^0(\boldsymbol{\beta}) := \lambda_0 \sum_{j=1}^{J} \sum_{0 \leq r < s \leq p_j} w_0^{(j,rs)} ||\beta_{j,r} - \beta_{j,s}||_0$$

is the CAS-L_0 penalty function. By appropriately choosing $\boldsymbol{K_j}$ in $P_\lambda^1(\boldsymbol{\beta})$, for example $\boldsymbol{K}_j := \tilde{w}_1^{(j)} \mathbf{1}_{p_j \times p_j}$, we can also obtain weights $w_1^{(j)} = \sqrt{\tilde{w}_1^{(j)}}$ in the Group Lasso part. Throughout our work we will adopt (in the non-adaptive version) the convenient choice $\boldsymbol{K_j} := p_j \mathbf{1}_{p_j \times p_j}$ which gives weights that account for factors having different numbers of levels. Using adaptive weights, we further introduce the adaptive L_0-FGL penalty function.

1.1 Existence and Theoretical Properties of the L_0-FGL Estimator

We show the existence of a minimum of the L_0-FGL objective function $M_{pen}(\boldsymbol{\beta})$, as well as that this choice of penalty performs levels fusion. We can also ensure this existence in a high dimensional setting with $p > n$. Having proven the existence of the L_0-FGL estimator (1), we prove its $\sqrt{n}$ consistency for the case of fixed $p < n$ and for diverging p_n, under respective regularity conditions. Further, under a true underlying sparse structure, we derive the asymptotic normality property for the L_0-FGL estimator (1). Two results (one for the fixed p case and one for the diverging p_n case) on selection consistency complete the investigation of theoretical properties. More details can be found in Kaufmann and Kateri (2022).

1.2 Computational Methods and Simulation Studies

To obtain L_0-FGL estimates, we regard two different approaches: the PIRLS algorithm (see Oelker and Tutz, 2013) and a block coordinate descent (BCD) procedure (see Meier et al., 2008). Since the L_0 part in L_0-FGL is not continuous, hence not differentiable, we use a quadratic approximation making it feasible for optimization. In PIRLS, also the Group Lasso part is quadratically approximated while in the BCD procedure it is not directly approximated since we use a quasi Newton step. Applying both computational methods, we investigate the performance of L_0-FGL in simulation studies. For the chosen high-dimensional design, our BCD approach seems to be beneficial, highly reducing the complexity of the problem. In a non high dimensional design we observe that L_0-FGL computed with PIRLS results in the most sparse model improving the selection performance of the known L_0 approach (see Kaufmann and Kateri, 2022).

2 Statistical Inference for L_0-Fused Group Lasso

It is well known that inference in high dimensions is challenging. Using penalized regression with L_0-FGL, we are able to reduce the dimensionality of the parameter space that we observe. But, the high dimensionality makes it in general impossible to directly apply likelihood ratio tests because of the missing consistency of the maximum likelihood estimator (see Ning and Liu,

2017). Consequently, it is not straightforward to obtain inferential results for L_0-FGL. We propose a two-stage procedure, based on a (single and multiple) sample splitting approach which allows us to perform statistical inference (see Wassermann and Roeder, 2009; Meinshausen et al, 2008).

2.1 Two-Stage L_0-FGL

The general idea is to first split the sample in two independent sets D_1 and D_2. Then, perform L_0-FGL to reduce the dimensionality on set D_1 and, proceeding with the reduced parameter space, perform maximum likelihood estimation on the other set D_2. Since with this procedure we reduced the dimension of the parameter set, we shifted the problem from a high dimensional to a non high dimensional one, making thus possible the application of likelihood ratio test theory. Of course, we have to impose some adequate assumptions to ensure that the dimension reduction is 'enough'. But, with the previous theoretical investigation of L_0-FGL, these assumptions are reasonable. The idea of sample splitting has its roots in Wassermann and Roeder (2009), while it was adjusted for the multiple split in Meinshausen et al. (2008). In both approaches, a linear regression setting is assumed and approaches performing variable selection. We extend this to logistic regression and using a regularization approach which performs variable selection as well as levels fusion. Further, since we adjust the mentioned approaches for categorical variables, we need to test sub-vectors in the likelihood ratio test instead of single (scalar) components, which is the case in the related literature. We use two approaches adjusting for multiplicity of testing (Bonferroni and Benjamini-Hochberg). We first show that in the single split case the type-I-error can be bounded asymptotically, using Bonferroni correction. Further results are investigated, also using Benjamini-Hochberg adjustment (see Benjamini and Hochberg, 1995) and considering the extension to the multiple split. We underline the importance of the theoretical properties for L_0-FGL that we have shown (see Section 1.1) since they will be essential for the proofs for statistical inference.

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.

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

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

Kim, Y. and Kim, J. and Kim, Y. (2006). Blockwise Sparse Regression. *Statistica Sinica*, 16: 375 – 390.

Kaufmann, L. and Kateri, M. (2022). Simultaneous Factors Selection and Fusion of Their Levels in Penalized Logistic Regression. *arXiv*.

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

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-Values for High-Dimensional Regression. *Journal of the American Statistical Association*, 104: 1671 – 1681.

Ning, Y. and Liu, H. (2017). A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models *The Annals of Statistics*, 45: 158 – 195.

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.

Oelker, M.-R., and Tutz, G. (2013). A General Family of Penalties for Combining Differing Types of Penalties in Generalized Structured Models. *Technical Report Number 139, LMU München, Germany*

Wasserman, L. and Roeder, K. (2009). High-Dimensional Variable Selection. *The Annals of Statistics*, 37: 2178 – 2201.

Advanced statistical modelling for polygenic risk scores based on large cohort data

Hannah Klinkhammer[1,2], Christian Staerk[2], Carlo Maj[2,3], Peter M. Krawitz[2], Andreas Mayr[1]

[1] Institute for Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Germany
[2] Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Germany
[3] Center for Human Genetics, Philipps University Marburg, Germany

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

Abstract: Polygenic risk scores (PRS) predict the individual genetic liability of a person to a certain trait and are expected to play an increasingly important role in future clinical risk stratification. Typically, PRS are constructed based on summing up univariate effect estimates derived from genome-wide association studies for risk alleles that are present for the individual. To overcome relying on univariate effect estimates and to directly enable multivariable statistical modelling for large and high-dimensional genotype data, we introduced a statistical boosting framework incorporating various loss functions to model a wide range of phenotypes. We discuss how `snpboost` can be used to construct prediction intervals via quantile regression and how sparse PRS models can be derived for the prediction of time-to-event data.

Keywords: statistical boosting; high-dimensional data; variable selection; polygenic risk scores; prediction.

1 Introduction

Polygenic risk scores (PRS) aim to capture an individual's genetic predisposition to a certain clinical outcome. They are often based on large-scale genotype data with both large n (hundred thousands of participants) *and* large p (millions of SNPs), comprising common genetic variants with low to medium effect sizes. As an additional hurdle, genetic variants which are close to each other often show high correlations (linkage disequilibrium, LD). To deal with the high-dimensionality of the data, the most common

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

methods to derive PRS such as LDpred, PRScs and PRSice use summary statistics of simple univariate association tests which had been assessed for each variant in genome-wide association studies (GWAS). Additional methods such as Bayesian approaches are then applied to account for LD and to combine these univariate effects. However, from a statistical perspective, it would be desirable to directly use multiple regression modelling to jointly estimate multivariably adjusted effects for the variants. To do so, we built the statistical boosting framework `snpboost` (Klinkhammer et al., 2023), which works on batches of variants to overcome the computational hurdle of large p and yields sparse prediction models while automatically selecting the most informative variants. The modular and flexible nature of statistical boosting enables us to extend our framework based on further loss functions. By incorporating loss functions for Gaussian, binary, time-to-event and count data, we can effectively use statistical modelling to optimize the PRS with respect to the clinical outcome.

2 Methods

For $n \in \mathbb{N}$ individuals, let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be the observed phenotype. Let $\boldsymbol{X} \in [0,2]^{n \times p}$ be the observed genotype matrix, where 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. For imputed genotype data, $x_{i,j}$ can fall into the continuous range $[0, 2]$. To apply statistical boosting, we model each variant via a single linear base-learner (component-wise boosting). We specify the loss function $\rho(\boldsymbol{y}, \hat{\boldsymbol{\eta}})$ depending on the type of phenotype; i.e. for quantitative outcomes we use the classical L_2 loss or the check-function for quantile regression, for binary outcomes the logarithmic loss, for count data the negative log likelihood of the Poisson distribution and for time-to-event data the negative log likelihood of an accelerated failure time (AFT) model (Schmid and Hothorn, 2008) with Weibull distributed survival times T_i given by

$$\log(T_i) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \sigma W, \quad W \sim \text{Gumbel}(0, 1) \tag{1}$$

with scaling parameter $\sigma > 0$ and standard Gumbel distributed W.
Aiming at minimizing the loss ρ, we estimate $\hat{\boldsymbol{\eta}} = \sum_{j=1}^{p} \hat{\beta}_j \boldsymbol{x}_j$ via statistical boosting by iteratively fitting the base-learners to the negative gradient vector of the loss as described in Bühlmann and Hothorn (2007). In order to deal with large-scale genotype data, we incorporated an additional batch-building step to restrict the set of base-learners in each boosting iteration (Figure 1). In the outer loop (shown in blue in Figure 1) we build a batch of variants by selecting the p_{batch} best fitting variants according to their correlation with the current negative gradient vector of the loss function.

Then, statistical boosting is applied on the chosen batch until the correlations with the current negative gradient of the variants in the batch are smaller than the latest updated maximal correlation of variants outside the batch (shown in grey in Figure 1). Finally, we simultaneously monitor the predictive performance on an independent validation set and the algorithm is stopped when the predictive performance on the independent validation set is no longer improving.

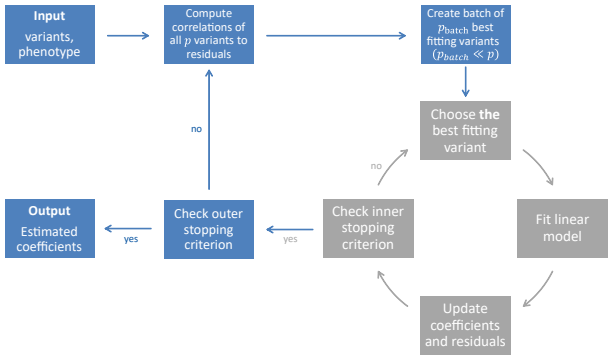


FIGURE 1. Illustration of the `snpboost` batch-wise boosting algorithm to make statistical boosting work on large individual-level genotype data. The grey boxes reflect the steps inside the batches.

3 Results

In Klinkhammer et al. (2023) we have shown that `snpboost` can effectively model the polygenic predisposition for continuous and binary outcomes. Here, we focus on extending the snpboost framework via new loss functions to two further types of outcomes.

First, we considered modelling time-to-event data, namely we wanted to model a PRS for the age of onset of asthma. From the UK Biobank resource we extracted $n = 395,644$ unrelated British individuals and used self-reports, ICD-9 codes and ICD-10 codes to identify $n = 56,436$ individuals affected by asthma. Additionally, we assessed the age of onset as the age when asthma was first diagnosed. For individuals not affected by asthma, the age at last follow-up visit or, if applicable, age at death was considered as censoring time. For all individuals, genotype data of $p = 604,967$ variants was available. The data set was randomly split into training (50%), validation (25%) and test set (25%). We applied an AFT model with a Weibull distribution via `snpboost` on the training and validation data, which yielded a PRS comprising $p = 1,622$ selected genetic variants with a non-zero effect size. The quantiles of the distribution of the estimated PRS on the training and validation data were used to classify

individuals into three classes: low PRS (PRS $<$ 10% PRS quantile of training and validation set, "bottom 10%"), medium PRS (PRS between 10% and 90% PRS quantile of training and validation set, "10-90%"), high PRS (PRS $>$ 90% PRS quantile of training and validation set, "top 10%"). Figure 2 shows the predicted lifetime risk for asthma on the test set stratified over the fitted PRS. Furthermore, we included the PRS as well as sex, age and the first 10 principal components of the genotype matrix in an AFT model with Weibull distributed outcome on the training and validation set. The corresponding prediction on the test set yielded a C-index of $C = 0.64$ (95% CI: 0.63-0.65).

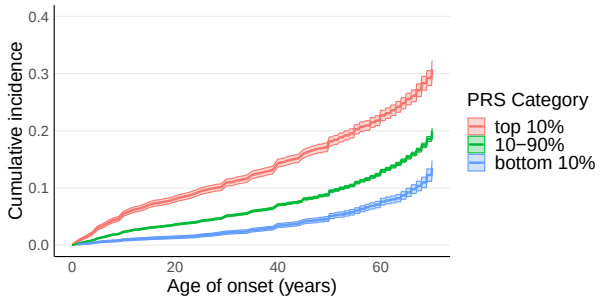


FIGURE 2. Cumulative incidence of asthma based on Kaplan-Meier estimates on test data stratified over three PRS groups, reflecting their genetic predisposition (stratification based on PRS distribution on training and validation set).

Secondly, we applied quantile regression via `snpboost` on a set of $n = 284,342$ individuals from the UK Biobank (divided into training, validation and test set) to construct predictive intervals for BMI. For 78% of the individuals in the test set ($n = 56,984$), the observed BMI was covered by intervals between the predicted 10% and 90% quantiles that were based on genetic information solely. Figure 3 shows the observed BMI data as well as the predicted quantiles ordered by predicted mean (50% quantile).

4 Discussion

Our extended boosting framework `snpboost` enables multivariable statistical modelling for polygenic risk scores on large and high-dimensional genotype data, providing a variety of loss functions for different traits. Quantile regression can be used to construct individual prediction intervals and additionally comprises median regression as a robust alternative to mean regression, which can be more suitable for outcomes that are susceptible to outliers. Regarding time-to-event outcomes, we were able to construct sparse PRS that can help to stratify the course of a disease. Noteworthy, using a PRS that was boosted to predict the *occurrence* of asthma (binary

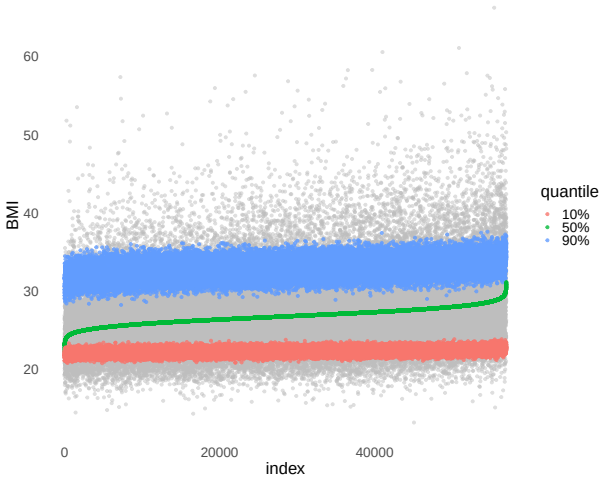


FIGURE 3. Observed BMI values as well as predicted 10%, 50% and 90% quantiles for a test set comprising $n = 56,984$ individuals.

outcome), neglecting the event time as it is common in the field of PRS modelling, showed very similar discriminatory power with respect to the age of onset. Further research is warranted to investigate the potential of advanced statistical modelling in the prediction of disease courses based on genetic information.

Acknowledgments: This study was conducted on data from the UK Biobank resource under application number 81202.

References

Bühlmann, P. and Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, **22**, 477–505.

Klinkhammer, H., Staerk, C., Maj, C., Krawitz, P., Mayr, A. (2023). A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Frontiers in Genetics*, **13**.

Schmid, M. and Hothorn, T. (2008). Flexible boosting of accelerated failure time models. *BMC Bioinformatics*, **9**.

Sparse modality regression

Chris Kolb[1,2], Bernd Bischl[1,2], Christian L. Müller[1,2] and David Rügamer[1,2,3]

[1] LMU Munich, Germany
[2] Munich Center for Machine Learning, Germany
[3] TU Dortmund, Germany

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

Abstract: Deep neural networks (DNNs) enable learning from various data modalities, such as images or text. This concept has also found its way into statistical modelling through the use of semi-structured regression, a model additively combining structured predictors with unstructured effects from arbitrary data modalities learned through a DNN. This paper introduces a new framework called sparse modality regression (SMR). SMR is a regression model combining different data modalities and uses a group lasso-type regularization approach to perform modality selection by zeroing out potentially uninformative modalities.

Keywords: Neural Networks; Statistical Modelling; Deep Learning; Modality Selection.

1 Introduction

Neural networks have become a highly significant area of study and development in recent years due to their ability to solve complex problems in a variety of fields, including computer vision, speech recognition, and natural language processing. This is largely owed to breakthroughs in deep learning techniques, which allow for efficient and large-scale training of neural network models that can accurately capture the underlying patterns and relationships in large amounts of data.
One of their key features is the flexibility in working with various data types and structures. Rügamer et al. (2023) proposed a framework for combining structured regression models and DNNs in a unifying network architecture (cf. Fig. 1). With their so-called *semi-structured regression (SSR)* model, the authors aim to extend the scope of statistical regression to incorporate non-tabular data modalities, but also to integrate interpretable additive

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

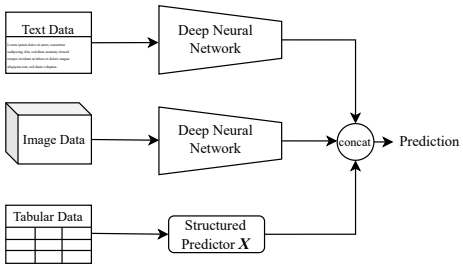


FIGURE 1. Architecture of a semi-structured regression model with one tabular input for structured predictors and two unstructured data sources (image, text), each transformed through a deep neural network.

predictors into DNN architectures. This paper presents a sparse extension of SSR models that enables principled modality selection and thereby allows to decide in a data-driven fashion which information sources should be retained in the model and which not.

2 Sparse Modality Regression

We define our setup using a linear mean regression, but note that this approach naturally extends to 1) additive models by replacing the linear predictor with smooth functions learned through basis representations, and 2) distributional regression with commonly used parametric distributions by learning each distribution parameter from multiple data modalities.

Let the different data modalities be defined by $\boldsymbol{x} \in \mathbb{R}^p$ for the tabular data and $\boldsymbol{z}_m, m = 1, \ldots, M$ for the M unstructured data sources with arbitrary shapes. We use these modalities to model the outcome of interest Y as

$$Y = \boldsymbol{x}^\top \boldsymbol{\beta} + \textstyle\sum_{m=1}^{M} \mathcal{D}_m(\boldsymbol{z}_m)^\top \boldsymbol{\gamma}_m + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the linear regression coefficients, $\mathcal{D}_m$ the mth (deep) neural network processing the mth modality to a q_m-dimensional latent representation, and $\boldsymbol{\gamma}_m \in \mathbb{R}^{q_m}$ the coefficients for the latent features in the final additive combination. Figure 1 depicts an exemplary SSR architecture for $M = 2$, with a text description $\boldsymbol{z}_1$ and an image input $\boldsymbol{z}_2$.

Sparsity in Semi-Structured Regression The model (1) can be embedded into one unifying neural network and estimated using first-order optimization routines, i.e., variants of stochastic gradient descent (SGD) as popularized in deep learning (see Rügamer et al., 2023). While considered a general-purpose tool, SGD does not allow for the optimization of

non-smooth objectives such as the one suggested in this paper:

$$\sum_{i=1}^{n} ||y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} - \textstyle\sum_{m=1}^{M} \mathcal{D}_m(\boldsymbol{z}_{m,i})^\top \boldsymbol{\gamma}_m||_2^2 + \lambda \left\{ ||\boldsymbol{\beta}||_2 + \textstyle\sum_{m=1}^{M} ||\boldsymbol{\gamma}_m||_2 \right\}, \quad (2)$$

where the first summand corresponds to the usual L_2 loss and the second summand penalizes the $M+1$ groups of different data modalities using an $L_{2,1}$ group lasso penalty with regularization parameter $\lambda \geq 0$.
It is, however, possible to transfer the above non-smooth optimization problem to a smooth surrogate using the general framework proposed by Kolb et al. (2023). The optimization transfer involves two steps: first, the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_m, m = 1, \ldots, M$, in the predictor are overparametrized using a type of Hadamard product, i.e., $\boldsymbol{\beta} = \boldsymbol{u}_0 \odot \boldsymbol{v}_0$ and $\boldsymbol{\gamma}_m = \boldsymbol{u}_m \odot \boldsymbol{v}_m$. The parameters $\boldsymbol{u}_m$ have the dimensionality of the respective original coefficient vector, whereas $\boldsymbol{v}_m = v_m \mathbb{1}$ induces one additional scalar parameter per modality. In the second step, the non-smooth penalty is replaced by a smooth quadratic penalty on the reparametrized coefficients. The quadratic penalty term in the overparametrized problem is defined as

$$\frac{\lambda}{2} \left\{ \textstyle\sum_{m=0}^{M} ||\boldsymbol{u}_m||_2^2 + v_m^2 \right\}, \quad (3)$$

and it can be shown that the minimum of this penalty, subject to the constraints imposed by the reparametrization, is precisely equal to the non-smooth group lasso penalty:

$$\min_{\substack{\boldsymbol{u}_0, \boldsymbol{v}_0 : \boldsymbol{\beta} = \boldsymbol{u}_0 \odot \boldsymbol{v}_0 \\ \boldsymbol{u}_m, \boldsymbol{v}_m : \boldsymbol{\gamma}_m = \boldsymbol{u}_m \odot \boldsymbol{v}_m}} \frac{1}{2} \left\{ \textstyle\sum_{m=0}^{M} ||\boldsymbol{u}_m||_2^2 + v_m^2 \right\} = \left\{ ||\boldsymbol{\beta}||_2 + \textstyle\sum_{m=1}^{M} ||\boldsymbol{\gamma}_m||_2 \right\}. \quad (4)$$

Optimization over the factorized coefficients $\boldsymbol{u}_m, \boldsymbol{v}_m$ using the smooth penalty (3) is then equivalent to solving the $L_{2,1}$ regularized group sparse problem (2) (Kolb et al., 2023).

3 Simulation

In order to examine the proposed Hadamard parametrization framework for SMR, we simulate $n = 2000$ observations according to a data-generating process involving two modalities (tabular and image). The tabular predictor is simulated based on uniformly drawn covariates $\boldsymbol{x}$, whereas the image predictor $\mathcal{D}(\boldsymbol{z})$ is derived from the face value of images of handwritten digits ranging from 0 to 9 (Deng, 2012). The noise is sampled from a standard Gaussian. To evaluate the modality support recovery of SMR, we combine the (initially balanced) structured and unstructured predictors as a convex combination using an influence parameter $\rho \in [0, 1]$, i.e., $Y = (1 - \rho)\boldsymbol{x}^\top \boldsymbol{\beta} + \rho \mathcal{D}(\boldsymbol{z})^\top \boldsymbol{\gamma} + \varepsilon$. Our models are implemented in the R

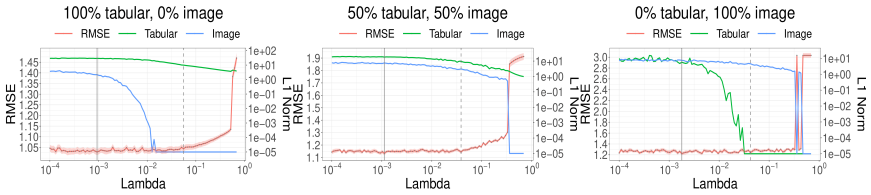


FIGURE 2. Out-of-sample RMSE and L_1 norm of coefficients per modality for $\rho \in \{0, 0.5, 1\}$ (left to right). The lines indicate the optimal λ^{min} (solid) and the λ^{1se} chosen by the one-standard-error-rule (dashed).

software package `deepregression` (Rügamer et al., 2022), and optimized using SGD with a batch size of 32, learning rate of 5e−3, and momentum parameter 0.9.

The corresponding results are shown in Figure 2, depicting the regularization paths and test errors of our approach for different ρ values. Results show that our SMR approach is able to correctly identify non-informative modalities (if any) without sacrificing predictive power, and thus provides a valid modality-sparse regression model for a suitable choice of the regularization parameter λ.

4 Application: Petfinder

We apply our approach to multi-modal data from `Petfinder.my`, a website that allows people to search for adoptable pets. The modalities comprise images from pet listings, a text description of the pets, and tabular data of various attributes such as breed, age, or health condition. We define a multi-modal SSR processing the images as well as the text, fusing both their latent representations with the additive predictor defined for the tabular data. The task is to predict the probability of new pets getting adopted within 100 days, i.e., a logistic regression version of the model presented in (1). We apply modality regularization as it is *a priori* unclear which of the modalities are informative. The resulting regularization path can be seen in Figure 3 (right), suggesting the least informative modality for the adoption probability is in fact the pets' images. This is in line with findings from the *PetFinder.my Adoption Prediction* challenge held in 2019.
In Figure 3 (left), we show the estimated non-linear effect of age for the sparse model with only tabular and text data, indicating that adoption probability is increasingly negatively influenced by the animal's age for elderly pets.

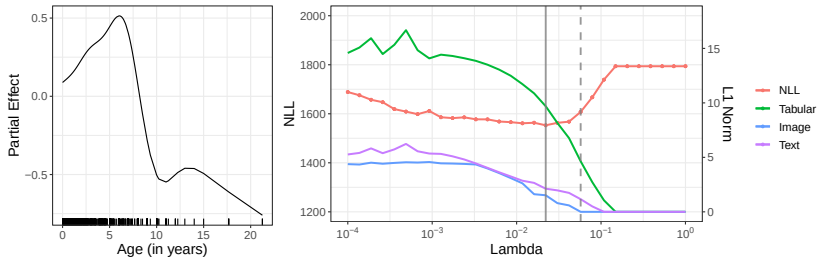


FIGURE 3. Left: Estimated non-linear effect of age. Right: Regularization path with out-of-sample neg. log-likelihood (left axis) and L_1 norm of coefficients per modality (right axis). The lines indicate the best model (solid) and a similarly performing, but sparse model with only structured and text modality (dashed).

5 Summary

In this work, we presented a novel framework that combines ideas of group sparse regularization and semi-structured networks to effectively enable the selection of informative data modalities in multi-modal applications. In order to learn a sparse representation using off-the-shelf SGD optimization, we leverage recent findings on regularization in overparametrized regression models. We demonstrate the effectiveness of our proposed framework in a simulation study and further apply it to real-world data to showcase its practical use.

References

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6), 141–142.

Kolb, C., Bischl, B., Müller, C. L., and Rügamer, D. (2023). Smoothing the Edges: A General Framework for Smooth Optimization in Sparse Regularization using Hadamard Overparametrization. *Unpublished work under review.*

Rügamer, D., Kolb, C., and Klein, N. (2023). Semi-Structured Distributional Regression. *The American Statistician*, **0**, 1–25. Taylor & Francis.

Rügamer, D., Kolb, C., Fritz, C., et al. (2022). deepregression: a Flexible Neural Network Framework for Semi-structured Deep Distributional Regression. *Journal of Statistical Software*, 105(1), 1–31.

On prediction via equal-tailed intervals with an application to sensor data analytics

Michele Lambardi di San Miniato[1], Ruggero Bellio[1], Luca Grassetti[1], Paolo Vidoni[1]

[1] Department of Economics and Statistics, University of Udine, Italy

E-mail for correspondence: michele.lambardi@uniud.it

Abstract: Prediction intervals with equal tail probabilities can be obtained by minimizing the Winkler loss function via the Hogg estimator. The prediction obtained this way can provide a more complete picture alongside alternative intervals centered around an available prediction rule. We illustrate the method with a dataset from an indoor environmental monitoring application.

Keywords: Hogg estimator; Interval prediction; Quantile regression; Sensor data.

1 Introduction

When predicting a random variable Y, uncertainty can be represented by means of an interval that predicts Y correctly with coverage probability $1-\alpha$. Here, the focus is on prediction intervals based on pairs of quantiles. Let y_p be the p-quantile of Y, such that $\Pr(Y \leq y_p) = p$. A prediction interval with equal tail probabilities (ETP) can be defined as

$$\mathcal{I}_{1-\alpha}^{\text{ETP}} = [y_{\alpha/2}, y_{1-\alpha/2}]\,,$$

which has some appealing invariance properties (see, e.g., Brehmer and Gneiting, 2021). However, a point prediction μ may be available, determined independently of the coverage target. In this case, one may favor the alternative interval

$$\mathcal{I}_{1-\alpha}^{\text{UTP}} = \mu \pm \rho_{1-\alpha}\,,$$

with unequal tail probabilities (UTP), where ρ_p is the p-quantile of the absolute prediction error $|Y-\mu|$. He (1997) uses the median prediction $\mu = y_{1/2}$ for the sake of robustness.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

We argue that ETP intervals can be more meaningful than UTP ones in some cases. ETP prediction error is less coarse, especially for high coverage and asymmetric error distributions. ETP prediction is also less biased towards groups defined by covariates, in the sense that the prediction error has sign uncorrelated with the explanatory variables.
We make ETP intervals more interpretable, as much as UTP ones, by reparameterizing them as

$$\mathcal{I}_{1-\alpha}^{\mathrm{ETP}} = \mu_\alpha \pm \delta_\alpha \,,$$

where $\mu_\alpha = (y_{1-\alpha/2}+y_{\alpha/2})/2$ and $\delta_\alpha = (y_{1-\alpha/2}-y_{\alpha/2})/2$. Next, we provide an efficient and approximate estimator of ETP intervals, as an alternative to the existing implementation of the Hogg estimator in the `R` package `quantreg` (Koenker, 2022). Moreover, we want to stress the usefulness of ETP prediction targeted at coverage $1-\alpha$ when this can be chosen soundly. Finally, we highlight the non-systematic nature of ETP prediction error, with an application to sensor data analytics.

2 Methodology

Consider $1-\alpha$ prediction of $Y = x^{\mathrm{T}}\beta_\alpha + \epsilon_\alpha$, where $x \in \mathbb{R}^d$ is a covariate vector, $\beta_\alpha \in \mathbb{R}^d$ is the regression coefficient, and ϵ_α is a noise term whose quantiles of order $\alpha/2$ and $1-\alpha/2$ are equally distant from zero. Then, ETP intervals will hold with $\mu_\alpha = x^{\mathrm{T}}\beta_\alpha$. Relatedly to this model, borrowing from support vector regression, the Δ-insensitive loss function (Vapnik, 1998) is defined depending on $\Delta \geq 0$ as

$$S(e;\Delta) = \max(0, |e| - \Delta) \,.$$

The Winkler loss function (see, e.g., Brehmer and Gneiting, 2021) is defined accordingly, also depending on the hyperparameter α, as

$$W_\alpha(e;\Delta) = \alpha\Delta + S(e;\Delta) \,.$$

The ground truth is given by the true parameter values $(\beta_\alpha, \delta_\alpha)$, which solve the risk minimization problem

$$\min_{(\beta,\delta)} \mathbb{E}_{(Y,x)}[W_\alpha(Y - x^{\mathrm{T}}\beta;\delta)] \,.$$

Then, an M-estimator can be defined as the solution to the empirical problem based on available data pairs $(Y_i, x_i)_{i=1}^n$ and defined as

$$\min_{(\beta,\delta)} \frac{1}{n}\sum_{i=1}^{n} W_\alpha(Y_i - x_i^{\mathrm{T}}\beta;\delta) \,,$$

This is a special case of the Hogg estimator (Koenker, 2005).

The Winkler loss is essentially a penalized version of the Δ-insensitive loss, with penalty term $\alpha\Delta$. The Winkler loss function allows to estimate δ_α thanks to the penalty term, while using the insensitive loss function would require to tune δ_α like a hyperparameter. We exploit this fact in the following with estimators for β_α and δ_α that depend on the only tuning aspect, namely, α.
We implemented the estimator in `R` based on an adaptive Newton method. Even median regression can be approximately estimated by choosing a suitably low coverage. At each iteration of the algorithm, the estimate of β_α is updated, then the one for δ_α is set based on the $1-\alpha$-quantile of current absolute residuals $|Y - \mu_\alpha|$. The Hessian for β_α is firstly approximated via Berndt-Hall-Hall-Hausman method and then corrected for information bias through a multiplicative constant that is estimated via optimization. This procedure could efficiently approximate median regression from the `R` package `quantreg`, which was too cumbersome for our long dataset.

3 Example

The dataset on focus collects data from twelve environmental sensors scattered across an office room in Villach, Austria. Readings relate to Y = temperature (in °C), X=clear light and Z=pressure. Each sensor provided data every ten seconds for six months. The dataset was shared by Brunello et al. (2021) and also analyzed with different aims in Lambardi di San Miniato et al. (2022). The first half of the data rows made up the training set, the remainder served as a test set. The prediction horizon was set equal to one hour. These data did feature a variety of anomalies that would hardly fit into a classical environmental model for regular system behavior. Any central prediction would not account for the extreme behavior of these sensors. The more flexible ETP approach would accommodate for them instead.
Each variable was centered around the instant cross-sensor median lagged by one hour. The median was adopted to account for spatial regularity across the office, while the lag was introduced to ease up the prediction. Y served as the response, while the covariates X and Z were lagged by one more hour for additional ease of prediction. The model is

$$\begin{aligned} Y_t =& \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-24} + \beta_3 Y_{t-25} \\ &+ \beta_4 X_{t-1} + \beta_5 X_{t-2} + \beta_6 X_{t-25} + \beta_7 X_{t-26} \\ &+ \beta_8 Z_{t-1} + \beta_9 Z_{t-2} + \beta_{10} Z_{t-25} + \beta_{11} Z_{t-26} + \epsilon_t \,, \end{aligned}$$

for $t = 27, \ldots, T$. The included lags account for seasonal auto-regression.
For $1-\alpha = 50\%$, ETP and UTP intervals are very similar and not reported here for space reasons. However the results are sharply different for $1-\alpha = 95\%$.
Figure 1 reports the cumulative distribution function (CDF) of absolute prediction error beyond thresholds, hence higher curves are better. The

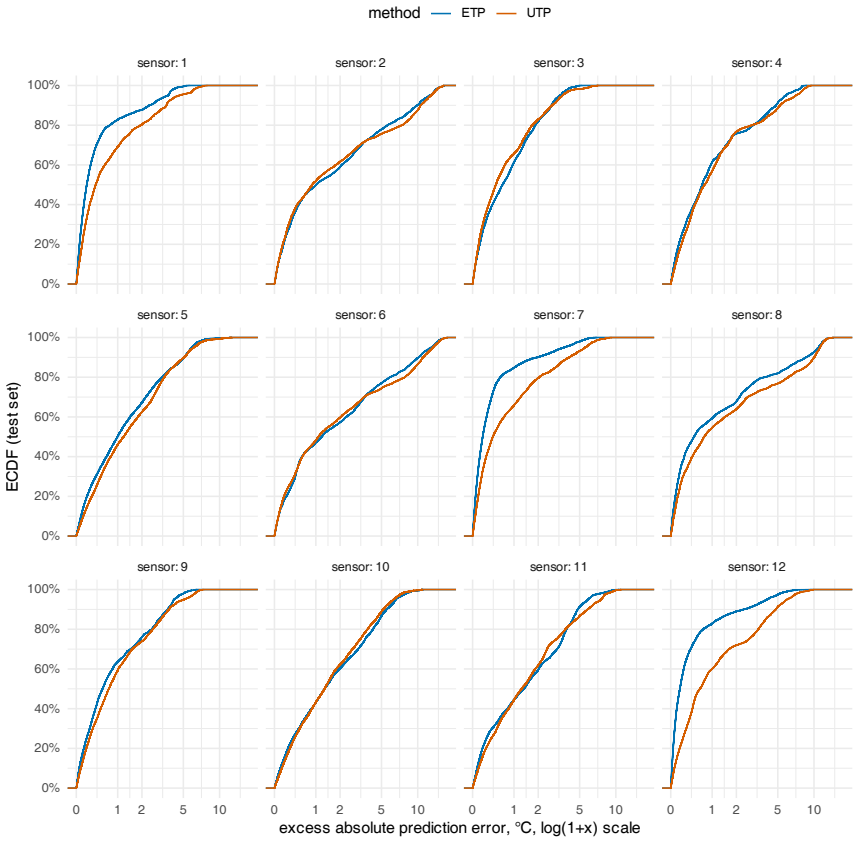


FIGURE 1. Distribution of test error per sensor, solid lines for ETP and dashed lines for UTP, 95% prediction.

UTP intervals perform better only occasionally and hardly control tail events. Sensors 1, 7, and 12 behave more regularly under the ETP prediction, net of all the remaining aspects. The UTP viewpoint is centered on some median sensor that cannot capture all the variability in the system on its own.

Residual diagnostics in Figure 2 show a correlation between prediction error and covariates, in the sense that the distribution of each covariate should not depend on the sign of prediction error, when this error exceeds thresholds. ETP is designed to remove the correlation, while UTP makes systematic prediction errors.

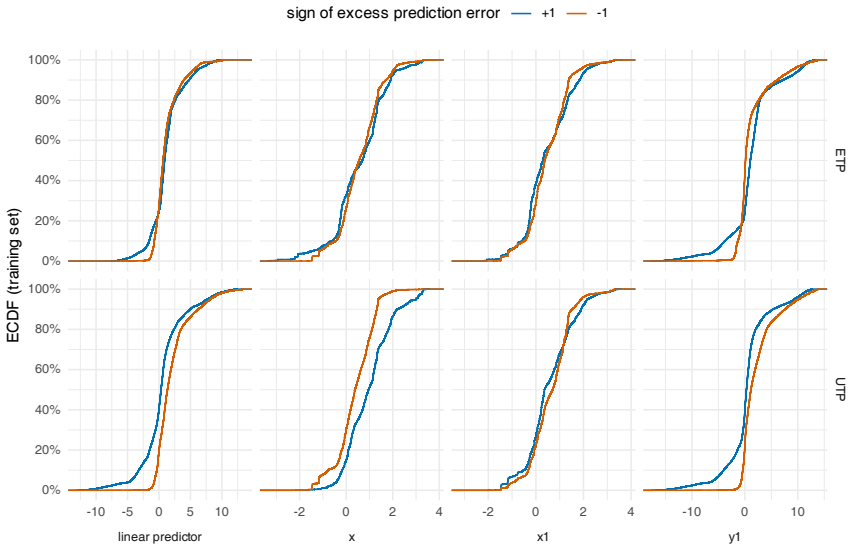


FIGURE 2. Distribution of covariates per error sign, ETP on top and UTP on bottom, 95% prediction. Here, `x1` and `y1` are X and Y lagged by one hour.

4 Closing remarks

In this paper, we highlighted some of the appealing features of ETP prediction intervals, such as robustness to outliers (as inherited from the quantile regression logic) and lower bias towards groups of observations. Further developments may focus on making the intervals calibrated within short time periods in the case of data streams, as is also a concern in conformal prediction and other resampling techniques.

As a plus, we see that even the Winkler loss function can be optimized by means of methods that assume some kind of regularity, like the Berndt-Hall-Hall-Hausman method. We only circumvented the computational issues arising from a non-smooth loss function, but any advances in this field will surely benefit also the estimation of ETP intervals.

Future developments may naturally relate to quasi-likelihood and other more complicated settings with bounded response variables and other model constraints. Heteroscedasticity may be effectively accounted for by also expressing the calibration parameter δ_α as a function of covariates. Care must be taken as to whether the quantile properties of $\mu_\alpha - \delta_\alpha$ and $\mu_\alpha + \delta_\alpha$ are retained under this generalization.

ETP prediction intervals prove capable to summarize a relevant regular pattern and can avoid several false positives in outlier detection. This is a major advantage in sensor data analytics, as it allows to keep system

information safe from data contamination and anomalies, due to a quantile-based approach.

References

Brehmer, J.R., and Gneiting, T. (2021). Scoring interval forecasts: Equal-tailed, shortest, and modal interval. *Bernoulli*, **27**, 1993 – 2010.

Brunello, A., Urgolo, A., Pittino, F., Montvay, A., and Montanari, A. (2021). Virtual sensing and sensors selection for efficient temperature monitoring in indoor environments. *Sensors*, **21**, 2728.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, **51**, 186 – 192.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. (2022). quantreg: Quantile Regression. R package version 5.94. https://CRAN.R-project.org/package=quantreg

Lambardi di San Miniato, M., Bellio, R., Grassetti, L., and Vidoni, P. (2022). Separable spatio-temporal kriging for fast virtual sensing. *Applied Stochastic Models in Business and Industry*, **38**, 806 – 829.

Vapnik, V.N. (1998). *Statistical Learning Theory*. New York: Wiley & Sons.

Asymmetry issues with non-penalized parameters in Laplace P-splines models

Philippe Lambert[1,2] and Oswaldo Gressani[3]

[1] Institut de Mathématique, Université de Liège, Liège, Belgium
[2] Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Louvain-la-Neuve, Belgium
[3] I-BioStat, Data Science Institute, Hasselt University, Belgium

E-mail for correspondence: `p.lambert@uliege.be`

Abstract: Laplace P-spline models (LPS) combine the P-spline smoother with Laplace approximations to perform fast Bayesian inference without the need for Monte Carlo methods. While this approach is appropriate for penalized parameters, inference may be misleading for others when there is little prior or empirical information about them. We propose an update of the LPS methodology by splitting the parameter space in two subsets. The first set involves parameters for which the joint posterior distribution of carefully selected linear projections is approximated using asymmetric families, while the conditional posterior distribution of penalised parameters can be treated using Laplace approximations. The method remains entirely sampling-free and enables fast inference in a Bayesian framework. The methodology is illustrated with an additive model for ordinal survey data.

Keywords: Additive model; Bayesian P-splines; Laplace approximation; Skewness.

1 Laplace approximation and Bayesian P-splines

Consider a regression model describing the conditional distribution of a response y for given covariates $\boldsymbol{x}$. Denote by $\boldsymbol{\xi}$ the model parameters: it includes the regression and spline parameters, plus possibly the (log of the) scale and (unconstrained transformed) shape parameters. Denote by $p(\boldsymbol{\xi}|\boldsymbol{\eta})$ the joint prior density of $\boldsymbol{\xi}$ conditionally on hyperparameters $\boldsymbol{\eta}$. In the context of a P-spline model, the latter might include J unknown smooth functionals $f_j(\cdot)$ $(j = 1, \ldots, J)$ specified as linear functions of B-splines spanning each argument range (Marx & Eilers 1998). We assume that the

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

joint conditional prior for the vector $\boldsymbol{\theta}$ stacking all the vectors of penalized B-spline coefficients in the model is $p(\boldsymbol{\theta}|\boldsymbol{\lambda}) \propto \exp\left(-\frac{1}{2}\,\boldsymbol{\theta}^\top \mathcal{P}_\lambda\, \boldsymbol{\theta}\right)$, where $\mathcal{P}_\lambda$ is a positive semi-definite matrix and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)$ is the vector of penalty parameters. The vector of model parameters can be reorganized as follows, $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top, \boldsymbol{\theta}^\top)^\top \in \mathbb{R}^{k_1+k_2}$, where $\boldsymbol{\gamma} \in \mathbb{R}^{k_1}$ denotes the vector of non-penalized parameters. If $\mathcal{D}$ generically denotes the available data and if $\boldsymbol{\lambda}$ stands for the vector of hyperparameters $\boldsymbol{\eta}$ in the context of P-spline models, then the joint posterior for $\boldsymbol{\xi}$ is

$$p(\boldsymbol{\xi}, \boldsymbol{\lambda}|\mathcal{D}) \propto \mathcal{L}(\boldsymbol{\xi}|\mathcal{D})\, p(\boldsymbol{\gamma})\, p(\boldsymbol{\theta}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda}).$$

It is typically explored using Markov chain Monte Carlo methods (MCMC) (Brezger & Lang 2006). We build up on the methodology described for additive models in Gressani & Lambert (2021) and in Lambert (2021) where Laplace approximations to the conditional posterior of $(\boldsymbol{\xi}|\boldsymbol{\lambda}, \mathcal{D})$ and an additional approximation to the marginal posterior of $(\boldsymbol{\lambda}|\mathcal{D})$ enable to bypass sampling algorithms. Thanks to the Gaussian Markov field (GMRF) prior (Rue & Held 2005) assumed for the penalized parameters $\boldsymbol{\theta}$, the Normal approximation to the conditional posterior, $(\boldsymbol{\xi}|\boldsymbol{\lambda}, \mathcal{D}) \dot{\sim} \mathcal{N}_{k_1+k_2}(\hat{\boldsymbol{\xi}}_\lambda, \Sigma_\lambda)$ with $\widehat{\Sigma}_\lambda^{-1} = -\partial^2 \log p(\hat{\boldsymbol{\xi}}_\lambda|\boldsymbol{\lambda}, \mathcal{D})/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^\top$, is usually excellent, yielding for the marginal posterior of $\boldsymbol{\lambda}$,

$$\widetilde{p}_\lambda(\boldsymbol{\lambda}|\mathcal{D}) = p(\boldsymbol{\xi}, \boldsymbol{\lambda}|\mathcal{D})/\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda}, \mathcal{D}) \propto p(\hat{\boldsymbol{\xi}}_\lambda, \boldsymbol{\lambda}|\mathcal{D}) \left|\widehat{\Sigma}_\lambda\right|^{\frac{1}{2}},$$

see Tierney & Kadane (1986) for the same strategy in the approximation of a marginal distribution and Wood & Fasiolo (2017) for related work. The Laplace approximation might not be suitable for some of the unpenalized parameters $\boldsymbol{\gamma}$ in $\boldsymbol{\xi}$, especially when the combined information coming from their prior and the likelihood is sparse.

2 Asymmetric posterior for non-penalized parameters

Using the same arguments as before, the conditional posterior of $(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D})$ is approximately Gaussian, $(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D}) \dot{\sim} \mathcal{N}_{k_2}\left(\mathbb{E}(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D}), \widehat{\Sigma}_\lambda^{\theta|\gamma}\right)$, where $\mathbb{E}(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D}) = \hat{\theta}_\lambda + \widehat{\Sigma}_\lambda^{\theta\gamma}\left(\widehat{\Sigma}_\lambda^{\gamma\gamma}\right)^{-1}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}_\lambda)$ and $\widehat{\Sigma}_\lambda^{\theta|\gamma} = \widehat{\Sigma}_\lambda^{\theta\theta} - \widehat{\Sigma}_\lambda^{\theta\gamma}\left(\widehat{\Sigma}_\lambda^{\gamma\gamma}\right)^{-1}\widehat{\Sigma}_\lambda^{\gamma\theta}$. Substituting that Normal approximation in the denominator of the following identity, $p_\gamma(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D}) = p(\boldsymbol{\gamma}, \boldsymbol{\theta}|\boldsymbol{\lambda}, \mathcal{D})/p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D})$, one gets

$$p_\gamma(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D}) \propto p\big(\boldsymbol{\gamma}, \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D})\big|\boldsymbol{\lambda}, \mathcal{D}\big) \left|\widehat{\Sigma}_\lambda^{\theta|\gamma}\right|^{\frac{1}{2}}.$$

Consider now the singular value decomposition (SVD) of $\widehat{\Sigma}_\lambda^{\gamma\gamma} = \mathbf{V}\mathbf{Z}\mathbf{V}^\top$ where $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_{k_1}]$ denotes the matrix of orthonormal eigenvectors, $\boldsymbol{\zeta}$

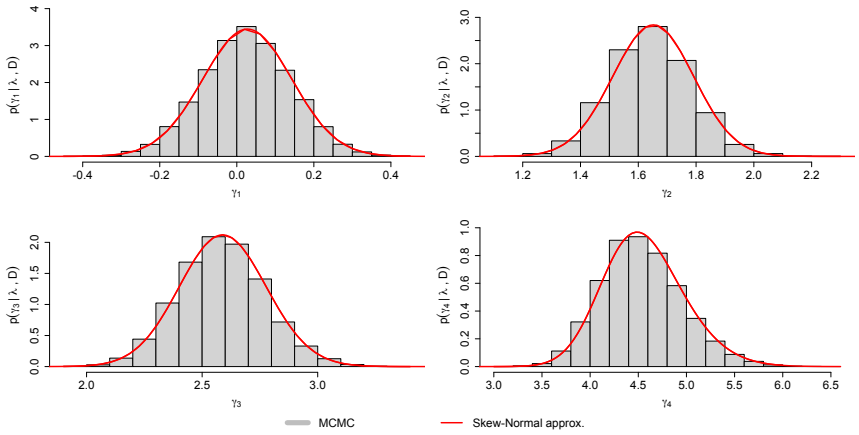


FIGURE 1. Approximated marginal posterior density for $(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D})$ compared to MCMC samples when $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$.

the eigenvalues and $\mathrm{Z} = \mathrm{diag}(\boldsymbol{\zeta})$. Let $\tilde{\boldsymbol{\gamma}} = \mathrm{Z}^{-\frac{1}{2}}\mathbf{V}^\top(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}_\lambda)$. Combined with the previous approximation, one gets $p_{\tilde{\gamma}_s}(\tilde{\gamma}_s|\boldsymbol{\lambda}, \mathcal{D}) \propto p_\gamma(\hat{\boldsymbol{\gamma}}_\lambda + \tilde{\gamma}_s\sqrt{\zeta_s}\mathbf{v}_s|\boldsymbol{\lambda}, \mathcal{D})$. Each of these densities can be approximated using a skew-Student (ST), yielding $p_\gamma(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D}) \approx \prod_{s=1}^{k_1} \frac{1}{\sqrt{\zeta_s}}\varphi(\tilde{\gamma}_s|\tilde{\psi}_s, \tilde{\omega}_s^2, \tilde{\alpha}_s)$. Therefore, one has the following stochastic representation for $(\boldsymbol{\xi}, \boldsymbol{\lambda}|\mathcal{D})$,

$$(\boldsymbol{\xi}, \boldsymbol{\lambda}|\mathcal{D}) \dot{\sim} \mathcal{N}_{k_2}\left(\mathbb{E}(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D}), \widehat{\Sigma}_\lambda^{\theta|\gamma}\right) \times \prod_{s=1}^{k_1} \mathrm{ST}(\tilde{\gamma}_s|\tilde{\psi}_s, \tilde{\omega}_s^2, \tilde{\alpha}_s) \times (\boldsymbol{\lambda}|\mathcal{D}) \ .$$

It can be used to quantify uncertainty on any function of the model parameters or to make predictions. We refer to Lambert & Gressani (2023) for more details.

3 Application on survey data

The proposed methodology is illustrated on data from the European Social Survey (ESS 2018) for Wallonia, one of the three regions in Belgium. Each of the 552 participants (aged at least 15) was asked to react to the following statement, *Gay men and lesbians should be free to live their own life as they wish*, with a positioning on a Likert scale going from 1 (=*Agree strongly*) to 5 (=*Disagree strongly*), with 3 labelled as *Neither agree nor disagree* (with relative frequencies 1: 54.9% ; 2: 30.4% ; 3: 8.2% ; 4: 5.4% ; 5: 1.1%). That ordinal response was analyzed using the proportional odds model with the number of completed years of education (14.1 ± 4.4 years) and age (47.3 ± 18.5 years) entering as additive terms,

$$\mathrm{logit}[P(Y \leq r|\mathbf{x})] = \eta_r = \gamma_r + f_1(\mathrm{educ}) + f_2(\mathrm{age}) \ .$$

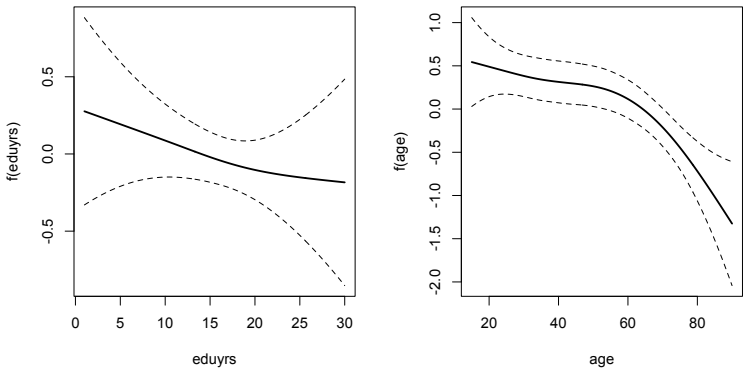


FIGURE 2. Fitted additive terms for `eduyrs` and `age` with pointwise 95% credible intervals.

There is a marked positive skewness in the marginal posterior distribution of γ_4, caused by the small proportion of respondents in the survey expressing an explicit disagreement with the submitted statement. The proposed approximation to the posterior distribution of $(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D})$ can be visualized on Fig. 1 (solid red curves) and confronted to MCMC samples (grey histograms) taken as a proxy for the true underlying distributions. The close agreement between the two results confirm the quality of the approximation bypassing the need for Monte Carlo sampling. The estimated additive terms in Fig. 2 suggest a non-significant effect of `eduyrs`, but a tolerant perception of homosexuality tending to decrease with `age`, with a marked change in attitude revealed beyond age 60.
The R-package `ordgam` to reproduce the results of the paper can be downloaded from https://github.com/plambertULiege/ordgam.

Acknowledgements

Philippe Lambert acknowledges the support of the ARC project IMAL (grant 20/25-107) financed by the Wallonia-Brussels Federation and granted by the Académie Universitaire Louvain.

References

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, **50**, 967–991.

ESS Round 9 (2018). European Social Survey Round 9. Data file Edition 3.1. Sikt - Norwegian Agency for shared services in education and

research, Norway - Data Archive and distributor of ESS data for ESS ERIC. doi: 10.21338/NSD-ESS9-2018.

Gressani, O. and Lambert, P. (2021). Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines. *Computational Statistics & Data Analysis*, **154**, 107088.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics and Data Analysis*, **51** (5), 2542–2558.

Lambert, P. (2021). Fast Bayesian inference using Laplace approximations in nonparametric double additive location-scale models with right- and interval-censored data. *Computational Statistics & Data Analysis*, **161**, 107250.

Lambert, P. and Gressani, 0. (2023). Penalty parameter selection and asymmetry corrections to Laplace approximations in Bayesian P-splines models. *Statistical Modelling* (in press).

Marx, B. D. and Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, **28**, 193–209.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: CRC Press.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81** (393), 82–86.

Wood, S. N. and Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*, **73**, 1071–1081.

Local moment matching with Gamma mixtures and automatic smoothness selection

Oskar Laverny[1], Philippe Lambert[1,2]

[1] Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Louvain-la-Neuve, Belgium
[2] Institut de Mathématique, Université de Liège, Liège, Belgium

E-mail for correspondence: oskar.laverny@uclouvain.be

Abstract: We consider the class of Erlang mixtures for the task of density estimation on the positive real line when the only available information is given as local moments, a histogram with potentially higher order moments in some bins. By construction, the obtained moment problem is ill-posed and requires regularization. Several penalties can be used for such a task, such as a lasso penalty for sparsity of the representation, but we focus here on a simplified roughness penalty from the P-splines literature. We show that the corresponding hyperparameter can be selected without cross-validation through the computation of the so-called effective dimension of the estimator, which makes the estimator practical and adapted to these summarized information settings. The flexibility of the local moments representations allows interesting additions such as the enforcement of Value-at-Risk and Tail Value-at-Risk constraints on the resulting estimator, making the procedure suitable for the estimation of heavy-tailed densities.

Keywords: Erlang mixtures, P-Splines, Local moment matching

1 Local moment matching problem

Consider $X_1, ...X_N$ a N-sample of a positive real random variable X, and let $\mathcal{B} = (B_1, ..., B_J)$ be a finite partition of $\mathbb{R}_+$, where $B_j = [b_{j-1}, b_j[$ $(j = 1, \dots, J)$ are called *bins*. Let $K_1, ..., K_J \in \mathbb{N}$ be the maximum orders of the observed empirical moments within bins and denote the set of these empirical moments by

$$\hat{\boldsymbol{\mu}} = \left\{ \hat{\mu}_{j,k_j} = \frac{1}{N} \sum_{i=1}^{N} X_i^k \mathbf{1}_{X_i \in B_j} : j = 1, \dots, J \ ; \ k_j = 1, \dots, K_j \right\}.$$

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

We denote separately by $\hat{\boldsymbol{\pi}} = \left\{\hat{\pi}_j = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{X_i \in B_j}\right\}$ the zeroth moments. In this paper, we discuss the reconstruction of the distribution of X from $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}})$. Table 1 gives a numerical example, constructed from $N = 750$ samples from a LogNormal$(\mu = 0, \sigma = 0.5)$. The exposed structure is quite common with insurance losses, where *attritional* and *large* losses are usually treated separately. The lack of detailed information in this case would be caused by confidentiality issues. This type of data may also occur in other contexts, see e.g. Lambert (2022).

TABLE 1. Summary statistics for the lognormal example.

j	$[b_{j-1}, b_j[$	n_j	K_j	$\hat{\pi}_j$	$\hat{\mu}_{j,1}$	$\hat{\mu}_{j,2}$	$\hat{\mu}_{j,3}$
1	$[0.00, 1.90[$	677	3	0.900	0.86	0.96	1.21
2	$[1.90, 3.20[$	64	3	0.085	0.20	0.46	1.10
3	$[3.20, +\infty[$	9	3	0.012	0.05	0.24	1.21

Note that zeroth and first order moment in the last bin conveniently corresponds to the value at risk and tail value at risk at level $\alpha = 0.01$. In this work, we propose to estimate the density of X in the semi-parametric class of Erlang mixtures from the observed local moments.

2 Erlang mixtures

A random variable X has a mixed Erlang distribution with *scale* $s \in \mathbb{R}_+$ and *mixing probability measure* $\nu \in \mathcal{P}(\mathbb{R}_+)$, denoted by $X \sim \mathrm{M}\Gamma(\nu, s)$, if and only if its moment generating function can be written as

$$M(t) = \int (1 - st)^{-\alpha} \nu(d\alpha).$$

Tijms (1994, Theorem 3.9.1) shows that this model is dense in the set of positive random variables, even under the restriction that $\mathrm{Supp}(\nu) \subseteq \mathbb{N}$. This result has generated a lot of interest and several extensions exists for multivariate, censored or truncated data, see e.g. Caussette *et al.* (2016) or Gui *et al.* (2021). Denote by $\boldsymbol{\pi}(\nu, s)$, $\boldsymbol{\mu}(\nu, s)$ and $\boldsymbol{\Sigma}(\nu, s)$ respectively the expectations of $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\mu}}$ and the variance of $\hat{\boldsymbol{\mu}}$ under the hypothesis that $X \sim \mathrm{M}\Gamma(\nu, s)$. Since the underlying raw data are i.i.d, the vector $\hat{\boldsymbol{\pi}}$ follows a multinomial distribution. Then, conditionally on the value of $\hat{\boldsymbol{\pi}}$, the vector $\hat{\boldsymbol{\mu}}$ can be approximated asymptotically by a Gaussian random vector. The log-likelihood of our model is then given (under this approximation) by:

$$\ell(\nu, s|\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}) = \hat{\boldsymbol{\pi}}^{\mathrm{T}} \ln(\boldsymbol{\pi}(\nu, s)) - \frac{1}{2}\ln|\boldsymbol{\Sigma}(\nu, s)| - \frac{1}{2}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}(\nu, s)\|^2_{\boldsymbol{\Sigma}(\nu,s)}.$$

Unfortunately, maximizing ℓ directly produces rough density estimates. To mitigate this behavior, we propose to penalize the roughness of ν directly.

3 Discrete smoothness penalization

Consider from now on that the measure ν is supported on $\{1, ..., n\}$, that is $\nu = \sum_{i=1}^{n} p_i \delta_i$. Enforcing smoothness in density estimation is usually done through a r^{th}-order penalty:

$$\text{Pen}(\nu, s) = \lambda_2 \int \left(f^{(r)}_{\text{M}\Gamma(\nu,s)}(x) \right)^2 dx = \lambda_2 s^{-2r-1} \boldsymbol{p}^{\text{T}} \boldsymbol{P}(r) \boldsymbol{p}.$$

If the factor s^{-2r-1} can simply be incorporated into λ_2, the $n \times n$ matrix $\boldsymbol{P}(r)$ is dense, which induces high computational cost. In addition, λ_2 cannot be calibrated through cross-validation since holding out testing data would not be acceptable in the considered sparse information context. Leveraging the P-spline literature, we propose to replace this penalty by a discrete penalization of the mixture weights $\{p_1, ..., p_n\}$. Neglecting an additional s^{-1} factor, the mode of the $(i+1)^{th}$ density in the mixture is at

$$(x_i, y_i) := \left(si, \frac{i^i e^{-i}}{i!} \right), i \in \{1, \ldots, n\}. \tag{1}$$

Note that $x_1, \ldots, x_n$ are regularly distributed on the positive real line. Therefore, we suggest to monitor the regularity of the density estimate by the regularity of the sequence of weighted modes, $p_1 y_1, ..., p_n y_n$. The corresponding penalty matrix, $\tilde{\boldsymbol{P}}(r) = \boldsymbol{D}(r)^{\text{T}} \boldsymbol{D}(r)$ (where $\boldsymbol{D}(r)$, is the r^{th} finite difference matrix) is sparse – only the $(2r-1)$ central diagonals are non-zero – which is computationally efficient. Furthermore, the penalty parameter λ_2 can be selected iteratively using the concept of effective dimension, see e.g. Eilers (2018):

$$\lambda_2 \longleftarrow \frac{\text{tr}\left(\left(\boldsymbol{H} + \lambda_2 \tilde{\boldsymbol{P}} \right)^{-1} \boldsymbol{H} \right) - r}{\boldsymbol{p}^{\text{T}} \tilde{\boldsymbol{P}}(r) \boldsymbol{p}}.$$

The trace in the numerator is the effective dimension of the model, constructed from the Hessian matrix $\boldsymbol{H}$ of the loss function w.r.t. $\boldsymbol{p}$. We provide detailed arguments for the proposed estimation techniques and detailed simulation studies to assess the performance of the proposed approach.

4 Illustration

From the set of moments depicted in Table 1, we obtain the optimal Erlang mixture depicted in Figure 1 (a,b,c). Figure 1 (d) shows the convergence of the procedure w.r.t. the number K of empirical moments used to estimate the density, as assessed using $S = 50$ resamples from the lognormal model. It shows that the distance between the density estimate and the true density (as measured by ℓ_1 and ℓ_2 distances between quantile functions and distribution functions, as well as the Kullback-Leibler divergence) decreases significantly with the number of available empirical local moments.

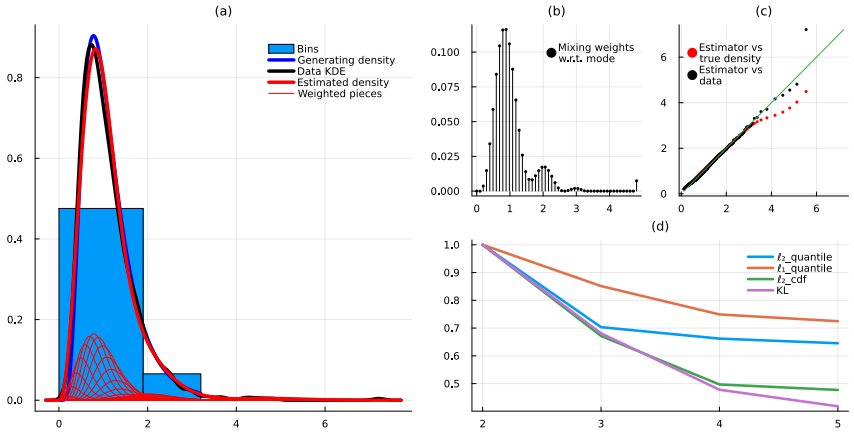


FIGURE 1. Panels (a), (b) and (c) provide information when 3 moments are observed within each bin on top of the frequencies. The histogram in (a) represents $\hat{\boldsymbol{\pi}}$, while the red curve represents the density estimate. Couples $(x_i, p_i y_i)$ from Equation (1) in panels (b) & (c) represents quantile-quantile plots of the estimated density against the underlying raw data and the true density. Panels (d) represents the median of the distance statistics across lognormal resamples for an increasing number of observed local moments, renormalized to be 1 for the least informative setting.

References

Cossette, H., Landriault, D., Marceau E., and Moutanabbir, K. (2016) Moment-based approximation with mixed Erlang distributions. In: *Variance, 2016, vol. 10, no 1, p. 161-182*

Eilers, P. H. C. (2018) The truth about the effective dimension. In: *Statistica Neerlandica 72, no 3: 201-9.*

Gui, W., Huang, R., and Lin, X.S. (2021) Fitting multivariate Erlang mixtures to data: a roughness penalty approach. In: *Journal of Computational and Applied Mathematics 386: 113216.*

Lambert, P. (2022) Nonparametric density estimation and risk quantification from tabulated sample moments. In: *Insurance: Mathematics and Economics, 108, 177-189.*

Tijms, H.C (1994) *Stochastic models: an algorithmic approach.* Vol. 303. John Wiley & Sons Incorporated

Linear mixed modelling of federated data when only the mean, covariance, and sample size are available

Marie Analiz April Limpoco[1], Christel Faes[1], Niel Hens[1,2]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Data Science Institute (DSI), Hasselt University, Hasselt, Belgium
[2] Centre for Health Economic Research and Modelling Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute, Antwerp University, Antwerp, Belgium

E-mail for correspondence: liz.limpoco@uhasselt.be

Abstract: Data confidentiality is becoming increasingly important, resulting in stricter policies regarding accessibility of individual records. A federated data setting wherein analysis is performed without the need to pool all samples from multiple sources, thus enabling the data to stay with the data custodian, offers a compromise. This paper proposes a framework for fitting a linear mixed model when only the mean, covariance, and sample size of the federated data are made available. This is largely anchored on the statistical sufficiency and likelihood principles applied to linear models with and without random effects. We apply this approach to a real data set and show that we can obtain identical inference as the strategy that uses the pooled unit-level data. Although similar to individual patient data meta-analysis settings, our approach has the benefit of accessing the correlation structure among relevant variables, which enriches the modelling process. Simplicity, computational efficiency, and potentially wider scope of implementation through any statistical software distinguish our approach from the existing strategies in the literature. Potential applications of this methodology include health research and ecological inference, to name a few.

Keywords: Linear Mixed Model (LMM); federated data; sufficiency principle; aggregate data; data confidentiality.

1 Background

A federated data setting serves as a compromise to preserve data privacy while still permitting data analysis. Classical statistical and machine learn-

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

ing models require unit-level samples, and in a typical federated data analysis, a network is set up such that the models are implemented at each data provider's node and only the parameters are sent to the analyst's node, which then performs some kind of aggregation to build the global model. This is typically an iterative process, requiring regular communication among the data providers and the analyst. This paper proposes a simple framework for fitting linear models and linear mixed models when only the mean, covariance, and sample size of the federated data are provided once by the data custodians. It exploits the principles of data reduction namely the sufficiency and likelihood principle (see e.g. Casella & Berger, 2002), which will be discussed in Section 2, along with the details of our proposed approach. Section 3 discusses its application to real patient data from hospitals and its potential in other fields before we provide a conclusion.

2 Proposed approach

2.1 Principles of data reduction

The core principles upon which this paper is anchored relate to the concept of statistical sufficiency and likelihood, which are discussed thoroughly in the book of Casella & Berger (2002). The sufficiency principle guarantees that the entire sample need not be available to make inferences about a parameter θ as long as a sufficient statistic $T(\mathbf{X})$ exists; that is, the inference about θ depends on the sample only through the sufficient statistic $T(\mathbf{X})$, such that a sample $\mathbf{x_1}$ having a sufficient statistic $T(\mathbf{x_1})$ will generate the same conclusion as another sample $\mathbf{x_2}$ if $T(\mathbf{x_1}) = T(\mathbf{x_2})$ even though $\mathbf{x_1} \neq \mathbf{x_2}$. In other words, even if the only information known is $T(\mathbf{x})$, inference about the parameter of interest θ can still be made, thus enabling data reduction without loss of important information contained in the sample about the parameter of interest. Casella & Berger (2002) proceed to argue that in a situation wherein only $T(\mathbf{x_1})$ and not the full sample $\mathbf{x_1}$ is available, the probability distribution given the sufficient statistic, denoted as $P(\mathbf{X} = \mathbf{x_2}|T(\mathbf{X}) = T(\mathbf{x_1}))$, can be used to draw a sample $\mathbf{x_2}$ and generate equivalent information about θ. However, this conditional probability distribution might be difficult to obtain in practice. In this case, the likelihood principle may provide additional support which does not necessitate a particular probability distribution. Taking for instance the likelihood for the population parameter μ of independent and identically distributed random variables $X_1, ..., X_n$ following a normal distribution $N(\mu, \sigma^2)$ with σ^2 known, we find that the likelihood principle implies that regardless of the values of $\mathbf{x_2}$, as long as its sample mean is identical to that of $\mathbf{x_1}$, the conclusion regarding μ will be the same. Hence, $\mathbf{x_2}$ can be randomly generated from any distribution as long as its sample mean is exactly equal to that of $\mathbf{x_1}$, and the inference regarding μ will still be equivalent.

2.2 Linear regression model

Given n observations, an intercept, and $p-1$ predictors, let $\mathbf{X}$ denote the $n \times p$ design matrix and $\mathbf{y}$ be the $n \times 1$ vector of responses. The least squares estimator can then be obtained using $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ which is also the maximum likelihood estimator under the normality assumption. Writing out the matrices reveals that the elements of $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{y}$ can be derived from the mean, covariance matrix, and sample size even without having access to the individual data points in $\mathbf{X}$. Extending to the case when there are m multiple sources of data but under the assumption that the observations coming from the same source are uncorrelated, the parameters are estimated by

$$\hat{\boldsymbol{\beta}} = \left(\sum_{h=1}^{m} \boldsymbol{X}_h^T\boldsymbol{X}_h\right)^{-1} \sum_{h=1}^{m} \boldsymbol{X}_h^T\boldsymbol{y}_h$$

wherein $\boldsymbol{X}_h$ and $\boldsymbol{y}_h$ are the design matrix and response vector, respectively, from data source h $(h = 1, ..., m)$ (Lee et. al., 2017).

For a practical implementation, we created functions in R which only require the sample size and the mean and covariance matrix of the response and predictors. We were able to generate exactly the same output (except for the residuals) as when the *lm* function is used on the pooled individual observations. The key to this is to adapt the QR decomposition implemented in the *lm* function to the case at hand; i.e. solving

$$R_{\boldsymbol{X}^T\boldsymbol{X}}\hat{\boldsymbol{\beta}} = (Q_{\boldsymbol{X}^T\boldsymbol{X}})^T\boldsymbol{X}^T\mathbf{y}$$

where $R_{\boldsymbol{X}^T\boldsymbol{X}}$ is an upper triangular matrix and $Q_{\boldsymbol{X}^T\boldsymbol{X}}$ is an orthogonal matrix obtained from the decomposition of $\boldsymbol{X}^T\boldsymbol{X}$. A major drawback of this strategy though is the need to use special functions to implement parameter estimation.

2.3 Linear mixed model

A more realistic assumption when handling federated data is that the observations from the same source are more similar than observations from different sources. To account for this, a linear mixed model is more appropriate. Let y_{hi} be the continuous response of individual i from data source h; $\mathbf{x}_{hi}$ is a p-dimensional vector consisting of an intercept and $p-1$ predictors; $\boldsymbol{\beta}$ is the vector of fixed effects; $\mathbf{z}_{hi}$ is the q-dimensional covariate vector corresponding to the q-dimensional random effects u_h representing the deviation of source h from the overall pattern; and ϵ_{hi} is the random error. The linear mixed model with source-level random effects is then given by

$$y_{hi} = \mathbf{x}_{hi}^T\boldsymbol{\beta} + \mathbf{z}_{hi}^T\mathbf{u}_h + \epsilon_{hi}$$

Parameter estimation through the log-likelihood involves

$$l(\boldsymbol{\beta}, \sigma^2, \mathbf{V}) = -\frac{1}{2}\sum_{h=1}^{m}\{\log|\Sigma_h| + (\mathbf{y}_h - \mathbf{X}_h\boldsymbol{\beta})^T\Sigma_h^{-1}(\mathbf{y}_h - \mathbf{X}_h\boldsymbol{\beta})\}$$

where $\mathbf{X}_h$ and $\mathbf{y}_h$ are the design matrix and response vector, respectively, of source h, $|.|$ is the matrix determinant and $\Sigma_h = \Sigma_h(\sigma^2, V) = \mathbf{Z}_h V \mathbf{Z}_h^T + \sigma^2 I_{n_h}$. Due to the seemingly entangled data and parameter matrices, it is not straightforward to identify the aggregate statistics that can be used in place of individual data points. Luo et. al. (2022) showed that by utilizing the Woodbury matrix identity and some linear algebra concepts, the data can be disentangled from the parameters to reconstruct the profile log-likelihood without the need for individual records. In their approach, only $\mathbf{X}_h^T\mathbf{X}_h$, $\mathbf{X}_h^T\mathbf{y}_h$, $\mathbf{y}_h^T\mathbf{y}_h$, and n_h from each data source h are required to perform parameter estimation either through maximum likelihood (ML) or restricted maximum likelihood (REML). This coincides with the idea of Papadimitropoulou et. al. (2018) who proposed a methodology in the context of meta-analysis for performing linear mixed modelling from the mean and standard deviation of the continuous outcome in the treatment and control group by generating what they coined as pseudo-IPD (individual patient data). These pseudo-IPD should have exactly the same mean and standard deviation as the ones provided in studies to exploit the sufficiency principle. We extend this approach to accommodate multiple variables and use it in the context of federated data. Given the mean vector ($\hat{\boldsymbol{\mu}}_h$), covariance matrix ($\hat{\boldsymbol{\Sigma}}_h$) and sample size n_h from data source h, the following algorithm generates pseudo-data for the $p-1$ predictors and response variable:

1. Generate $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_i, ...\mathbf{w}_n]^T$ which is an $n \times p$ matrix where each column is independently distributed as $N(0, 1)$ (although any distribution will do).

2. Compute the mean vector $\hat{\boldsymbol{\mu}}_{\boldsymbol{W}}$ $(p \times 1)$ and the covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{W}}$ $(p \times p)$ of $\mathbf{W}$.

3. Generate the ith pseudo-data point as

$$\mathbf{x}_i = \hat{\boldsymbol{\mu}}_h + L_{\hat{\boldsymbol{\Sigma}}_h}(L_{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{W}}})^{-1}(\mathbf{w}_i - \hat{\boldsymbol{\mu}}_{\mathbf{W}})$$

   where $L_{\hat{\boldsymbol{\Sigma}}_h}$ and $L_{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{W}}}$ are the lower triangular matrices of the Cholesky decomposition of the given covariance matrix $\hat{\boldsymbol{\Sigma}}_h$ and the covariance matrix of $\mathbf{W}$, respectively.

Linear mixed model estimation with a random intercept can then be performed using the generated pseudo-data in any statistical package.

3 Application to real data

We apply the proposed approach to model the [log] length of hospital stay after COVID-19 infection (in days) regressed on age (in years) and gender (0-female, 1-male), including an interaction effect. We estimated a linear mixed model on the actual data pooled from 98 hospitals and compared the results to those obtained using the proposed framework. Figure 1 displays the results, where we see that only the residuals are different between the two. Needless to say, residuals require the individual responses, hence the difference.

```
> summary(full_lmm)
Linear mixed model fit by REML ['lmerMod']
Formula: log(0.001 + as.numeric(length_stay_hosp)) ~ age * gender +
(1 |      hospital_name_admis)
   Data: full_data

REML criterion at convergence: 136858.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5820 -0.6156 -0.0151  0.5987  5.6178

Random effects:
 Groups              Name        Variance Std.Dev.
 hospital_name_admis (Intercept) 0.07431  0.2726
 Residual                        0.87057  0.9330
Number of obs: 50568, groups:  hospital_name_admis, 98

Fixed effects:
                  Estimate Std. Error t value
(Intercept)      0.7353458  0.0344059  21.373
age              0.0210421  0.0002749  76.557
genderHomme      0.2510726  0.0265238   9.466
age:genderHomme -0.0030813  0.0003902  -7.898

Correlation of Fixed Effects:
            (Intr) age    gndrHm
age         -0.526
genderHomme -0.388  0.663
age:gndrHmm  0.361 -0.689 -0.949
> # Compute AIC
> AIC(full_lmm)
[1] 136870.4
```

```
> summary(lmm_pseudo_ipd)
Linear mixed model fit by REML ['lmerMod']
Formula: log_length_stay_hosp ~ age + genderHomme + ageXgenderHomme
+      (1 | hospital_name_admis)
   Data: pooled_simdata

REML criterion at convergence: 136858.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.1098 -0.6663 -0.0008  0.6658  4.1765

Random effects:
 Groups              Name        Variance Std.Dev.
 hospital_name_admis (Intercept) 0.07431  0.2726
 Residual                        0.87057  0.9330
Number of obs: 50568, groups:  hospital_name_admis, 98

Fixed effects:
                  Estimate Std. Error t value
(Intercept)      0.7353458  0.0344059  21.373
age              0.0210421  0.0002749  76.557
genderHomme      0.2510726  0.0265238   9.466
ageXgenderHomme -0.0030813  0.0003902  -7.898

Correlation of Fixed Effects:
            (Intr) age    gndrHm
age         -0.526
genderHomme -0.388  0.663
ageXgndrHmm  0.361 -0.689 -0.949
> # Compute AIC
> AIC(lmm_pseudo_ipd)
[1] 136870.4
```

FIGURE 1. Comparing results of *lmer* function applied to the full actual data (left) and to the pseudo-data generated from the sufficient statistics (right).

Aside from generating exactly the same parameter estimates for the fixed and random effects, our approach can also replicate the results of performing partial F test and model selection via AIC. In contrast to the study of Luo et. al. (2022) which also yields exactly the same estimates for LMM, our approach is a simple one in the sense that it does not need sophisticated programmed functions or packages. Additionally, the concept can be applied using any statistical software that can estimate LMM parameters, thus enabling a wider scope of implementation among all data practitioners alike. Another advantage of our approach, aside from its simplicity, is the computational efficiency gained from generating only one set of synthetic data compared to methodologies that simulate data multiple times and aggregate the estimates to form a single parameter estimate. We are also spared from the question of how many simulations to run and which aggregation method to best implement. Lastly, in contrast to federated learning algorithms in the literature, our approach does not require more than one communication iteration among the multiple sources and the data analyst's computer, nor do we need to set up a network among the databases, hence

significantly minimizing, if not totally eliminating, the risk of disclosing sensitive data.

A major limitation of our proposed approach is the inability to compute residuals, which require individual response values from the original data. Another consequence of our approach is the inability to perform training and testing since partitioning the original data is not possible. A future endeavour is to generate pseudo-data with similar distributional properties as the unknown actual data, but this might require more than just the summary statistics.

Meta-analysis is a field related to federated learning in terms of constructing a common global model to analyze and synthesize the information from multiple studies or sources. Hence, we borrow some ideas to deal with the challenges of federated data analyses. However, Papadimitropoulou et. al. (2018) discuss that having access to individual-level data when performing meta-analysis is still preferred. Thus, although meta-analysis and FDA have similarities, using an aggregate data meta-analysis method directly to perform FDA may not necessarily be the best option.

4 Conclusion

In this paper, we have demonstrated that parameter estimation of a linear mixed model can be performed on federated data by generating synthetic data from the mean vector and covariance matrix provided by each data source. The principles of statistical sufficiency and likelihood provide a good theoretical support to the validity of the proposed framework. Estimates achieved from this approach are identical to those obtained from the pooled individual-level data. Extending this approach to generalized linear mixed models is a current work in progress. Potential applications of this methodology not only include health research. Fields such as ecological inference which is dealt with by social scientists, political scientists, economists, and ecologists can benefit from this approach.

References

Casella, G. and Berger, R.L. (2002). *Statistical Inference*. California: Duxbury Press.

Lee, J.Y.L., Brown, J.J., and Ryan, L.M. (2017). Sufficiency revisited: Rethinking statistical algorithms in the big data era *The American Statistician*, **71**, 202 – 208.

Papadimitropoulou, K., Stijnen, T., Dekkers, O.M., and le Cessie, S. (2019). One-stage random effects meta-analysis using linear mixed models for aggregate continuous outcome data. *Research Synthesis Methods*, **10**, 360 – 375.

Feedforward neural networks from a statistical-modelling perspective

Andrew McInerney[1], Kevin Burke[1]

[1] University of Limerick, Ireland

E-mail for correspondence: andrew.mcinerney@ul.ie

Abstract: Feedforward neural networks (FNNs) have many similarities to the models typically used in statistical modelling. Although they are often viewed as "black boxes", when embedded in a statistical framework, these models can become more interpretable. A statistical-modelling approach to FNNs is demonstrated, with significance testing and covariate-effect plots used for explainability.

Keywords: Neural networks; Hypothesis testing; p-values; Covariate effects.

1 Introduction

In recent years, neural networks have experienced great success in the prediction of complex problems (LeCun et al., 2015). However, while neural networks exhibit strong predictive performance, they are viewed as "black-box" algorithms, i.e., their predictions are not easily understood and difficult to interpret. However, feedforward neural networks (FNNs) can be viewed through a statistical lens and seen as an alternative statistical model for non-linear regression. Thus, we aim to leverage the inherent intelligibility present in statistical modelling and demonstrate the inferential capabilities of FNNs, highlighting how these models can be used for problems beyond "pure prediction". The testing of the irrelevant-input-node hypothesis can inform us whether the effect of a given covariate on the response is significantly different from zero (White, 1989). Combining this with covariate-effect plots and their associated uncertainty, the outputs of FNNs become more akin to classical statistical models.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Feedforward Neural Network

We assume a model of the form

$$y_i = \mathrm{NN}(x_i) + \varepsilon_i,$$

where y_i is the response for the ith individual with covariate vector $x_i = (1, x_{1i}, x_{2i}, \ldots, x_{pi})^{\mathrm{T}}$, ε_i is a random error that we assume to have a $N(0, \sigma^2)$ distribution, and

$$\mathrm{NN}(x_i) = \gamma_0 + \sum_{k=1}^{q} \gamma_k \phi \left(\sum_{j=0}^{p} \omega_{jk} x_{ji} \right)$$

is the neural network regression model, where p is the number of covariates (input nodes), and q is the number of hidden nodes. The parameters are: ω_{jk}, the weight that connects the jth input node to the kth hidden node; γ_k, the weight that connects the kth hidden node to the output node; and γ_0, the bias term associated with the output node. The function $\phi(\cdot)$ is the activation function for the hidden layer, which is often a logistic function. Given our assumption that $\varepsilon_i \sim N(0, \sigma^2)$, maximum likelihood is used to estimate the parameters.

3 Hypothesis Testing

Hypothesis tests can be used to determine the statistical significance of individual parameters, or more appropriately for neural networks, they can be used to determine the statistical significance of groups of parameters. As each input node has multiple weights associated with it, we can make use of the multiple-parameter Wald test to test a single hypothesis on each of these parameters, i.e., test the overall relevance of covariate x_j by testing $H_0 : \omega_j = 0_q$, where $\omega_j = (\omega_{j1}, \omega_{j2}, \ldots, \omega_{jq})^T$ is the vector of weights that connects input node j to the hidden layer and 0_q is a zero vector of length q. Using the fact that (asymptotically) $\hat{\omega}_j \sim N(\omega_j, \Sigma_{\hat{\omega}_j})$, where $\Sigma_{\hat{\omega}_j}$ is the relevant $q \times q$ sub-matrix of the variance-covariance matrix, $\hat{\Sigma}$ (which is the inverse of the observed information matrix). Then, we have that

$$(\hat{\omega}_j - \omega_j)^T \Sigma_{\hat{\omega}_j}^{-1} (\hat{\omega}_j - \omega_j) \sim \chi_q^2$$

from which a p-value can be obtained by setting $\omega_j = 0_q$ and comparing this statistic to the χ_q^2 distribution. However, the estimation of $\hat{\Sigma}$ can be problematic due to the issue of parameter redundancy in FNNs, which leads to unidentifiability in some of the parameters.

4 Covariate Effects

When modelling the relationship between a covariate and a response, there are two natural questions to ask. First, is there any relationship? This is covered by the significance testing presented above. Second, if there is a relationship, what is the nature of this relationship? To this end, we consider a graphical approach to understanding the potentially complex covariate effects that are captured by the neural network model. A common approach to assess the relationship between a covariate and the response is using partial dependence plots (Friedman, 2001). The "partial dependence" of the response on x_j can be estimated from the data using

$$\overline{\mathrm{NN}}_j(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{NN}(x_{(i,1)}, \ldots, x_{(i,j-1)}, x, x_{(i,j+1)}, \ldots, x_{(i,p)}), \qquad (1)$$

where $x_{(i,j)}$ is the value of the jth covariate for the ith individual. Equation 1 can be computed for a set of x values, and the pairs of points, $(x, \overline{\mathrm{NN}}_j(x))$, can be used to construct a plot.

However, while the partial dependence plot provides the change in the average predicted response value as x_j varies, it is also useful to consider the *difference* in the average predicted response for a d-unit increase in x_j. This then plays the same role as a regression coefficient obtained from classical statistical models. Thus, we define the effect of a d-unit increase in x_j on the response as

$$\hat{\beta}_j(x, d) = \overline{\mathrm{NN}}_j(x + d) - \overline{\mathrm{NN}}_j(x) \qquad (2)$$

where d is often set to one or the standard deviation of x_j. Again, the $(x, \hat{\beta}_j(x, d))$ pairs can be used to construct a plot, which we term a Partial Covariate-Effect plot (PCE).

As the weights of the neural network are estimated using maximum likelihood, there are a number of methods available to estimate the associated uncertainty of these functions, e.g., the delta method and bootstrapping.

5 Application to Data

The Boston housing data set is available in James et al. (2022). It contains information relating to housing in 506 communities in the Boston area in 1970. The aim of the study was to examine the relationship between twelve explanatory variables and the median house price for each community (`medv`). For this analysis, we will fit a neural network with all explanatory variables, however, for brevity, we focus on only two explanatory variables: the proportion of the population that fall into a 'lower status' categorisation (`lstat`), and the average number of rooms per dwelling (`rm`).

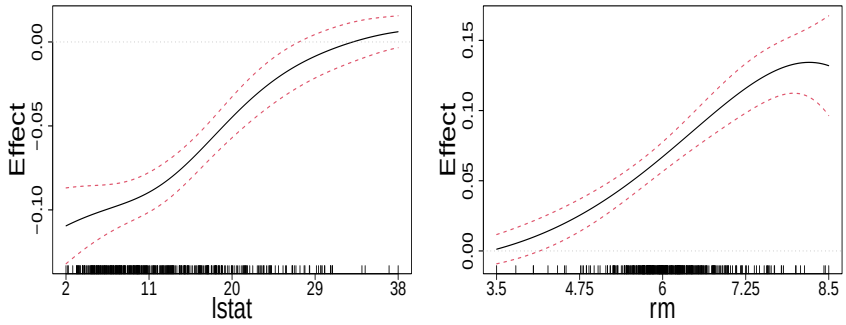


FIGURE 1. PCE plots for `lstat` (p-value < 0.001) and `rm` (p-value < 0.001).

The PCE plots for `lstat` and `rm`, and their associated p-values from the Wald test, are displayed in Figure 1. Both terms are deemed statistically significant. From the plots, we see that an increased value of `lstat` has a negative effect on `medv`, which weakens as `lstat` increases. On the other hand, `rm` has a positive effect, which increases as `rm` increases. All code for performing the Wald test and visualising the covariate-effect plots are available in the R package `statnn` (McInerney and Burke, 2022).

6 Discussion

Viewing neural networks as statistical models, and embedding them in a maximum-likelihood-based framework, can help improve their overall interpretability. This leads to more statistically-based outputs that are more familiar in the statistical modelling context. However, as mentioned above, there can be issues in the estimation of the variance-covariance matrix. Results from simulation studies, which will be discussed in our presentation, will highlight the scenarios where $\hat{\Sigma}$ is valid, and, show that, when this is the case, the aforementioned methods of uncertainty quantification perform as expected.

Acknowledgments: This work has emanated from research conducted with the financial support of Science Foundation Ireland under Grant numbers 18/CRT/6049 and 16/RC/3918.

References

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

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2022). *ISLR2: Introduction to Statistical Learning, Second Edition*, R package version 1.3-1.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, **521**, 436 – 444.

McInerney, A., and Burke, K. (2022). *statnn: Feedforward neural networks as statistical models*, R package version 0.0.0.9.

White, H. (1989). Learning in Artificial Neural Networks: A Statistical Perspective. *Neural Computation*, **1**, 435 – 464.

Modelling medical claims data using Markov-modulated marked Poisson processes

Sina Mews[1], Bastian Surmann[1], Lena Hasemann[1], Svenja Elkenkamp[1]

[1] Bielefeld University, Germany

E-mail for correspondence: sina.mews@uni-bielefeld.de

Abstract: We explore Markov-modulated marked Poisson processes (MMMPPs) as a natural framework for modelling patients' disease dynamics over time based on medical claims data. The approach is illustrated by modelling drug use and interval lengths between consecutive physician consultations of patients diagnosed with chronic obstructive pulmonary disease (COPD).

Keywords: continuous time; disease process; hidden Markov model (HMM); informative observation times; maximum likelihood.

1 Introduction

In medical claims data, observations do not only occur at random points in time but are also informative, i.e. driven by unobserved disease levels, as poor health conditions usually lead to more frequent and hence clustered healthcare interactions over time. Neglecting such an informative observation process in the analysis of disease dynamics potentially leads to biased parameter estimates (see, e.g., Pullenayegum and Lim, 2016). While joint models incorporating both informative observation times and disease processes exist, they mostly rely on data with directly observed disease stages and assume pre-scheduled examinations with informative missingness instead of patient-initiated visit times.
To jointly model the informative event times in claims data and additional data like patients' drug use collected at these event times, we propose to use Markov-modulated marked Poisson processes (MMMPPs) comprising two state-dependent processes: the observation process (corresponding to the event times) and the mark process (corresponding to event-specific information). Both processes are governed by an underlying state process,

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

which is modelled as a continuous-time Markov chain, whose states serve as proxies for the patients' latent disease levels. While MMMPPs are not new in themselves, to our knowledge, only Lange et al. (2015) and Alaa et al. (2017) use similar modelling approaches in the medical context. Our contribution extends these existing methods to a more general representation of the marks' state-dependent distributions and focuses on extracting information on disease dynamics from claims data.

2 Methods

We consider (claims) data containing information on the random observation times $T_0, T_1, \ldots, T_n$, $0 = T_0 < T_1 < \ldots < T_n$, which occur at irregularly spaced points in time, as well as additional data $Y_{t_0}, \ldots, Y_{t_n}$ collected at the realised observation times. These sequences of random variables are referred to as the observation process and the mark process, respectively, and depend on an underlying, unobserved state process $\{S_t\}_{t\geq 0}$. The state process is modelled as an N-state continuous-time Markov chain. Transitions between the states are governed by a transition intensity matrix $\mathbf{Q} = (q_{ij})_{i,j=1,\ldots,N}$, whose off-diagonal elements $q_{ij} \geq 0$, $i, j = 1, \ldots, N$, $i \neq j$, can be interpreted as the rates at which transitions from state i to state j occur. The duration in each state $i = 1, \ldots, N$ is exponentially distributed with parameter $q_{ii} = \sum_{j\neq i} q_{ij}$, where $-q_{ii}$ is the i-th diagonal entry in $\mathbf{Q}$. Furthermore, the initial distribution of the state process is denoted by $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_N)$, where $\delta_i = \Pr(S_0 = i)$.
The observation process is modelled as a doubly stochastic point process, namely a Markov-modulated Poisson process (MMPP) whose event rates $\lambda_i, i = 1, \ldots, N$, are selected by the underlying Markov chain. Within each state, the waiting times between consecutive events $X_{t_\tau} = T_\tau - T_{\tau-1}, \tau = 1, \ldots, n$, are exponentially distributed with parameter λ_i. Specifying the observation process as an MMPP thus accounts for the time-varying intensity of the observations and their temporal dependence. For the mark process, we assume the distribution of a variable (i.e. mark) Y_{t_τ} collected at observation time t_τ to be fully determined by its underlying state, with the Markov chain selecting which state-dependent distribution $f_i(y_{t_\tau}) = f(y_{t_\tau}|s_{t_\tau} = i)$ is active at time t_τ. As the state-dependent distributions can take on any (parametric) form, various data types like binary, count, or continuous variables can be considered in the mark process.
Subject to the unobserved Markovian state process, the MMMPP jointly models the mark process and the observation process. To evaluate the corresponding likelihood of the model, inferential tools from the hidden Markov model (HMM) framework, in particular the corresponding efficient algorithms for parameter estimation, can be applied. Let the integer $\tau = 1, 2, \ldots, n$ denote the index of the observation in the sequence and define the observation process by its waiting times. Then the likelihood

of the observed sequence $\{(x_\tau, y_\tau)\}_{\tau\in\{0,1,\dots,n\}}$ can be calculated using the HMM-based forward algorithm (Lu, 2012):

$$\mathcal{L} = \boldsymbol{\delta}\mathbf{P}(y_0)\Big(\prod_{\tau=1}^{n}\exp\big((\mathbf{Q}-\boldsymbol{\Lambda})x_\tau\big)\boldsymbol{\Lambda}\mathbf{P}(y_\tau)\Big)\mathbf{1}, \tag{1}$$

where $\mathbf{P}(y_\tau) = \mathrm{diag}\big\{f_1(y_\tau),\dots,f_N(y_\tau)\big\}$ and $\boldsymbol{\Lambda} = \mathrm{diag}\big\{\lambda_1,\dots,\lambda_N\big\}$ are diagonal matrices and $\mathbf{1}\in\mathbb{R}^N$ denotes a column vector of ones. To estimate the model parameters, we numerically maximise the joint likelihood over all patients, which is the product of the individual likelihoods given in (1).

3 Case study

3.1 Data and model formulation

We consider data from one of the largest statutory health insurance (SHI) companies in Germany, covering the years 2005 to 2020. Our study population consists of patients initially diagnosed with chronic obstructive pulmonary disease (COPD) in 2008 who have a mild to moderate age-adjusted Charlson comorbidity index (ACCI). The final data set includes 470 persons (141 males and 329 females) with 112,297 observations in total, covering a mean period of 12 years per person (min: 6; max: 13) after initial COPD diagnosis.

For modelling COPD patients' general health condition over time, we consider the interval length between consecutive physician consultations (i.e. the waiting times) and patients' drug use measured in daily defined doses (DDDs) — a standardised unit for drug consumption — based on physicians' prescriptions contained in the SHI data. These are jointly modelled as an MMMPP, where the waiting times follow state-dependent exponential distributions, while the DDDs are assumed to follow a zero-adjusted gamma distribution with state-specific parameters. As we are interested in inter-individual differences in the state-switching dynamics, we model the state transition intensities as a function of patients' sex, their age at initial diagnosis ($ageD$), and the ACCI (dichotomised into either mild or moderate age-adjusted comorbidities):

$$q_{ij} = \exp\left(\beta_0^{(ij)} + \beta_1^{(ij)} sex + \beta_2^{(ij)} ACCI + \beta_3^{(ij)} ageD\right), \quad \text{for } i \neq j.$$

For simplicity, we restrict ourselves to a 2-state model, noting that the methodology is generally applicable for any finite number of states.

3.2 Results

The model results regarding the estimated parameters of the observation process and the mark process show that overall, state 1 is characterized

TABLE 1. Parameter estimates for the observation process and the mark process.

parameter		**state 1**	**state 2**
δ_i	(initial state prob.)	0.707	0.293
λ_i	(rate of exponential dist.)	0.123	0.029
μ_i	(mean of gamma dist.)	80.9	124.5
σ_i	(std. deviation of gamma dist.)	85.0	116.9
$\pi_0^{(i)}$	(prob. at zero)	0.726	0.370

by frequent (i.e. roughly weekly) healthcare interactions and high drug use over long periods (cf. Table 1). This could be interpreted as a state of poor health condition or, alternatively, a period in which a treatment needs to be appropriately adjusted to a patient (e.g. right after disease diagnosis) — an interpretation supported by the estimated initial state probabilities, as it is 2.5 times more likely that a person is in state 1 rather than state 2 right after their initial COPD diagnosis. In contrast, state 2 consists of, on average, approximately one healthcare interaction per month with higher DDDs at a single consultation. Because of the low interaction rate, the relative (e.g. per month) drug use here is lower than in state 1. Therefore, we tentatively describe state 1 as the high and state 2 as the low disease level. Importantly, however, the disease states are derived in a data-driven way and as such should not be expected to match disease stages postulated in the literature.

Based on the estimated coefficients in the state process, we can calculate the transition intensity matrix $\boldsymbol{Q}$ for different covariate values, from which in turn we derive the expected durations within each state presented in Table 2. The results show that the reference group of female patients with moderate comorbidities and age 49 at initial diagnosis is expected to spend roughly 48 days in state 1 compared to 94 days in state 2. In contrast, male patients with otherwise the same characteristics spend considerably more time in the low disease level (i.e. 81 days more in state 2, on average), while

TABLE 2. Expected duration (in days) in each state with 95% confidence intervals, which were obtained using Monte Carlo simulation. The reference group are female patients with moderate ACCI and age 49 at initial diagnosis.

group	**state 1**		**state 2**	
reference group	48.2	[45; 51.7]	94.2	[88.3; 100.3]
male	49.1	[44.7; 54.3]	175.1	[158.4; 194.4]
mild ACCI	53.5	[47.3; 60.6]	148.3	[131.7; 167.8]
min. age: 21	61.3	[52.0; 72.7]	77.0	[65.9; 90.1]
max. age: 69	40.7	[36.2; 45.7]	108.4	[96.9; 121.4]

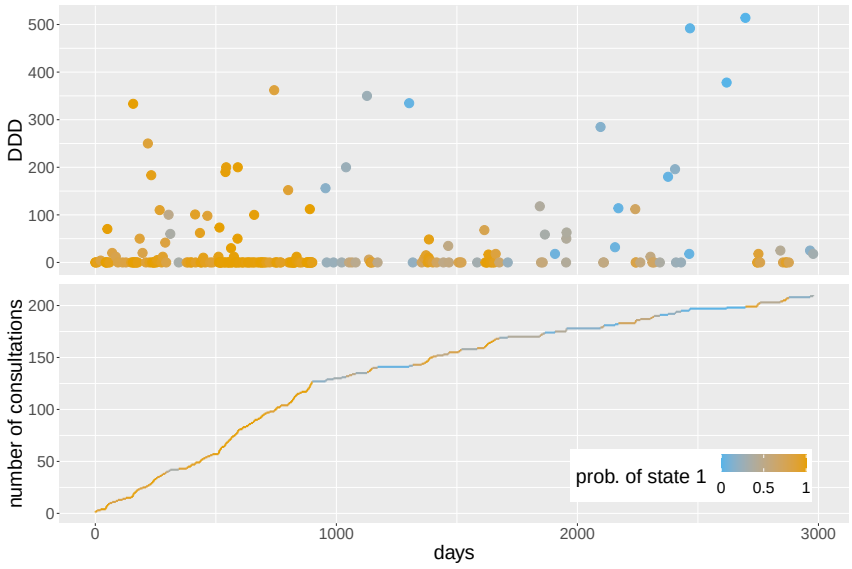


FIGURE 1. Example DDD sequence (upper plot) and step function of the number of physician consultations over time (lower plot) for one patient, coloured according to the local state probabilities of the high disease level (i.e. state 1).

a similar pattern is found for mild in contrast to moderate comorbidities. Concerning age at diagnosis, younger persons spend both more time in state 1 and less time in state 2 compared to the reference group, which is reversed in older age. This effect, however, is caused by a selection bias due to our use of the age-adjusted comorbidity index.

Based on the fitted MMMPP, we can decode the patients' latent state sequences, which provide insight into the individual course of a disease. The local state probabilities, which are calculated using the forward-backward algorithm, allow us to further quantify uncertainty in the decoded state sequences (see Figure 1). In particular, uncertainty in state allocation is to be expected, as the state-dependent distributions of the observation process and the mark process overlap substantially (cf. Table 1), reflecting that disease dynamics within the data are less distinct. Nevertheless, the model appears to adequately distinguish (qualitatively) different periods of healthcare utilisation and drug use, while taking into account the temporal dependence structure of the data.

4 Discussion

As illustrated in the case study, MMMPPs can detect distinct patterns of healthcare utilisation related to disease processes and reveal inter-individual

differences in the state-switching dynamics. In particular, their flexible model structure offers manifold possibilities to analyse claims data; MMMPPs not only operate in continuous time and allow for (potentially multivariate) observations consisting of various data types but can also include covariate effects on the disease dynamics, the event rates, or the mark distributions. By jointly modelling observations and their informative time points, the continuous-time latent-state approach of MMMPPs thus offers a natural framework to analyse the evolution of patients' disease activity underlying claims data.

References

Alaa, A. M., Hu, S. and van der Schaar, M. (2017). Learning from clinical judgments: semi-Markov-modulated marked Hawkes processes for risk prognosis. In: *Proceedings of the 34th International Conference on Machine Learning*, **70**, Precup, D. and Teh, Y. W. (Eds), pp. 60 – 69.

Lange, J. M., Hubbard, R. A., Inoue, L. Y. T. and Minin, V. N. (2015). A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics*, **71**, 90 – 101.

Lu, S. (2012). Markov modulated Poisson process associated with state-dependent marks and its applications to the deep earthquakes. *Annals of the Institute of Statistical Mathematics*, **64**, 87 – 106.

Pullenayegum, E. M. and Lim, L. S. (2016). Longitudinal data subject to irregular observation: a review of methods with a focus on visit processes, assumptions, and study design. *Statistical Methods in Medical Research*, **25**, 2992 – 3014.

Estimating what is under the tip of gender-based violence: A statistical modelling approach

Isabel Millán[1], Amanda Fernández-Fontelo[2], Pere Puig[2], David Moriña[1]

[1] Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, Universitat de Barcelona, Spain
[2] Department of Mathematics, Universitat Autònoma de Barcelona, Spain

E-mail for correspondence: ana.isabel.millan@estudiantat.upc.edu

Abstract: One in three women worldwide experience physical or sexual violence, mostly by an intimate partner. Despite the large number of complaints, it is suspected that they represent only a small portion of the cases occurring, as for very different reasons, the victims often decide not to report the assaults they suffer. In this work, the number of weekly diagnoses related to gender-based violence registered in the Primary Care system in one of the largest areas in Catalonia (Spain) is analysed, estimating its underreporting and considering the different behavior of the phenomenon due to the Covid-19 mitigation measures undertaken by the Spanish government and the impact of a training activity to sensibilize practitioners conducted in late 2019 in the considered area. The proposed methodology is also capable of reconstructing the most likely evolution of the process.

Keywords: gender-based violence; count data; underreporting.

1 Introduction

According to the United Nations, gender-based violence (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 the victim can require medical assistance after an episode of GBV or not. 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.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

Despite the large number of complaints, it is suspected that they represent only a small portion of the cases occurring, as for very different reasons, the victims often decide not to report the assaults they suffer. As reported in the document on tackling sexual violence in Catalonia (Toledo-Vásquez and Pineda-Lorenzo, 2016), being a victim of sexual violence is often fraught with guilt that can lead to denial of sexual violence. Therefore, it is reasonable to think that the number of diagnoses related to gender-based violence recorded in the public health system databases may be underestimating the magnitude of the problem (Fernández-Fontelo et al., 2019).

2 Methods

2.1 Data

The proposed approach is used to reconstruct the most likely evolution of the weekly number of confirmed diagnoses of GBV from January 2010 to December 2021 in the North Metropolitan Health area (Catalonia, Spain). This area is divided in 6 subareas (25, 26, 27, 31, 34 and 35), each one with a particular behavior.

It is known that the outbreak of the Covid-19 pandemic and the measures undertaken by governments to deal with it (as mandatory home confinements) resulted in an increase in the cases of GBV in many countries, also in Spain (Rodriguez-Jimenez et al., 2021). Additionally, by late 2019 a training activity was carried out by the catalan Department of Health in order to sensibilize practitioners with the issue, which is expected to reduce the underreporting of cases.

In Spain, the official statistic to measure the prevalence of GBV is the violence against women Macro Survey (done every 4 years). The 2019 Macro Survey's main objective was to find out the percentage of women aged 16 or over residing in Spain who have suffered GBV (Delegación del Gobierno contra la Violencia de Género, 2019). This type of surveys give us a more realistic idea of the number of GBV in Spain whether they have been reported or not. Given that the interview is anonymous, the social stigma and lack of access to resources and support systems are not an obstacle to have a better aproximation of the actual magnitude of the issue.

2.2 Model

Let's assume that the actual weekly number of GBV cases X_t follows a Poisson distribution with mean λ, which is increased in a factor β in the mandatory confinement period (2020 March 14th to 2020 June 24th), i.e., $E(X_t) = \lambda + I(t) \cdot \beta$ where $I(t)$ takes the value 1 if t falls within the mandatory confinement period and 0 otherwise. The evolution of the phenomenon in each subarea is shown in Figure 1, jointly with the reconstructed most likely actual process according to Equation 2.

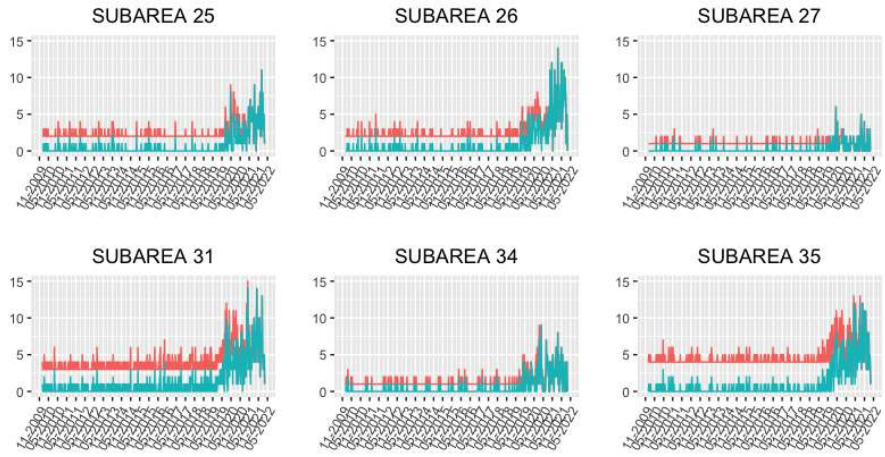


FIGURE 1. Evolution of the weekly number of gender-based violence diagnoses in each subarea of the North Metropolitan Health area (Catalonia, Spain) (in green) and reconstructed most likely process following Eq 2 (in red).

The number of cases diagnosed within the public primary care system, Y_t, is just a part of the actual process, expressed as

$$Y_t = \begin{cases} q_0 \circ X_t, t \leq t' \\ q_t \circ X_t, t > t' \end{cases} \tag{1}$$

where $\circ$ is the *binomial thinning* operator, defined as $q_t \circ X_t = \sum_{i=1}^{X_t} Z_i$, with Z_i independent and identically distributed Bernoulli random variables with probability of success q_t and $q_t = q_0 + \frac{t-t'}{\frac{\alpha-t'}{(1-q_0)}}$ for $t > t'$, where t' is the changing point at which the sensibilization training for primary care professionals starts impacting the weekly number of diagnoses.
It should also be noted that α is the moment when $q_\alpha = 1$, i.e., the registered and observed processes coincide. It is important to note that the number of GBV cases X_t is not directly observed, and only the number of diagnosed cases Y_t is observed. Model (1) assumes that Y_t only reports a fraction q_t of the total number of GBV cases. All the parameters (q_0, λ, β, α and t') are estimated by Gibbs sampling using the *R2jags* package (Yu-Sung and Masanao, 2021), using appropriate priors based on the available information. In order to avoid non-identifyability of the model (1), the actual average number of GBV cases in each subarea on the non-Covid period (parameter λ) has a normal prior distribution with mean based on the expected cases according to the Macro Survey results provided by the Spanish Minsitry of Equality. It is worth noticing that this is a conservative approach, as we are assuming that the results of the survey are not

underestimating the prevalence of GBV cases.
Once the parameters have been estimated, the most likely process can be reconstructed taking into account that $Y_i \mid X_i \sim Binom(x_i, q_t)$. At each time t with j reported cases, the most likely number of gender-based violence cases is the value ν that maximizes the probability

$$f(\nu) = P(X = \nu \mid Y = j) \propto P(Y = j \mid X = \nu) \cdot P(X = \nu) = \\ = \begin{cases} 0, j > \nu \\ \binom{\nu}{j} \cdot q_t^j \cdot (1 - q_t)^{\nu - j} \cdot \frac{e^{-(\lambda + I(t)\cdot\beta)} \cdot (\lambda + I(t)\cdot\beta)^\nu}{\nu!}, j \leq \nu \end{cases} \tag{2}$$

A thorough simulation study reproducing the described structure with different parameter values has been conducted in order to assess whether the original values can be recovered by using this estimation method and to assess the model performance. Preliminary results of this simulation study are summarized in Section 3.2.

3 Results

3.1 Catalonia Primary Health Care System

Table 1 summarizes the parameter estimates in each subarea. It can be seen that the underreporting at the beggining of the period is severe (q_0 ranging from 0.05 to 0.14), and that the impact of the training for professionals contributes in reducing the underreporting from early dates, so the actual number of cases is being registered between 2022-03-04 (subarea 25 and 26) and 2022-05-13 (subarea 27).

TABLE 1. Parameter estimates (posterior median and 95% credible interval) for each subarea.

	q_0	λ	β	α	t'
25	0.08 (0.07, 0.1)	2.31 (2.19, 2.55)	3.34 (1.34, 5.9)	636 (636, 640)	520 (517, 522)
26	0.09 (0.07, 0.11)	2.57 (2.39, 2.74)	3.73 (1.81, 6.15)	636 (636, 639)	503 (499, 506)
27	0.09 (0.06, 0.12)	1.19 (1.03, 1.36)	1.44 (0.4, 3.07)	643 (636, 671)	507 (482, 516)
31	0.14 (0.11, 0.16)	4.27 (4.08, 4.45)	4.12 (1.91, 6.78)	637 (636, 640)	500 (494, 506)
34	0.12 (0.09, 0.16)	1.47 (1.31, 1.64)	2.95 (1.37, 5.02)	637 (636, 643)	501 (495, 509)
35	0.05 (0.04, 0.06)	5.21 (5.02, 5.39)	2.73 (0.78, 5.45)	637 (636, 640)	503 (495, 510)

In all cases we have run 5 MCMC chains and 50,000 iterations, and convergence is reached for all parameters while not having any patterns in the trace plots. Figure 2 correspond to subarea 27 for illustration, but the behavior is consistent for all subareas.
In addition, the Potential Scale Reduction Factor corresponding to all parameters in all subareas were below 1.01, also indicating acceptable convergence of the chains.

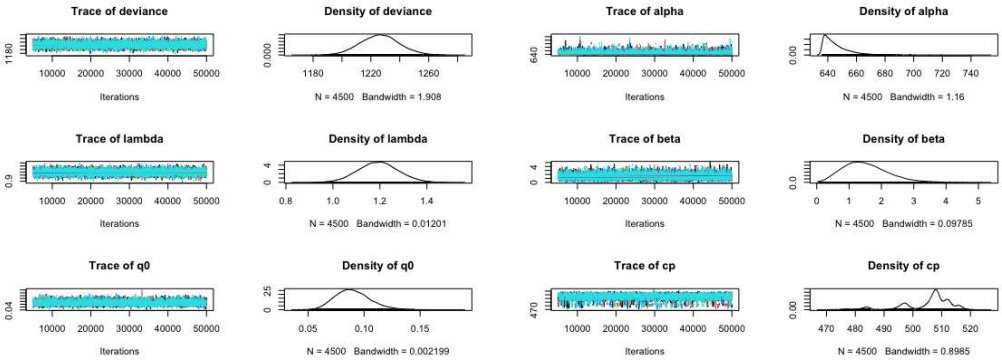


FIGURE 2. MCMC diagnostics: Convergence and posterior density functions for subarea 27.

3.2 Preliminary performance results

In order to assess the performance of the model (1), a simulation study has been conducted. The theoretical values for the parameter q_0 ranged from 0.1 to 0.9; $\alpha = 1200, 1500, 2000$; $\beta = 0.5, 2, 5$; $\lambda = 5, 7, 10$; $t' = 100, 500, 900$. For each parameters combination, 100 random samples of size n=1000 have been generated. Average relative bias, average interval length (AIL) and average 95% credible interval coverage are shown in Table 2. To summarize model robustness, these values are averaged over all combinations of parameters, considering their prior distribution is a Dirac's delta with all probability concentrated in the corresponding parameter value. Due to computational burden, only partial results can be provided here.

TABLE 2. Model performance measures (average relative bias, average interval length (AIL) and average coverage) summary based on a simulation study

Parameter	Bias (%)	AIL	Coverage (%)
α	4.14	1656	86.2
β	0.006	3.95	96.6
t'	0.61	436	98.1
λ	0.001	0.38	87.7
q_0	0.0001	0.07	90.8

As shown in Table 2, the model behaves as expected. In general, 95% credible intervals coverage is reasonable, even with relatively thin intervals and low relative bias.

Acknowledgments: This research has been funded by the Social Obser-

vatory of the "la Caixa" Foundation as part of the project LCF/PR/SR22/52570005. A.F-F acknowledges Agencia Estatal de Investigación for the financial support IJC2020-045188I/AEI/10.13039/501100011033 and María Zambrano scholarship.

References

Delegación del Gobierno contra la Violencia de Género (2019). Macroencuesta de violencia contra la mujer 2019. https://violenciagenero.igualdad.gob.es/

Fernández-Fontelo, A., Cabaña, A., Joe, H., Puig, P., Moriña, D. (2019). Untangling serially dependent underreported count data for gender-based violence. *Statistics in Medicine*, **22**, 4404 – 4422.

Rodriguez-Jimenez, R., Fares-Otero, N. E., García-Fernández, L. (2021). Gender-based violence during COVID-19 outbreak in Spain. *Psychological Medicine*, **7**, 1 – 2.

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.*

Yu-Sung, S., Masanao, Y. (2021). R2jags: Using R to Run 'JAGS'. *R package version 0.7-1.*

A bivariate Poisson regression model for radiation dose estimation

Dorota Młynarczyk[1], Pedro Puig[1,2], Carmen Armero[3], Virgilio Gómez-Rubio[4], Jayne Moquet[5]

[1] Universitat Autònoma de Barcelona, Barcelona, Spain
[2] Centre de Recerca Matemàtica, Barcelona, Spain
[3] Universitat de València, València, Spain
[4] Universidad de Castilla-La Mancha, Albacete, Spain
[5] UK Health Security Agency, Didcot, UK

E-mail for correspondence: dorotaanna.mlynarczyk@uab.cat

Abstract: Radiation dose estimation is a topic that is constantly looking for new methodological and applied solutions. Accurate dose estimation is essential for minimizing the health risks associated with exposure to ionizing radiation. Over the years, various biomarkers have been used for radiation dose estimation, including chromosome aberrations in peripheral blood lymphocytes. Instead of using only one blood cell-based biomarker, we suggest a novel approach that makes use of two biomarkers: dicentrics and chromosomal translocations. One of the statistical methods that enables us to fit two correlated count variables, as is the case, is the bivariate Poisson regression model. By combining these two types of chromosomal aberrations in a bivariate model, we can potentially overcome the limitations of each biomarker and improve the accuracy of radiation dose estimation.

Keywords: biological dosimetry; chromosomal translocations; dicentrics.

1 Introduction

Nowadays, nuclear technology is widely employed around the world, particularly in the fields of industry, medicine, and energy production. Unintentional radiation exposure unfortunately occurs sometimes despite rigorous restrictions and safety precautions. It is necessary to assess the radiation dose that the exposed person has absorbed in order to give them the best and most immediate medical care. This need will be much more urgent in the event of a large-scale disaster involving hundreds of individuals. Then

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

the key would be to identify persons who need immediate care, i.e. the so-called triage. Estimating the radiation dose is the aim of the field known as biological dosimetry.
Biodosimetry is based on the quantification of the magnitude of radiation damage at the cellular level, that is chromosomal aberrations, which frequency is used in statistical models to estimate a radiation dose (of an exposed person). Dicentrics, chromosomes with two centromeres, are the most studied chromosomal abnormalities. Translocations, which involve the swapping of two chromosomal fragments, are a labor-intensive method that is rarely performed. However, dicentrics are also observed when translocations are detected by fluorescence in situ hybridization (FISH) assay, hence we provide mathematical models that enable their simultaneous usage.
In biodosimetry, linear and quadratic models of radiation dose are usually investigated. For paired count data relating to dicentrics and translocation information, bivariate Poisson regression models may be a good choice. Examples of applications for these types of regression models include the analysis of healthcare data, insurance ratemaking, or football games (Karlis & Ntzoufras, 2003). Although, as far as we know, no studies have employed a bivariate model to estimate radiation doses in biodosimetry.

2 Model

Let denote by X_j and Y_j the number of dicentrics and translocations in j-th cell, $j = 1, \ldots, M$. Then assume that X_j and Y_j follow jointly a bivariate Poisson distribution, $BP(\lambda_{1j}, \lambda_{2j}, \lambda_{3j})$, with probability function

$$P_{(X_j,Y_j)}(x_j, y_j) = \exp(-(\lambda_{1j} + \lambda_{2j} + \lambda_{3j}))\frac{\lambda_{1j}^{x_j}}{x_j!}\frac{\lambda_{2j}^{y_j}}{y_j!} \times \sum_{k=0}^{\min(x_j,y_j)} \binom{x_j}{k}\binom{y_j}{k} k!(\frac{\lambda_{3j}}{\lambda_{1j}\lambda_{2j}})^k, \quad x_j, y_j = 0, 1, 2, \ldots,$$

where $\lambda_{ij} > 0$, $i = 1, 2, 3$, $E(X_j) = \lambda_{1j} + \lambda_{3j}$ and $E(Y_j) = \lambda_{2j} + \lambda_{3j}$. Then $\lambda_{1j}, \lambda_{2j}, \lambda_{3j}$ can be modelled using some regressors. We will examine two models, linear and quadratic, that include the radiation dose as a covariate, which are typical in the field of biodosimetry (see IAEA, 2011 and Młynarczyk et al., 2022). In the case of a quadratic model we have

$$\begin{aligned} \lambda_{1j} &= \beta_{11} + \beta_{12} \cdot dose_j + \beta_{13} \cdot dose_j^2, \\ \lambda_{2j} &= \beta_{21} + \beta_{22} \cdot dose_j + \beta_{23} \cdot dose_j^2, \\ \lambda_{3j} &= \beta_{31}, \end{aligned} \tag{1}$$

where $dose_j$ denotes the radiation dose received by j-th cell, β_{kq} denotes the corresponding regression coefficients, for $k = 1, 2, q = 1, 2, 3$. Moreover,

λ_{3j} is the covariance between the two random variables X_j and Y_j. The model is completely linear when β_{13} and β_{23} are fixed to zero.
Typically, these types of models are used to estimate the radiation dose of a potentially exposed person. Assuming that we are now analyzing a new blood sample from an irradiated patient, this implies that there are new observations of X_j dicentrics and Y_j translocations in cells $j = M + 1, \ldots, N$. In contrast to the data for $j = 1, \ldots, M$ for which the doses are known, the dose received by the patient is unknown and will be denoted by $\boldsymbol{D}$. The main goal is to estimate the dose based on the observed data. Since we are working within Bayesian framework, the interest is in finding the posterior distribution for the model

$$\pi(\boldsymbol{\beta}, \boldsymbol{D} \mid (\mathbf{x}, \mathbf{y})) \propto \pi((\mathbf{x}, \mathbf{y}) \mid \boldsymbol{\beta}, \boldsymbol{D})\pi(\boldsymbol{D}, \boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{31})$ and $\mathbf{x} = (x_1, \ldots, x_N)$, $\mathbf{y} = (x_1, \ldots, x_N)$. The likelihood is given by

$$\pi((\mathbf{x}, \mathbf{y}) \mid \boldsymbol{\beta}, \boldsymbol{D}) = \prod_{j=1}^{N} \exp(-(\lambda_{1j} + \lambda_{2j} + \lambda_{3j}))\frac{\lambda_{1j}^{x_j}}{x_j!}\frac{\lambda_{2j}^{y_j}}{y_j!} \times \sum_{k=0}^{\min(x_j, y_j)} \binom{x_j}{k}\binom{y_j}{k} k!\left(\frac{\lambda_{3j}}{\lambda_{1j}\lambda_{2j}}\right)^k,$$

where λ_{ij} are defined as in (1) for $j = 1, \ldots, N$.
We need to specify a prior distribution $\pi(\boldsymbol{D}, \boldsymbol{\beta})$. We assume a prior independence between dose and regression coefficients so $\pi(\boldsymbol{D}, \boldsymbol{\beta}) = \pi(\boldsymbol{D})\pi(\boldsymbol{\beta})$. For $\pi(\boldsymbol{\beta})$ we opt for non-informative uniform priors on a positive interval because that λ_{ij} is positive. Prior for dose, $\pi(\boldsymbol{D})$, should be chosen with some knowledge of the radiation accident and the patient's symptoms. If such information is not available, it is possible to consider a uniform distribution on the interval corresponding to the range of doses in the dataset, as was done in this study.
Marcov chain Monte Carlo (MCMC) methods can be used to approximate the posterior distribution. In particular, using Gibbs sampling method, the samples from $\pi(\boldsymbol{\beta}, \boldsymbol{D} \mid (\mathbf{x}, \mathbf{y}))$ are constructed from the conditional posterior distribution of each element in $(\boldsymbol{D}, \boldsymbol{\beta})$ given the rest of them. The full conditional distributions can be clearly defined in this case. For instance, the conditional distribution for β_{12} is given by

$$\pi(\beta_{12} \mid \beta_{-12}, \boldsymbol{D}, (\mathbf{x}, \mathbf{y})) \propto \pi(\beta_{12}) \cdot \prod_{j=1}^{N} \exp(-\lambda_{1j})\lambda_{1j}^{x_j} \times \sum_{k=0}^{min(x_j, y_j)} \frac{1}{k!(x_j - k)!(y_j - k)!}\left(\frac{\lambda_{3j}}{\lambda_{1j}\lambda_{2j}}\right)^k,$$

where $\beta_{-12} = (\beta_{11}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{31})$, and $\pi(\beta_{12})$ is the prior distribution for β_{12}. Identically, other conditional distributions can be defined for the rest of the $\boldsymbol{\beta}$ parameters. Since the primary objective is to estimate the dose $\boldsymbol{D}$, the posterior marginal distribution $\pi(\boldsymbol{D} \mid (\mathbf{x}, \mathbf{y}))$ is the main interest. It can be obtained by integrating the posterior density over $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{D} \mid (\mathbf{x}, \mathbf{y})) = \int \pi(\boldsymbol{\beta}, \boldsymbol{D} \mid (\mathbf{x}, \mathbf{y}))d\boldsymbol{\beta}.$$

The most likely range of dose received by the patient, can then be determined looking at the 95% credible interval.

3 Data & Results

In this study we apply the bivariate Poisson model to data from Finnon et al. (1999). The blood sample, from a healthy 31-year-old male, was exposed to γ-rays in a laboratory at radiation doses ranging from 0.25 to 4 Gy. Then the samples were painted according to the FISH protocol, which allows the detection of bicoloured dicentrics and translocations in chromosomes. The number of scored cells varies for different doses from 200 to 2000. For lower doses the sample mean of the number of dicentrics and of tranlocations is much lower (0.01 for dose 0.25 Gy for both aberrations types) than for dose 4 Gy (0.27 for dicentrics and 0.37 for translocations). Table 1 displays details regarding the data.

TABLE 1. Number of cells, mean of dicentrics and mean of translocations for different doses. The dose 2 Gy was chosen as test data.

Dose (Gy)	0.25	0.50	0.75	1.00	2.00	3.00	4.00
Number of cells	2000	1000	1000	1001	500	299	200
Mean of dicentrics	0.01	0.01	0.02	0.02	0.10	0.19	0.27
Mean of translocations	0.01	0.01	0.02	0.02	0.11	0.22	0.37

The frequency of aberrations for dose 2 Gy was chosen as test data to check the performance of the model (i.e. it was irradiated in the same conditions as calibration data, but it was not included in the calibration part of the model). The mean number of dicentrics found in this sample was 0.1 and translocations 0.106.

Figure 1 provides a graphic representation of the result for both estimate linear and quadratic models. The curves have been constructed using the mean values of the posterior distribution of the parameters β_{kq} for $k = 1, 2, q = 1, 2, 3$. The marginal posterior dose density of both models can be seen in the Figure 2. The quadratic model estimates the mean value of dose 2.186 Gy, and the 95% credible interval is between 2.088 Gy and 2.28 Gy.

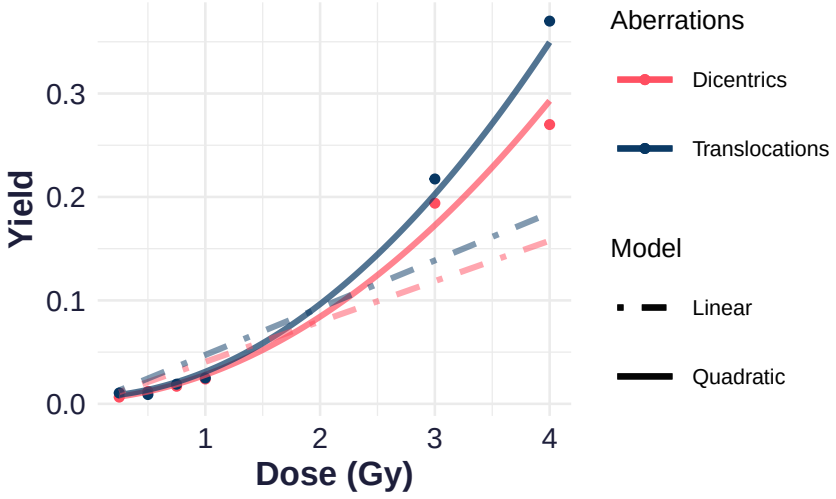


FIGURE 1. Result of the linear (dotdash line) and quadratic (solid line) model. The red lines represent the curves for dicentrics and the blue ones for translocations with regard the dosis received. The points are the actual values of yield of dicentrics and translocations that were used to fit the model.

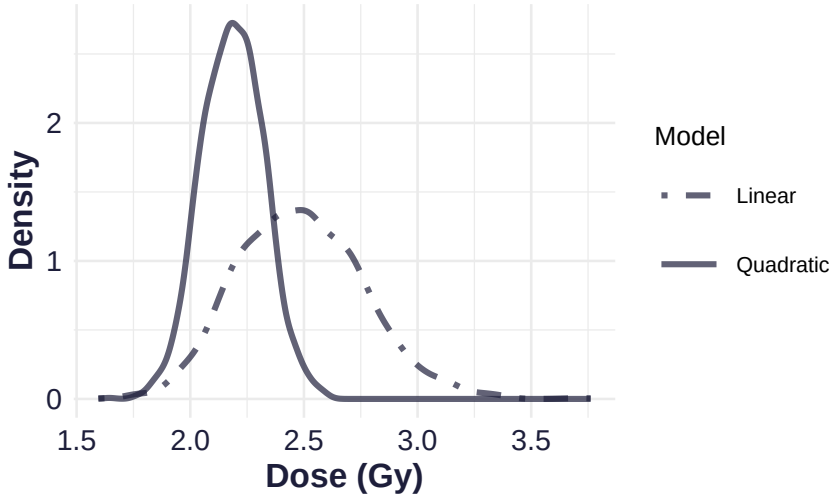


FIGURE 2. Marginal posterior dose density given by linear and quadratic model.

On the other hand, the linear model produces a mean estimate of 2.486 Gy, with the 95% credible interval of 2.286 Gy to 2.674 Gy. Thus the quadratic model gives better estimation, although it overestimates the real dose. This

is most likely a result of the limited calibration data that are available; the more calibration data used, the more accurate the estimation could be. The advantage of the proposed model is that it may be simply expanded to add other covariates (such as age, sex, etc.) that may affect the frequency of dicentrics or translocations. The model may also be employed in scenarios involving partial body exposure, but further research is needed.

Acknowledgments: Special Thanks to Cytogenetics Group from UK Health Security Agency for explaining all biological details of the assays.

References

Finnon, P., Moquet, J.E., Edwards, A.A., and Lloyd, D.C. (1999). The 60Co gamma ray dose-response for chromosomal aberrations in human lymphocytes analysed by FISH; applicability to biological dosimetry. *International journal of radiation biology*, **75**, 1215 – 1222.

International Atomic Energy Agency (2011). Cytogenetic Dosimetry: Applications in Preparedness for and Response to Radiation Emergencies. In: *Emergency Preparedness and Response*, IAEA, Vienna.

Karlis, D., and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **52**, 381 – 393.

Młynarczyk, D., Puig, P., Armero, C., Gómez-Rubio, V., Barquinero, J. F., and Pujol-Canadell, M.(2022). Radiation dose estimation with time-since-exposure uncertainty using the γ-H2AX biomarker. *Scientific reports*, **12**(1), 19877.

Bayesian spatio-temporal conditional overdispersion models proposals

Mabel Morales-Otero[1], Vicente Núñez-Antón[2]

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

E-mail for correspondence: vicente.nunezanton@ehu.eus

Abstract: In this work, we propose a direct spatio-temporal extension of the spatial conditional overdispersion models, where we include the spatial lag of the response variable for each time unit in the linear predictor. The proposed models are able to capture both spatial and temporal correlations that may be present in the data under study. In addition, we also propose temporally varying spatial lag coefficient models, which allow us to study the variation in time of the spatial term. In order to illustrate their performance, we apply our proposals, for Poisson distributed responses, to the Glasgow respiratory hospital admissions data set, where we compare their performance with the widely used Knorr-Held's models.

Keywords: Bayesian models; Overdispersion; Spatio-temporal models.

1 Introduction

Spatio-temporal data arise in many fields of study, since researchers are often interested in studying a phenomenon observed in several locations and time periods. This type of data often exhibit correlation among regions and time units that need to be taken into account when fitting regression models. The spatial conditional overdispersion models were proposed by Cepeda-Cuervo et al. (2018) to fit spatial count data, as they are able to take overdispersion into account, and also model the possible existing spatial dependence by including the spatial lag of the response variable under study in the regression structure for the mean (see also Morales-Otero and Núñez-Antón, 2021). In this work, we propose some extensions of these models to allow for the modelling of spatio-temporal count data.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Methods

One of the most frequently used models for fitting spatio-temporal data was proposed by Knorr-Held (2000). In this model, it is assumed that the random variables Y_{ij} represent counts for $i = 1, \ldots, n$ regions in $j = 1, \ldots, J$ time periods. It is also assumed that Y_{ij}, conditioned on the random effects $\nu_i, \eta_i, \delta_j, \phi_j$ and ϵ_{ij}, follows a Poisson distribution with means μ_{ij}; that is $(Y_{ij}|\nu_i, \delta_j, \phi_j, \epsilon_{ij}) \sim \text{Poi}(\mu_{ij})$, with regression structure given by:

$$\log(\mu_{ij}) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \nu_i + \eta_i + \delta_j + \phi_j + \epsilon_{ij}, \tag{1}$$

where $\mathbf{x}_{ij}$ is the $k \times 1$ vector of explanatory variables for the i-th area in the j-th time period, and $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown regression parameters that needs to be estimated. In addition, ν_i and δ_j are unstructured random effects for space and time (i.e., $\nu_i \sim N(0, \tau_\nu)$, $\tau_\nu > 0$, and $\delta_j \sim N(0, \tau_\delta)$, $\tau_\delta > 0$), respectively, η_i is a spatially structured random effect following an intrinsic conditionally autoregressive (ICAR) distribution and ϕ_j is a temporal effect, following either a random walk or an autoregressive process. Finally, ϵ_{ij} is a spatio-temporal interaction term for which an unstructured normal prior distribution (i.e. $\epsilon_{ij} \sim N(0, \tau_\epsilon)$, $\tau_\epsilon > 0$) is often assumed.
We propose an extension of the spatial conditional models that includes, for each time period, the lag term of the response variable under study. The parameter associated to this term would represent the strength of the global spatial autocorrelation that can be present in the data. In this sense, positive significant values would suggest positive spatial autocorrelation in the whole time period considered, and negative significant values, negative spatial autocorrelation. In particular, we assume that, for each time period j, the response variables Y_{ij}, conditioned on the values of all the neighbours of the i-th region, but not including the i-th region itself (i.e., $Y_{\sim ij}$), and on the random effects ν_i, δ_j, ϕ_j and ϵ_{ij}, follow a Poisson distribution; that is $(Y_{ij}|Y_{\sim ij}, \nu_i, \delta_j, \phi_j, \epsilon_{ij}) \sim \text{Poi}(\mu_{ij})$. Here, the conditional means μ_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$, follow the regression structure:

$$\log(\mu_{ij}) = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \rho\mathbf{W}_i\mathbf{y}_j + \nu_i + \delta_j + \phi_j + \epsilon_{ij}, \tag{2}$$

where $\mathbf{W}_i$ is the i-th row of the $n \times n$ spatial weights matrix $\mathbf{W}$ modelling the spatial dependence, $\mathbf{y}_j$ is the $n \times 1$ vector of observations for all n spatial units for time period j, ρ is the parameter that captures the strength of the spatial association, and $\mathbf{x}_{ij}$, $\boldsymbol{\beta}$, ν_i, δ_j, ϕ_j and ϵ_{ij} are as before.
Finally, we also propose the temporally varying spatial lag coefficient model where, in equation (2), we assume that the coefficient for the spatial lag is the sum of a fixed coefficient, ρ_0, and a random coefficient, ρ_j, that varies according to the time units $j = 1, \ldots, J$. More specifically, we propose different specifications for the temporal varying coefficient, such as an unstructured normal distribution, a random walk process or an autoregressive process. The estimated value obtained for ρ_0 would represent the strength

of the spatial dependence among the regions for the whole time period under study, whereas the estimated values obtained for ρ_j would indicate whether the spatial association increases or decreases with time. This would allow us to examine the variability of the coefficient of the spatial lag from one time unit with regard to the others. Taking into account the value obtained for ρ_0, a positive estimated value of ρ_j would suggest that for time period j, the strength of the spatial association is larger than that indicated by ρ_0, and a negative estimated value of ρ_j would indicate that, for the j-th time period, the spatial autocorrelation is weaker. If $\hat{\rho_j} \approx 0$, this would mean that there are no significant changes in the spatial correlation pattern for the j-th time period, with respect to that of ρ_0.

3 Application

We study a data set to assess the impact of air pollution on the respiratory health of the population living in each of the $n = 271$ regions or statistical sectors belonging to the Scotland National Health System's board of Greater Glasgow and the Clyde Valley, Scotland, for a time period of $J = 5$ years (i.e., from 2007 to 2011) (Lee et al., 2018). The variables available for each region and time period are the observed number of respiratory hospital admissions (i.e., variable Y), the expected number of respiratory hospital admissions (i.e., variable E), the yearly average modelled concentrations of particulate matter less than 10 microns (i.e., variable PM10), the average property price in each region and year (i.e., variable Price), and the proportion of the working age population who are in receipt of job seekers allowance (i.e., variable JSA). In addition, we can obtain the standardized incidence ratio (SIR) for each region and time period so that $\text{SIR}_{ij} = Y_{ij}/E_{ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$.
We have fitted the spatio-temporal model proposed by Knorr-Held (2000) in equation (1), where we assume that $(Y_{ij}|\nu_i, \eta_i, \delta_j, \phi_j, \epsilon_{ij}) \sim \text{Poi}(\mu_{ij})$, with means $\mu_{ij} = E_{ij}\theta_{ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$, following the regression structure:

$$\begin{aligned}\log(\mu_{ij}) = &\log(E_{ij}) + \beta_0 + \beta_1 \text{JSA}_{ij} + \beta_2 \text{Price}_{ij} + \beta_3 \text{PM10}_{ij} \\ &+ \nu_i + \eta_i + \delta_j + \phi_j + \epsilon_{ij}\end{aligned} \tag{3}$$

In addition, we have also fitted our proposed spatio-temporal conditional model in equation (2), where we include the spatial lags of the SIR's in the regression structure for the conditional means. Therefore, we assume that $(Y_{ij}|Y_{\sim ij}, \nu_i, \delta_j, \phi_j, \epsilon_{ij}) \sim \text{Poi}(\mu_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$, with means $\mu_{ij} = E_{ij}\theta_{ij}$ following the regression structure:

$$\begin{aligned}\log(\mu_{ij}) = &\log(E_{ij}) + \beta_0 + \beta_1 \text{JSA}_{ij} + \beta_2 \text{Price}_{ij} + \beta_3 \text{PM10}_{ij} \\ &+ \rho \mathbf{W}_i \mathbf{SIR}_j + \nu_i + \delta_j + \phi_j + \epsilon_{ij},\end{aligned} \tag{4}$$

where $\mathbf{W}_i$ is the i-th row of the spatial weights matrix and $\mathbf{SIR}_j$ is the vector of observations of the SIR's for the j-th time period. In addition, the random effects ν_i, δ_j, ϕ_j and ϵ_{ij} are as before. Results obtained after fitting the models in equations (3) and (4) to the respiratory hospital admissions in Glasgow data are reported in Table 1.

The variables PM10, JSA and Price are statistically significant in both models, results that are consistent with the ones obtained by Lee et al. (2018). In addition, for the spatio-temporal conditional model, the spatial term is statistically significant and its coefficient has a positive value, indicating the presence of positive spatial autocorrelation in the data, which is being properly captured by this term. Furthermore, the smallest information criteria values are given for our proposed spatio-temporal conditional model, a fact that suggests, for this specific case, a better fit for this model over the Knorr-Held model.

TABLE 1. Results obtained after fitting the Knorr-Held and the spatio-temporal conditional models to the respiratory hospital admissions in Glasgow data.

	Knorr-Held			Spatio-temporal conditional		
	Mean	SD	CI	Mean	SD	CI
β_0	-0.441	(0.102)	(-0.642,-0.240)	-0.800	(0.083)	(-0.964,-0.636)
PM10	0.019	(0.007)	(0.004,0.033)	0.017	(0.006)	(0.006,0.028)
JSA	0.057	(0.006)	(0.046,0.068)	0.052	(0.005)	(0.042,0.062)
Price	-0.187	(0.023)	(-0.233,-0.141)	-0.182	(0.021)	(-0.222,-0.141)
ρ	-	-	-	0.463	(0.048)	(0.368,0.557)
τ_ν	0.015	(0.004)	(0.008,0.025)	0.025	(0.003)	(0.020,0.031)
τ_η	0.052	(0.018)	(0.025,0.094)	-	-	-
τ_δ	0.003	(0.010)	(-1.48e-04,0.021)	0.002	(0.009)	(2.91e-05,0.015)
τ_ϕ	0.008	(0.009)	(5.72e-04,0.031)	0.007	(0.009)	(6.280e-04,0.029)
τ_ϵ	0.011	(0.001)	(0.009,0.014)	0.011	(0.001)	(0.009,0.013)
	DIC = 10389, WAIC = 10352			DIC = 10373, WAIC = 10343		

We have also fitted the proposed spatio-temporal varying spatial lag coefficient models to these data. In particular, we assume $(Y_{ij}|Y_{\sim ij}, \nu_i, \epsilon_{ij}) \sim \text{Poi}(\mu_{ij})$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. For the means μ_{ij}, with $\mu_{ij} = E_{ij}\theta_{ij}$, we specify the following regression structure:

$$\begin{aligned} \log(\mu_{ij}) = \log(E_{ij}) + \beta_0 + \beta_1 \text{JSA}_{ij} + \beta_2 \text{Price}_{ij} + \beta_3 \text{PM10}_{ij} \\ + (\rho_0 + \rho_j)\mathbf{W}_i\mathbf{SIR}_j + \nu_i + \delta_j + \epsilon_{ij}, \end{aligned} \tag{5}$$

where we assume that $\rho_j \sim N(0, \tau_\rho)$, $\tau_\rho > 0$, $\nu_i \sim N(0, \tau_\nu)$, $\tau_\nu > 0$, $\delta_j \sim N(0, \tau_\delta)$, $\tau_\delta > 0$, $\epsilon_{ij} \sim N(0, \tau_\epsilon)$, $\tau_\epsilon > 0$, and the rest of the terms are as before. We have fitted the reduced versions of this model, obtaining the best fit for the one that does not include the temporal random effect δ_j. This model takes into account the temporal correlation only by means of the

random coefficient ρ_j for the spatial lag. Results obtained after fitting this model to the respiratory hospital admissions in Glasgow data are reported in Table 2. Here, the spatial coefficient is also positive and statistically significant, capturing the positive spatial autocorrelation present in the whole time period. In addition, the red line in Figure 1 represents the estimated mean obtained for ρ_j, according to the year, and the green bands correspond to its 95% credible interval. Here, we can see how the effect of the spatial lag over the response, which is the number of respiratory hospital admissions, changes with time. In particular, in this case the estimated mean of ρ_j has the largest value for the year 2008 and then, it decreases from that year on. This suggests that in this year is where the strongest spatial autocorrelation is found in the data and, that it becomes weaker for the following years. Only for the year 2009, the effect is nearly zero, meaning that in this year, the spatial dependence is well explained by the fixed parameter ρ_0.

TABLE 2. Results obtained after fitting the temporally varying spatial lag coefficient model in equation (5) to the respiratory hospital admissions in Glasgow data.

	Mean	SD	CI
β_0	-0.809	(0.081)	(-0.968,-0.651)
PM10	0.018	(0.006)	(0.007,0.030)
JSA	0.056	(0.005)	(0.045,0.066)
Price	-0.180	(0.021)	(-0.221,-0.139)
ρ_0	0.430	(0.049)	(0.335,0.526)
τ_ρ	0.011	(0.010)	(0.002,0.038)
τ_ν	0.024	(0.003)	(0.019,0.030)
τ_ϵ	0.011	(0.001)	(0.009,0.013)
	DIC = 10371, WAIC = 10342		

4 Conclusions

We have proposed extensions of the spatial conditional overdispersion models for fitting spatio-temporal count data. We have illustrated their usefulness in the study of the respiratory hospital admissions in Glasgow. We have compared our results with those obtained from the fitting of the Knorr-Held model, resulting in a better fit in terms of information criteria and estimates for our proposed models, with the spatial lag coefficient being statistically significant for all the models considered. Moreover, it has a positive value, indicating that this term is properly capturing the positive spatial autocorrelation present in the data. We have also fitted our proposed temporally varying spatial lag coefficient model to these data, which allowed us to examine the temporal variation of the spatial correlation and to identify the

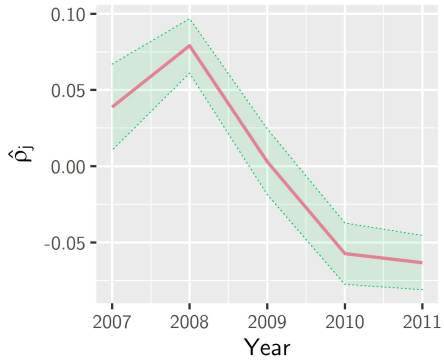


FIGURE 1. Temporal variation of the spatial autoregressive parameter.

year 2008 as the one where the spatial autocorrelation in the data was the strongest.

Acknowledgments: This research has been funded by Ministerio de Ciencia e Innovación (MCIN, Spain), Agencia Estatal de Investigación (AEI/10.13039/501100011033/) and Fondo Europeo de Desarrollo Regional (FEDER) "Una manera de hacer Europa" under the I+D+i research grant PID2020-112951GB-I00 and also by the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group) under research grant IT-1508-22.

References

Cepeda-Cuervo, E., Córdoba, M. and Núñez-Antón, V. (2018). Conditional overdispersed models: Application to count area data. *Statistical Methods in Medical Research*, **27**, 2964 – 2988

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **19**, 2555 – 2567.

Lee, D., Rushworth, A. and Napier, G. (2018). Spatio-temporal areal unit modeling in R with conditional autoregressive priors using the CARBayesST package. *Journal of Statistical Software*, **84**, 1 – 39

Morales-Otero, M. and Núñez-Antón, V. (2021). Comparing Bayesian spatial conditional overdispersion and the Besag-York-Mollié models: Application to infant mortality rates. *Mathematics (Special issue on Spatial Statistics with its Applications)*, **9(3)**, 282

Lasso-based order selection in hidden Markov models: a case study using stock market data

Marius Ötting[1], Roland Langrock[1]

[1] Bielefeld University, Germany

E-mail for correspondence: marius.oetting@uni-bielefeld.de

Abstract: In hidden Markov models (HMMs), the selection of an adequate number of states — also referred to as order selection — is commonly made based on information criteria, despite well-known problems and pitfalls. We explore an alternative approach to order selection in HMMs, considering a penalised likelihood comprising a group lasso penalty on the entries of the transition probability matrix. The feasibility of the approach is demonstrated in a real-data case study on financial share returns, where we compare the predictive performance of the HMMs fitted using the lasso penalty with the common benchmarks.

Keywords: financial time series; group lasso; maximum penalised likelihood; model selection.

1 Introduction

Hidden Markov models (HMMs) are flexible tools for modelling time series driven by underlying states. More specifically, in an HMM, each observation is assumed to be generated by a distribution selected by an underlying latent state. The state process is modelled as a finite-state Markov chain in discrete time, and could for example represent the states of the economy (growth vs. recession; Hamilton, 2008) or the volatility level of a financial market (high vs. low volatility; De Angelis & Paas, 2013).
To select the number of states of an HMM, information criteria such as the AIC or the BIC are often used. However, these are known to often favour models with an unreasonably large number of states, which can be undesirable for example when the states shall be interpretable entities (Pohle et al., 2017). Here we explore an alternative approach using a lasso-type penalty on the state-switching probabilities, thereby reducing the number

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

of states whenever this is supported by the data. In a case study considering daily share returns, we compare the performance of the new approach to the common benchmarks, i.e. order selection based on AIC or BIC.

2 Methods

In an HMM, the observations $y_1, \ldots, y_T$ are driven by an unobserved state process $s_1, \ldots, s_T$, modelled as an N-state Markov chain, in the sense that each y_t is generated by one of N distributions as selected by the state s_t. The state transitions are governed by the transition probability matrix (t.p.m.) $\boldsymbol{\Gamma} = (\gamma_{ij})$, with $\gamma_{ij} = \Pr(s_t = j \mid s_{t-1} = i)$. Assuming the Markov chain starts in its stationary distribution $\boldsymbol{\delta}$, the likelihood of the HMM is

$$\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{\delta}\mathbf{P}(y_1)\boldsymbol{\Gamma}\mathbf{P}(y_2)\ldots\boldsymbol{\Gamma}\mathbf{P}(y_T)\mathbf{1}\,,$$

with $\mathbf{P}(y_t)$ an $N \times N$ diagonal matrix with the state-dependent probabilities (or densities) $f(y_t \mid s_t = i)$, $i = 1, \ldots, N$, on the diagonal, column vector $\mathbf{1} = (1, \ldots, 1)^\intercal \in \mathbb{R}^N$, and with $\boldsymbol{\theta}$ collecting all unknown model parameters (see Zucchini et al., 2016).

The number of states N is typically chosen before model fitting, then comparing several fitted models (e.g. with $N = 2, 3, 4, 5$) via information criteria. However, it is well-known that information criteria in many settings tend to favour models with larger numbers of states than seem adequate given the subject matter (Pohle et al., 2017). This can be explained by the possibility to use extra states for compensating any relevant structure not accounted for in the model formulation (e.g. when using Gaussian state-dependent distributions despite a heavy-tailed empirical distribution, or assuming a first-order Markov chain despite higher-order dependence in the data). Indeed, to date there is no formal approach to order selection in HMMs that would yield reliable results in practice, such that empirical research instead resorts to pragmatism when choosing N.

Without claiming to fill this gap, we here explore a completely novel approach to order selection, which may have some advantages over information criteria. Specifically, we propose a penalised likelihood approach based on the group lasso penalty to shrink the entries of the t.p.m., thereby allowing for an automatic data-driven reduction in the number of states whenever adequate. In particular, we consider the penalised log-likelihood

$$\ell_{\text{pen}}(\boldsymbol{\theta}) = \log\big(\mathcal{L}(\boldsymbol{\theta})\big) - \lambda \sum_{j=2}^{N} \|\boldsymbol{\Gamma}_{\cdot,j}\|_2,$$

with the L_2 norm $\|\cdot\|_2$ and with $\boldsymbol{\Gamma}_{\cdot,j}$ the j-th column of the t.p.m. $\boldsymbol{\Gamma}$. The tuning parameter $\lambda > 0$ governs the amount of penalisation: for $\lambda \to \infty$, the vector $\hat{\boldsymbol{\Gamma}}_{\cdot,j}$ will be zero (Hastie et al., 2015), thus leading to the disappearance of state j. If instead some elements of $\hat{\boldsymbol{\Gamma}}_{\cdot,j}$ are nonzero, then

state j is selected into the model. The first column is not penalised since the row sums of the t.p.m. must be one, such that we cannot shrink all elements to zero. The simplest model thus would be a 1-state HMM, where the first column of the t.p.m. contains only ones and the remaining elements of the t.p.m. are all zero. To select the tuning parameter λ, we consult either the AIC or the BIC, where the number of non-zero parameters is used as a proxy of the effective degrees of freedom.

3 Application

To demonstrate the feasibility of the proposed approach, we consider an application in finance, specifically the prediction of share returns. We consider the daily adjusted closing prices p_t of four stocks from the German stock market index DAX, Bayer, BMW, Deutsche Bank (DB), and Volkswagen (VW), downloaded from `finance.yahoo.com`. We model the time series of daily log-returns, given by $y_t = \log(p_t/p_{t-1})$. Figure 1 shows the time series of DB's daily returns, indicating that in the time window considered, there were some periods with more volatile trading (for example during the European debt crisis in the early 2010s), but also such where the market was relatively calm (e.g. around 2014).

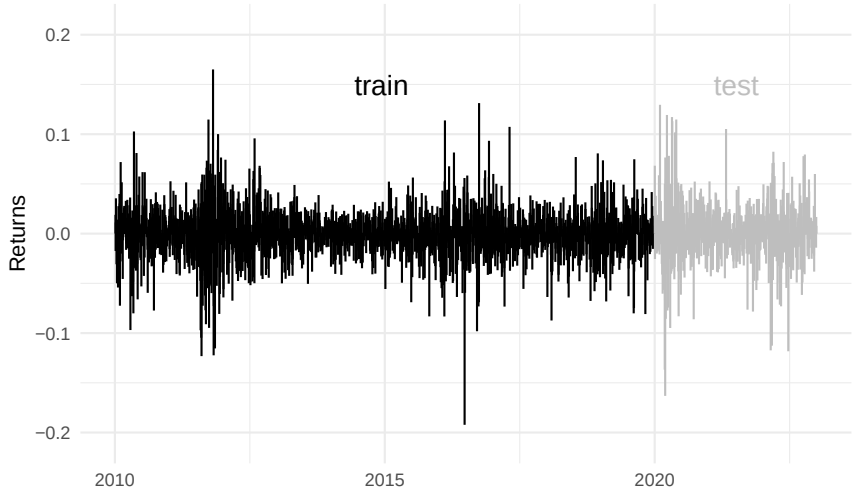


FIGURE 1. Deutsche Bank (log-)returns from January 2010 until December 2022.

We split our data into a training (Jan 2010 – Dec 2019, 2515 observations) and a test set (Jan 2020 – Dec 2022, 760 observations). We fit HMMs with $2-8$ states to the training data, using AIC and BIC to select the

optimal N. These are the benchmarks for the new approach using group lasso penalisation. The different models are compared in terms of their predictive performance on the test set, as measured by the out-of-sample likelihood.
For the four stocks considered, Table 1 lists the number of states preferred by the different approaches and the out-of-sample (log-)likelihood values. The results are inconclusive but indicate that the group lasso could improve the forecast performance (at least on average).

TABLE 1. Out-of-sample (log-)likelihood values on the test data for the four stocks considered, with bold values indicating the maximum. Values in parentheses show the number of states preferred by the respective approach.

	Bayer	BMW	DB	VW
AIC	1949 (5)	1916 (5)	1641 (5)	**1757** (4)
BIC	1956 (3)	1859 (3)	1642 (3)	**1757** (4)
Lasso (λ selected via AIC)	**1983** (4)	**1922** (5)	**1648** (4)	1728 (4)
Lasso (λ selected via BIC)	1983 (4)	1893 (3)	1629 (3)	1728 (4)

4 Outlook

Current research focuses on simulation experiments to further investigate the proposed approach's properties and potential (dis)advantages compared to selecting the number of states via information criteria. In addition, we implement further out-of-sample criteria for evaluating the predictive performance in financial applications, e.g. considering the value at risk.

References

De Angelis, L. and Paas, L.J. (2013). A dynamic analysis of stock markets using a hidden Markov model. *Journal of Applied Statistics*, **40**, 1682 – 1700.

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357 – 384.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity*. Boca Raton: CRC Press.

Pohle, J., Langrock, R., Van Beest, F.M., and Schmidt, N.M. (2017). Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, **22**, 270 – 293.

Zucchini, W., MacDonald, I.L., and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton: Chapman & Hall-CRC.

Bayesian survival analysis using pseudo-observations

Léa Orsini[1], Caroline Brard[1], Emmanuel Lesaffre[2], David Dejardin[3], Gwénaël Le Teuff[1]

[1] CESP, INSERM U1018, Université Paris-Saclay, UVSQ, Villejuif, France
[2] I-Biostat, KU-Leuven, Leuven, Belgium
[3] Product Development, Data Sciences, F. Hoffmann-La Roche AG, Basel, Switzerland.

E-mail for correspondence: lea.orsini@gustaveroussy.fr

Abstract: In the frequentist framework, pseudo-observations analysis offers an alternative to the Cox proportional hazard model and is particularly interesting for complex survival modeling (multi-state model, recurrent events, interval censored data). Its advantage lies in its ability to break free from the complexity of censored data modeling using Generalized Estimating Equations (GEE). Yet Bayesian analysis of pseudo-observations may offer an alternative to the Bayesian survival analysis, which faces complexity either using non-parametric methods or full parametric Bayesian models depending on the baseline hazard assumption. Using pseudo-observations may result in a more straightforward formulation of the Bayesian model without making additional assumptions on the baseline hazard. This paper extends the analysis of pseudo-observations to the Bayesian framework using the Bayesian generalized method of moments. Similarly to the frequentist framework, this new approach gave valid estimates with similar performances compared to the Cox, GEE, and piecewise exponential models with large sample sizes. This approach may benefit other complex Bayesian survival models where censoring causes a substantial computational burden.

Keywords: Bayesian analysis; Survival analysis; Pseudo-observations; Generalized estimating equations; Generalized method of moments.

1 Methods

Pseudo-observations, defined in Andersen et al. (2003), may be used as an alternative approach to the Cox model and, more generally, in complex survival modeling, such as multi-states models. The K pseudo-observations of

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

the i-th individual are defined as $Y_{ik} = n\widehat{S}(t_k) - (n-1)\widehat{S}^{-i}(t_k)$ with n the sample size, $\widehat{S}(t)$ the Kaplan-Meier estimator of the survival probability at time t and $\widehat{S}^{-i}(t)$ the Kaplan-Meier estimator after excluding i-th individual. In practice, $K = 5$ time points equally spaced on the event time scale are sufficient to capture all the information from the Kaplan-Meier curve, see e.g. Klein et al. (2014).

Pseudo-observations are analyzed as an outcome variable in a regression model using generalized estimated equations, defined by Liang and Zeger (1986). This marginal approach is based on quasi-likelihood functions where only the moments are defined.
Suppose $X_i = (X_{i1}, ..., X_{iK})^{\mathrm{T}}$, $y_i = (y_{i1}, ..., y_{iK})^{\mathrm{T}}$ and $\mu_i = (\mu_{i1}, ..., \mu_{iK})^{\mathrm{T}}$ the covariates matrix, outcome vector, and mean vector for the i-th individual. $\beta = (\beta_1, ..., \beta_P)^{\mathrm{T}}$ is the vector of parameters to estimate. The mean model is defined as $\eta(\mu_{ik}) = X_{ik}^{\mathrm{T}}\beta$ where η is a monotone differentiable link function. The marginal variance is assumed to be a function of the mean $var(y_{ik}) = \phi v(\mu_{ik})$, see e.g. McCullagh et al. (1991).
The βs are estimated by solving the score equations:

$$U_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} u_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} D_i^{\mathrm{T}} V_i^{-1}(y_i - \mu_i) = 0,$$

where $D_i = \partial\mu_i/\partial\beta^{\mathrm{T}}$ and $V_i = \phi A_i^{1/2} R(\alpha) A_i^{1/2}$ with
$A_i = diag\{v(\mu_{i1}), ..., v(\mu_{iK})\}$. The working correlation matrix $R(\alpha)$ is assumed among specific forms. The nuisance parameters (ϕ, α) are alternately estimated with the βs, using moment estimations and a modified Fisher scoring algorithm, see e.g. Liang and Zeger (1986).

Yin (2009) proposed an approach that can be viewed as the Bayesian counterpart of GEE. It relies on the quadratic inference functions given in Qu et al. (2000). This approach is an extension of GEE, where the inverse of the working correlation matrix is expressed as a linear combination of basis matrix, $R^{-1} \approx \sum_{j=1}^{J} a_j M_j$. Contrary to GEE, the βs are estimated by applying the generalized method of moments (GMM), defined by Hansen (1982). The minimization problem of the generalized method of moments is equivalent to an MCMC sampling problem, see e.g. Chernozhukov and Hong (2003). Yin (2009) defined the pseudo-likelihood function as follows:

$$\tilde{L}(y|\beta) \propto \exp\{-\frac{1}{2} U_n^{\mathrm{T}}(\beta)\Sigma_n^{-1}(\beta)U_n(\beta)\},$$

where $\Sigma_n(\beta) = \frac{1}{n^2}\sum_{i=1}^{n} u_i(\beta)u_i^{\mathrm{T}}(\beta) - \frac{1}{n}U_n(\beta)U_n^{\mathrm{T}}(\beta)$,
with $U_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} u_i(\beta)$ and $u_i(\beta)$ a $(J \times p)$-dimensional score vector written as

$$u_i(\beta) = \begin{Bmatrix} D_i^{\mathrm{T}} A_i^{-1/2} M_1 A_i^{-1/2}(y_i - \mu_i) \\ D_i^{\mathrm{T}} A_i^{-1/2} M_2 A_i^{-1/2}(y_i - \mu_i) \\ \cdots \\ D_i^{\mathrm{T}} A_i^{-1/2} M_J A_i^{-1/2}(y_i - \mu_i) \end{Bmatrix}.$$

This paper extends the above method by implementing the Bayesian generalized method of moments to analyze pseudo-observations. We wrote our model using the Stan software (Carpenter et al., 2017) using a cloglog link function to interpret the estimates as hazard ratios. The working correlation matrix was assumed independent ($R(\alpha)$ equals the identity matrix), as is usually the case when applying GEE to pseudo-observations, see e.g. Klein et al. (2008).

2 Simulation study

2.1 Simulation settings

We performed a simulation study of 2-arm randomized clinical trials with a time-to-event outcome to assess the performance of the Bayesian generalized method of moments model applied to pseudo-observations. Simulated data were generated with a Weibull and a uniform distribution for event and censoring times, respectively. For the Bayesian GMM, MCMCs were performed using the rstan package with 3 chains of each 5000 iterations after a warm-up of 1000 iterations, thinning of 5, yielding 3000 iterations overall. Its performance was compared to the Cox model, the GEE model, and the Bayesian piecewise exponential model. All pseudo-observations-based models include an intercept, a treatment factor (X_1) and $K-1$ dummy variables for the time factor ($X_2, ..., X_5$). Bias, average standard error, root-mean-square error, and coverage were calculated from 1000 generated datasets with different simulation parameters of sample size, actual treatment effect, and censoring rate.

2.2 Priors elicitation

Convergence issues occur for some simulations when using non-informative priors $\beta_p \sim N(0, 1000)$. After some iterations, one of the three chains starts to diverge either with large variability of estimates or fixed to extreme estimates. One of the reasons may be that, unlike classical likelihood functions, the pseudo-likelihood is not defined for all values of β, but rather on a restricted support where Σ_n is invertible, see Figure 1. Convergence issues seem to occur when parameter values fall outside this local support. The inverse link function being $x \to \exp(\exp(x))$ results in extreme values. Consequently, Σ_n becomes quickly non-invertible when β values differ strongly from the actual simulation settings. This convergence issue may be resolved by truncating the priors of all coefficient regressions (except the one of the treatment effect) to reasonable values. Another solution is to choose more appropriate priors for the cloglog scale. Gelman et al. (2008) proposed to use Cauchy(0, 2.5) as default priors for generalized linear regression models after centering and re-scaling all the input variables. These weakly informative priors reflect the fact that large changes on the logit or cloglog scale

are rare. Using weak Gaussian priors such as $N(0, 10)$ or $N(0, 1)$, recommended by the Stan Development Team (2020), can provide an alternative to Cauchy priors. They may be more adapted to pseudo-likelihood defined on small support because they have lighter distribution tails. Finally, in our simulation settings, weakly informative priors $N(0, 10)$ were specified for all parameters when applied Bayesian GMM.

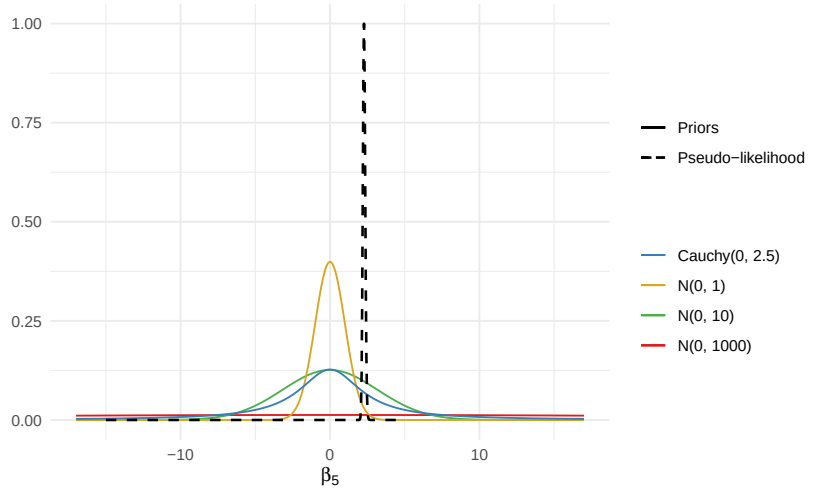


FIGURE 1. Pseudo-likelihood (dashed line) as a function of the last time point (β_5), all the other parameters are fixed to their GEE estimations. Solid lines represent different priors that have been investigated.

2.3 Results

Table 1 shows a slight bias and a larger variance with the GEE method compared to the Cox model for $n = 500$ patients and a 20% censoring rate. These results are consistent with Andersen et al. (2003). The bias and variance were larger with Bayesian GMM than with GEE; however, both decreased when the sample size increased (results not shown). As illustrated in Figure 2, the trace plots suggest the chains mixed well and appear to be stationary for the 1000 simulations.

3 Discussion

This paper presents a Bayesian approach based on pseudo-observations as an alternative to Bayesian survival proportional hazards models. Although the estimates are less efficient, this approach allows direct hazard ratio

TABLE 1. Comparison of log hazard ratio estimates between Cox model, GEE model, piecewise exponential model (PEM), and Bayesian GMM model.

$\beta_1^{true} = -0.5$		*Frequentist*		*Bayesian*	
Sample size		Cox	GEE	PEM	GMM
n = 500	Bias	0.00089	0.00197	0.00772	-0.00874
	ASE[1]	0.10262	0.11590	0.10277	0.11887
	RMSE[2]	0.10417	0.11423	0.10307	0.11693
	Coverage	94.7	95.2	94.6	95.0

[1] Average standard error [2] Root Mean Square Error

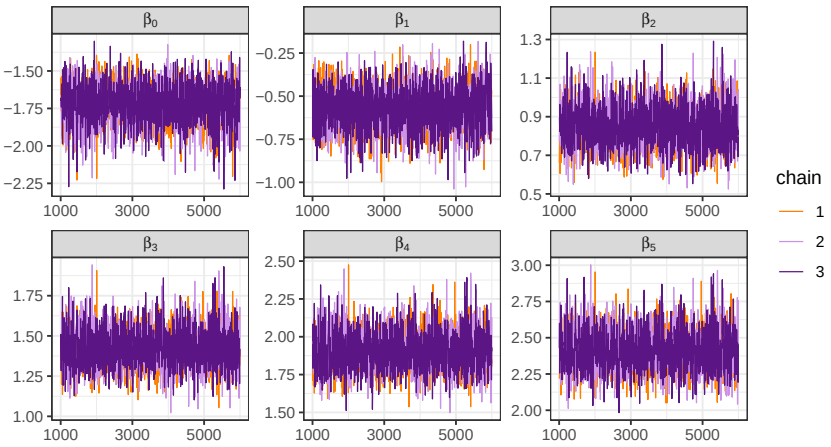


FIGURE 2. Post warm-up MCMCs, using the Bayesian GMM for the first simulated dataset, with n = 500. β_0 is the intercept, β_1 is the parameter of the treatment factor and $(\beta_2, ...\beta_5)$ are for the $K-1$ dummy time points variables.

estimations without specifying a full likelihood. Using pseudo-observations simplifies the Bayesian modeling of survival data since it does not rely on any assumption on the baseline hazard function. Complementary analysis is ongoing to assess this approach under different working correlation matrix assumptions through a comprehensive simulation study. Further research will extend this work to other applications where pseudo-observations are used in the frequentist framework, such as for restricted mean survival time, multi-state models, and more generally, complex survival models.

References

Andersen, P. K., Klein, J. P. and Rosthøj, S.(2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*. **90**(1), 15 – 27.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*. **76**(1), 1 – 32.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*. **115**(2), 293 – 346.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*. **2**(4), 1360 – 1383.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*. **50**(4), 1029-1054.

Klein, J. P., Gerster, M., Andersen, P. K., Tarima, S., and Perme, M. P. (2008). SAS and R functions to compute pseudo-values for censored data regression. *Computer Methods and Programs in Biomedicine*. **89**(3), 289 – 300.

Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H (2014). *Handbook of survival analysis*. CRC Press, Taylor and Francis Group.

Liang, K.-Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*. **73**(1), 13 – 22.

McCullagh, P., and Nelder, J. A. (1991). *Generalized linear models (2nd ed)*. London: Chapman & Hall.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*. **87**(4), 823 – 836.

Stan Development Team. (2020) Prior Choice Recommendations. Available online (accessed on 25 April 2023):

https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis*. **4**(2), 191 – 208.

Clustering anterior cruciate ligament reconstruction patients using functional walking biomechanics

Garritt L. Page[1], Matthew K. Seeley [1], Brian G. Pietrosimone[1]

[1] Brigham Young University, USA
[2] University of North Carolina, USA

E-mail for correspondence: page@stat.byu.edu

Abstract: We develop a method that clusters subjects based on two functional biomechanic outputs simultaneously. This produces a novel exploratory analysis technique of walking biomechanics that clusters 196 anterior cruciate ligament reconstruction (ACLR) patients into five distinct clusters. Patients in different clusters exhibited different walking biomechanics and more desirable patient-reported outcomes (PRO), clarifying potential relationships between walking biomechanics and desirable PRO post-ACLR.

Keywords: clustering; functional data; biomechanics.

1 Introduction

Exercise science researchers have previously used traditional regression analyses to relate biomechanical characteristics of walking and patient reported outcomes (PRO) among post-anterior cruciate ligament reconstruction (ACLR) subjects. But this approach limits researchers' abilities to discover possibly interesting relationships between PRO and walking biomechanics throughout the gait cycle. This study proposes employing the entire curve of two commonly encountered biomechanic variables (vertical ground reaction force (vGRF) and knee flexion angle (KFA)) to simultaneously inform cluster formation. Clustering subjects based on these two curves will permit connecting the PRO of post-ACLR patients to their gait which could highlight interesting patterns and lead to new discoveries about how to better treat post-ACLR patients.
This study consists of 196 post-ACLR patients. Each participant performed five walking trials in a biomechanics laboratory at the University of North

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

Carolina at Chapel Hill. For each trial, participants walked across a walkway, contacting a force platform that was set flush to the walking surface, facilitating measurement of vertical ground reaction force during walking. The participants were also instrumented with retroreflective markers placed over specific anatomical landmarks. Near infrared high-speed video was used to track the 3D position of these markers and estimate 3D knee joint kinematics such as KFA. The resulting vGRF and KFA mean curves for each subject are displayed in Figure 1.

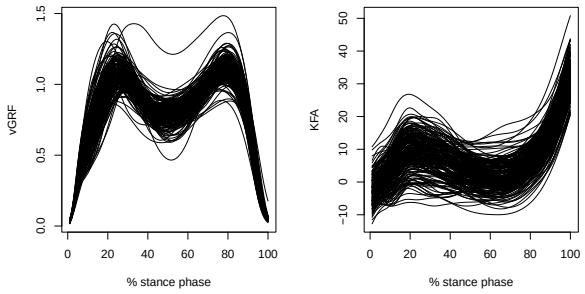


FIGURE 1. Vertical ground reaction force and knee flexion angle curves.

2 Methods

Let $y_{ij}(t)$ denote the jth variable's output for the ith individual at time t with $i = 1, \ldots, n$ and $t \in \mathcal{T}$. We note that $j = 1$ corresponds to vGRF an $j = 2$ to KFA. We assume that both vGRF and KFA are realizations of unknown subject-specific functions resulting in the following bivariate functional model

$$y_{ij}(t) = \beta_{0ij} + f_{ij}(t) + \epsilon_{ij}(t), \text{ for } j = 1, 2. \tag{1}$$

We assume that $\epsilon_{ij}(t) \sim N(0, \sigma^2_{ij})$ and $\beta_{0ij} \sim N(\mu_0, \sigma^2_0)$. Further, we approximate each $f_{ij}(\cdot)$ for $i = 1, \ldots, n$ using a B-spline basis so that

$$f_{ij}(t) = h'(t)\boldsymbol{\beta}_{ij}. \tag{2}$$

Here $h()$ is a known basis function and $\boldsymbol{\beta}_{ij}$ correspond to spline coefficients for subject i and curve j (Fahrmeir and Kneib 2005). We are interested in formulating a model that produces clusters based on $f_{ij}(t)$ but that also preserves subject-specific curve fits. To make our clustering approach concrete, let $c_1, \ldots, c_n$ denote each subjects cluster label such that $c_i \in \{1, \ldots, K\}$ where K denotes the number of clusters that is *a priori* unknown. Then, given each subjects cluster label, the following hierarchical

model for $\boldsymbol{\beta}_{ij}$ ensures good subject-specific fits while permitting flexibility in cluster formation

$$\boldsymbol{\beta}_{ij}|\boldsymbol{\theta}^{\star}_{j}, \lambda^{2\star}_{j}, c_i \sim N(\boldsymbol{\theta}^{\star}_{c_i j}, \lambda^{2\star}_{c_i j}\boldsymbol{I})$$
$$\boldsymbol{\theta}^{\star}_{jk} \sim N(\boldsymbol{\mu}_j, \tau^{2\star}_{j}\boldsymbol{P}^{-1}).$$

Note that $\boldsymbol{\theta}^{\star}_{kj}$ along with $\lambda^{2\star}_{kj}$ are cluster-specific means and variances with $k = 1, \ldots, K$. (Objects with a super-script '$\star$' are all cluster-specific.) As a result, the subject-specific spline coefficients vary around a cluster-specific mean spline coefficient vector. $\lambda^{2\star}_{kj}$ plays the crucial role of regulating the homogeneity of curves assigned to a particular cluster. To avoid knot selection and overfitting, we model $\boldsymbol{\theta}^{\star}_{kj}$ using Bayesian P-splines (Lang and Brezger 2004) so that $\boldsymbol{P}^{-1}$ denotes a penalty matrix from a lag-1 random-walk that has been adjusted so that it is full rank (i.e., the first column first row entry is changed from 1 to 2). $\tau^{2\star}_{j}$ is a cluster-specific smoothing parameter that determines the smoothness of the spline. We employ a product partition model (PPM) to model $c_1, \ldots, c_n$ (Quintana and Iglesias 2003). To introduce the PPM, note that alternative to cluster labels, $\rho = \{S_1, \ldots, S_k\}$ will be used to denote a clustering of n units so that $i \in S_k$ implies that $c_i = k$. Then the PPM has the following product form

$$Pr(\rho) \propto \prod_{k=1}^{K} c(S_k), \tag{3}$$

where $c(S_k) = (S_k - 1)!$ which favors clusterings with a small number of large clusters. The model is finished by employing the following prior distributions: $\sigma_{ij} \sim UN(0, 0.01)$, $\lambda^{\star}_{kj} \sim UN(0, 1)$, $\boldsymbol{\mu}_j \sim N(\mathbf{0}, 100^2 P^{-1})$, $\tau^{2\star}_{kj} \sim IG(1, 1/0.05)$, $\mu_0 \sim N(0, 100^2)$, and $\sigma^2_0 \sim IG(1, 1)$.

TABLE 1. Cluster means for symptomatic, mass, and months since surgery.

Cluster	Symptomatic	Mass(kg)	Months Since Surgery
red	0.58	72.39	20.14
blue	0.61	73.99	20.15
green	0.51	75.42	27.54
orange	0.68	72.49	22.42
purple	0.43	64.74	33.82

3 Results

Using `salso` (Dahl et. al 2021) to estimate ρ resulted in five distinct patient clusters, each with different biomechanical patterns (see Figure 2). The percentage of symptomatic patients significantly differed between the purple

cluster and the others (Table 1). The orange cluster contained the greatest percentage of symptomatic patients (68%), as well as GRF patterns reflecting an under-loading strategy (top right plot in Figure 2). While the purple cluster resulted in over-loading yet were the lightest group in terms of mass and exhibited the longest time since surgery. The purple cluster patients displayed increased GRF peaks and increased knee KFA, relative to the patients in the orange cluster. These results support the idea that PRO, post-ACLR, are at least partly related to walking biomechanics, including GRF, and knee joint kinematics and kinetics.

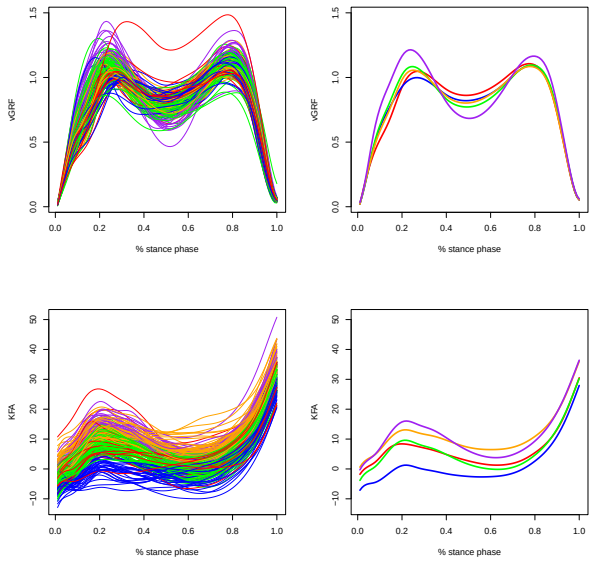


FIGURE 2. Clustered individual (left) and mean (right) vGRF and KFA curves.

References

Dahl, D., Johnson, D., and Mueller, P. (2021). *salso: Search Algorithms and Loss Functions for Bayesian Clustering. R package version 0.3.0.* https://CRAN.R-project.org/package=salso.

Fahrmeir, L. and Kneib, T. (2005). *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. New York: Oxford University Press, 1st ed.

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

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian Clustering and Product Partition Models. *Journal of the Royal Statistical Society Series B*, **65**, 557 – 574.

Forecasting insect abundance using time series embedding and environmental covariates

Gabriel R. Palma[1], Rodrigo F. Mello[2], Wesley A. Godoy[3], Eduardo Engel[3], Douglas Lau[4], Charles Markham[1], Rafael A. Moral[1]

[1] Hamilton institute, Maynooth University, Maynooth, Ireland
[2] Diretoria Estratégica de Dados, Itaú Unibanco SA, São Paulo, Brazil
[3] Entomology and Acarology, University of São Paulo, Piracicaba, Brazil
[4] Brazilian Agricultural Research Corporation (Embrapa Trigo), Passo Fundo, Rio Grande do Sul, Brazil;

E-mail for correspondence: gabriel.palma.2022@mumail.ie

Abstract: Implementing insect monitoring systems provides an excellent opportunity to create accurate interventions for insect control. Growers can use methods enlightened by Integrated Pest Management to prevent economic damage to their crops. However, selecting the appropriate time for applying an intervention is still an open question. This decision is even more critical with insect species that can abruptly increase their population size, such as the aphid *Rhopalosiphum padi* (Hemiptera: Aphididae). Moreover, studies involving the causal effect of other covariates are required to predict insect outbreaks accurately. Therefore, this research paper proposes a new approach to address this problem by combining statistics, machine learning, and time series embedding. We used a time series of aphids and climate data collected weekly in Coxilha (RS-Brazil) for eight years. We pre-processed the data using our newly proposed approach and more straightforward approaches. Using a Random Forests algorithm, we showed that our novel approach yields competitive forecasts when looking at the Root Mean Squared Error obtained from test data.

Keywords: Insect outbreak; Integrated Pest Management; Machine Learning; Forecasting; Causality.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

1 Introduction

Insect outbreaks have frequently been documented in pest populations. This ecological disturbance can affect forests and agroecosystems, resulting in economic and environmental damage. Yield loss is one of the many impacts of insect outbreaks due to the consumption of plants by the pest species. More specifically, arthropod pests are responsible for 20% of global annual crop losses. Also, pests may transmit diseases; for instance, *Rhopalosiphum padi* (Hemiptera: Aphididae) is the vector of barley yellow dwarf virus. Therefore, it is vital to carry out correct management of insect outbreaks to avoid economic damage.

These facts motivate developing and implementing of forecasting methods to prevent insect outbreaks. Especially methods that predict the best moment to proceed with chemical, or other types of interventions in the field, which are the focus of Integrated Pest Management (IPM). The vast number of alternative solutions based on IPM provides an excellent toolkit for growers; however, methods supporting accurate decisions for preventing insect outbreaks and reducing losses still lead to open research questions.

This problem also provides an opportunity for Machine Learning (ML) applications. However, many studies do not consider the cause-effect relationship, solely focussing on feature selection and ML methods prediction. This paper introduces a novel approach for forecasting insect abundance by combining statistics, machine learning, and time series analysis techniques. We present a framework for understanding the causal effects of insect abundance. Then we use a Random Forests (RF) method for predicting crop pest dynamics based on the causality studies. Finally, we compared the RF performance using the original dataset with all features, the dataset obtained from the proposed causal approach, and two datasets based on the insect abundance, with a delay of three and six time steps.

2 Methods

We used a time series of 211 observations of the total number of weekly sampled aphids obtained from a monitoring system in Coxilha (RS/ Brazil). Each observation of the time series is accompanied by climate covariates. We reconstructed the time series by unfolding time dependencies among observations so that the convergence provided by the Uniform Law of Large Numbers is held, a necessary condition to ensure supervised learning. This is obtained by applying Takens' embedding theorem, which reconstructs each observation $x(t)$ from a time series $\mathbf{x}$, for all $t = 0, \ldots, T$, in phase space, by creating a coordinate matrix $\mathbf{\Phi}$ such that the $t-$row is given by:

$$\phi_t = (x(t), x(t+d), \ldots, x(t+(m-1)d)),$$

Here, m is the embedding dimension (or the number of spatial axes), and d is the time delay between consecutive observations. We have that ϕ_t

corresponds to a position vector, or state, in phase space $\mathbf{\Phi}$. Given our interest in analysing cause-effect relationships among time series observations, we employ Granger's causality to map how a given exogenous or explanatory variable (other time series such as temperature or humidity) influences or anticipates events on the target time series. For this, we use a time delay of $d = 1$, and determine the embedding dimension m using AR (auto-regressive) models.

To compare the performance of our novel method, we (i) used only the target time series (population of insects), with up to 3 or 6 lags behind to forecast future observations and (ii) used all exogenous time series (environmental covariates) as predictors. We performed validation by obtaining one-step ahead forecasts for the entire time series (apart from the first few observations, in case of lagged predictors). We trained Random Forests for each set of features (reconstructed series, exogenous time series, 3- and 6-step lagged series) and obtained their performance based on the Root Mean Squared Error (RMSE). We computed the performance of the Random Forests algorithm 50 times, considering the stochastic nature of the method to obtain statistics from the RMSE values.

3 Results and discussion

Figure 1 shows the predictions of the Random Forests method based on the datasets containing all features, compared to a naive approach using the target time series solely with a delay of 3 and 6, as well as the novel method using reconstructed features. Averages RMSEs of 112.1 with a standard deviation of 0.60, 63.6 with a standard deviation of 0.60, 67.3 with a standard deviation of 0.54, and 65.6 with a standard deviation of 0.53 were obtained for these datasets, respectively, showing that the novel methodology proposed here is competitive.

4 Conclusions

The proposed reconstruction procedure presented a competitive performance in terms of predictive power. It shows the feasibility of applying these techniques to forecasting insect pest abundance, and therefore, this study will be a basis for developing new techniques to predict insect outbreaks.

Acknowledgments: This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Science Foundation Ireland.

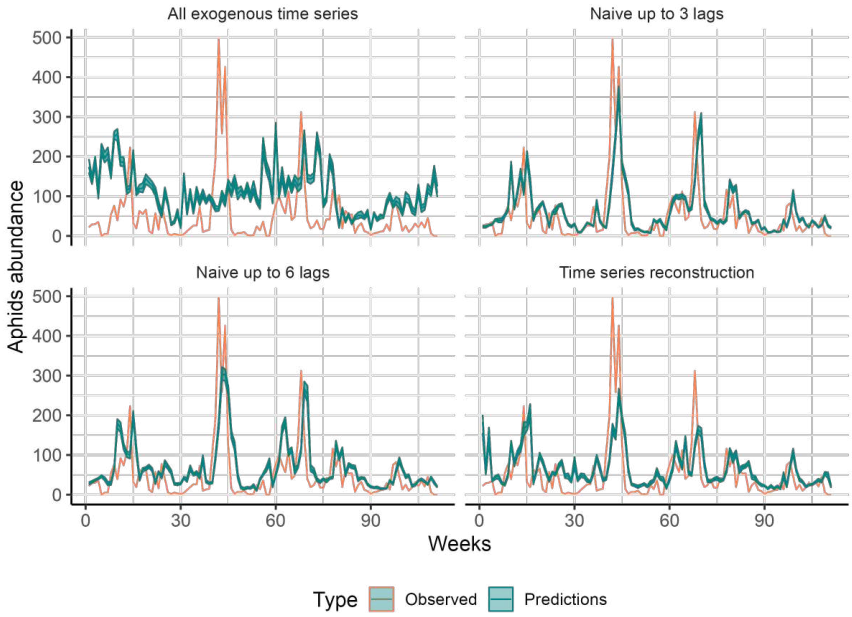


FIGURE 1. One-step forecasts for the time series obtained from the aphid monitoring system at Coxilha (RS-Brazil) based on ranfom forests trained with different sets of features. Our novel proposed method is the 'Reconstructed' series.

References

Santos, S., Specht, A., Carneiro, E., Paula-Moraes, S., Casagrande, M. (2017). Interseasonal variation of chrysodeixis includens (walker, [1858]) (lepidoptera: Noctuidae) populations in the Brazilian savanna In: *Revista Brasileira de Entomologia*, 294 – 299.

Lynch, A.N. (2018). Socioecological impacts of multiple forest insect outbreaks in the Pinaleño spruce–fir forest, Arizona. In: *Journal of Forestry*, 117.

Mateos Fernández, R., et al. (2022). Insect pest management in the age of synthetic biology. In: *Plant Biotechnology Journal*, 20(1), 25,–,36.

Smyrnioudis, I., Harrington, R., Clark, S., Katis, N. (2001). The effect of natural enemies on the spread of barley yellow dwarf virus (BYDV) by Rhopalosiphum padi (Hemiptera: Aphididae). In: *Bulletin of Entomological Research*, 91(4), 301,–,306.

Mello, R., Ponti, M. (2018). *Machine Learning: A Practical Approach on the Statistical Learning Theory*. Springer.

Studying animal interactions with Markov-switching step-selection models

Jennifer Pohle[1], Johannes Signer[2], Jana A. Eccard[1], Melanie Dammhahn[3], Ulrike E. Schlägel[1]

[1] University of Potsdam, Germany
[2] University of Göttingen, Germany
[3] University of Münster, Germany

E-mail for correspondence: jennifer.pohle@uni-potsdam.de

Abstract: Studying animal behaviour towards other conspecific or heterospecific individuals can provide new insights into local population dynamics and biodiversity patterns. We use Markov-switching step selection models to link the movement of an animal to occurrence estimates of other individuals obtained by kriging. This allows for the detection of interactions such as attraction or avoidance, while also accounting for temporal variation in the animals unobserved behaviour. We illustrate our approach in a case study on bank vole interactions.

Keywords: Animal movement; Hidden Markov models; Latent variables.

1 Introduction

Inter- and intraspecific interactions between animals, such as attraction or avoidance, influence local population and community dynamics. This in turn affects species distributions and biodiversity patterns. To detect and study such interactions based on concurrent movement data, Schlägel et al. (2019) propose a step-selection model which links the animals' movement decisions to dynamic occurrence estimates of other individuals obtained through kriging of their movement paths (Fleming et al., 2016). However, an animal's response to occurrence of other individuals can vary over time and depend on other, usually unobserved behavioral/biological states. For example, a resting animal may show no reaction at all, and a female's response to male occurrence may depend on its oestrous cycle phase (Schlägel et al., 2019). We therefore combine the approach with Markov-switching

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

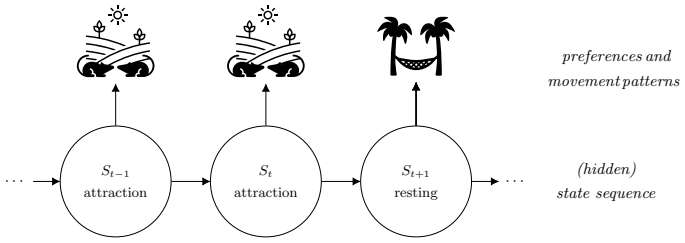


FIGURE 1. Illustration of the Markov-switching step selection model.

integrated step selection analysis (MS-iSSA, Pohle et al., 2022) which incorporates the state-switching patterns using an underlying latent Markov chain. We illustrate the method in a case study on fine-scale bank vole interactions.

2 Methods

Let $\{\mathbf{x}_{k,t}\}_{t=1}^{T}$ be the sequence of locations of individual k ($k = 1, \ldots, K$) observed at regular time intervals. In the Markov-switching step selection model (Figure 1), we assume the steps from location $\mathbf{x}_{k,t}$ to $\mathbf{x}_{k,t+1}$ to be driven by an underlying latent N-state Markov chain $\{S_t\}_{t=1}^{T}$ with transition probability matrix $\Gamma = (\gamma_{ij})$ and $\gamma_{ij} = \Pr(S_t = j \mid S_{t-1} = i)$. Given the spatial covariate map $\mathbf{Z}$ and the current state $S_t = i$, the distribution for a step to location $\mathbf{x}_{k,t+1}$ is modelled as:

$$f_i(\mathbf{x}_{k,t+1} \mid \mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}, \mathbf{Z}; \boldsymbol{\theta}_i, \boldsymbol{\beta}_i) = \frac{\overbrace{\Phi(\mathbf{x}_{k,t+1} \mid \mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}; \boldsymbol{\theta}_i)}^{\text{selection-free movement kernel}} \cdot \overbrace{\exp(\mathbf{Z}_{k,t+1}^{\mathrm{T}} \boldsymbol{\beta}_i)}^{\text{movement-free selection function}}}{\underbrace{\int_{\tilde{\mathbf{x}} \in \mathbb{R}^2} \Phi(\tilde{\mathbf{x}} \mid \mathbf{x}_{k,t}, \mathbf{x}_{k,t-1}; \boldsymbol{\theta}_i) \cdot \exp(\tilde{\mathbf{Z}}^{\mathrm{T}} \boldsymbol{\beta}_i) d\tilde{\mathbf{x}}}_{\text{normalising constant}}}.$$

$\Phi(\cdot)$ describes the space use density in absence of any habitat selection and is usually defined in terms of step length and turning angle. It is weighted by a log-linear function of the state-dependent selection coefficient vector $\boldsymbol{\beta}_i$ which indicates possible preference for or avoidance of the location-specific characteristics stored in the covariate vector $\mathbf{Z}_{k,t+1}$. Usually, $\mathbf{Z}_{k,t+1}$ contains classical habitat variables such as landcover type. For studying interactions, however, it mainly contains the occurrence estimates $\{O_{-k,t+1}\}$ of the other $K-1$ simultaneously tracked individuals. These are obtained from the corresponding location data by kriging within a rolling time period of length T_{krige} (Schlägel et al., 2019). Kriging assumes the movement process to be reasonably modelled by a Gaussian stochastic process and provides

a space use density estimate for the given time interval (Flemming et al., 2016).
The integral in the denominator of f_i is usually intractable. MS-iSSAs circumvent its evaluation by using a case-control design in which each observed location is mapped with M randomly drawn control locations. Parameters can then be estimated using a Markov-switching conditional logistic regression with a numerical maximisation of the corresponding log-likelihood (Pohle et al., 2022).

3 Bank vole interactions

We applied MS-iSSAs to movement data of synchronously tracked bank vole individuals (*Myodes glareolus*), a subset of the data analysed in Schlägel et al. (2019). Our data set contained 6-minute locations of $n = 12$ individuals split into 4 replicates with 2 males and 1 female each. Individuals within a replicate were synchronously tracked in fenced outdoor enclosures of 50m $\times$ 50m for $3 - 5$ days. Due to daily system maintenance, this resulted in $708 - 1,200$ locations per individual, split into $3 - 5$ bursts of around 23 hours each.
To study interactions between the bank voles, with a special interest in interactions between individuals of opposite sex, we applied 2-state MS-iSSAs to the individual movement data with state-dependent gamma distributions for step length and uniform distributions for turning angle. Occurrence estimates ($T_{\text{krige}} = 4\text{h}$) of all conspecifics within the same replicate were used as covariates for the selection part of the model. For parameter estimation, we used $M = 500$ control locations per observed location and tested 50 sets of random starting values to initialise the numerical optimisation. We further applied corresponding integrated step-selection analyses (iSSAs; no state-switching) and hidden Markov models (HMMs; no selection coefficients) to the same case-control data sets.
For most individuals, the MS-iSSA approach could reasonably distinguish between two activity levels. State 1 was associated to shorter step length compared to state 2 (Figure 2) and could thus reflect a rather inactive behaviour. This is in line with the estimated selection coefficients for occurrence of opposite-sex individuals: For all but one bank vole individual, the coefficients of state 1 were non-significant (p-values < 0.05; Figure 2) which indicates neutral behaviour towards the conspecifics. However, especially for male 1 in replicate D interpretation must be taken with care. Here, the Viterbi-decoded state sequence assigned over 96% of the observations to state 1 and the MS-iSSA showed larger mean step lengths in the estimated state-dependent gamma distributions than for all other individuals. Thus, the second state either captured rare events or outlying observations.
The selection estimates of state 2 ("active") indicated similar interactions as the iSSA (model without state-switching; Figure 2), except for the fe-

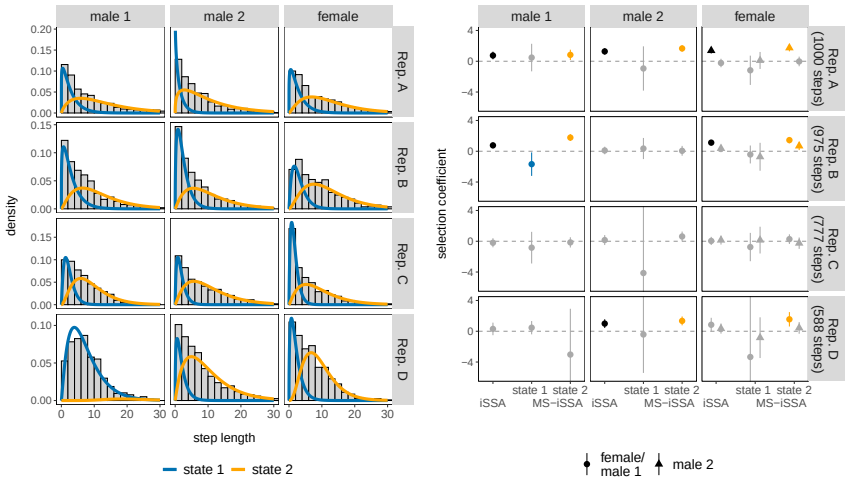


FIGURE 2. Results for the bank vole case study. The right panel shows the estimated state-dependent gamma-distributions for steps length (movement kernel) for each bank vole in each replicate, weighted by the relative Viterbi state frequencies. The left panel shows the estimated iSSA (black) and MS-iSSA (in colour) selection coefficients for occurrence of opposite-sex individuals with 95% confidence intervals. Non-significant coefficients ($p < 0.05$) are greyed out. Zero, positive and negative coefficients indicate neutrality, attraction and avoidance, respectively.

males in replicate B and D. Nevertheless, the information criteria AIC and BIC usually pointed to the Markov-switching step-selection model. Only for the female in replicate C both criteria selected the HMM (no selection coefficients). The BIC further preferred an HMM for the female in replicate D, and the iSSA (no state-switching) for male 1 in replicate D.

4 Discussion

Interactions between individuals are complex and often difficult to study, especially for animals in their natural environment. We provide an approach that uses movement data to detect interactions such as attraction or avoidance while also accounting for the temporal variation in the animals unobserved behaviour. In the bank vole case study, the MS-iSSA provides reasonable results and we are currently working on a second case study on yellowhammer interactions. It is important to note that the spatial and temporal resolution of the movement data can strongly influence the model results and interpretation. For example, some behaviours and interactions might not be expressed on very coarse (spatial, temporal) or very fine (temporal) scales.

Acknowledgments: We thank Sophie Eden, Angela Puschmann and Pauline Lange for help with the bank vole data collection and maintenance of the outdoor enclosures.

References

Fleming, C.H., Fagan, W.F., Mueller, T., Olson, K.A., Leimgruber, P. and Calabrese, J.M. (2016). Estimating where and how animals travel: An optimal framework for path reconstruction from autocorrelated tracking data. *Ecology*, **97**, 576 – 582.

Pohle, J., Signer, J., Schlägel, U.E. (2022). Markov-switching step selection analysis. *Proceedings of the 36th International Workshop on Statistical Modelling*, 284- 288.

Schlägel, U.E., Signer, J., Herde, A., Eden, S., Jeltsch, F., Eccard, J.A. and Dammhahn, M. (2019). Estimating interactions between individuals from concurrent animal movements. *Methods in Ecology and Evolution*, **10**, 1234 – 1245.

Prediction-based variable selection for component-wise gradient boosting

Sophie Potts[1], Elisabeth Bergherr[1], Constantin Reinke[2], Colin Griesbach[1]

[1] University of Goettingen, Goettingen, Germany
[2] University of Rostock, Rostock, Germany

E-mail for correspondence: sophie.potts@uni-goettingen.de

Abstract: Model-based component-wise gradient boosting is a popular tool for data-driven variable selection. In order to improve its prediction and selection qualities even further, several modifications of the original algorithm have been developed, that mainly focus on different stopping criteria, leaving the actual variable selection mechanism untouched. We investigate different prediction-based mechanisms for the variable selection step. These approaches include Akaikes Information Criterion (AIC) as well as a selection rule relying on the component-wise test error computed via cross-validation (CV). The AIC and CV routines are implemented for Generalized Linear Models and evaluated regarding their variable selection properties and predictive performance. The simulation study revealed improved selection properties whereas the prediction error could be lowered in a real world application with age-standardized Covid-19 incidence rates.

Keywords: Gradient Boosting; Variable Selection; Prediction Analysis; High-dimensional Data; Sparse Models.

1 Introduction

Regression settings with a large number of possible covariates necessitate tools for proper variable selection. One popular tool for data-driven variable selection is model-based component-wise gradient boosting (Bühlmann and Hothorn, 2007), from now on referred to as boosting.
Boosting is a statistical learning mechanism, that iteratively estimates the coefficient vector of a regression model. It refits simpler estimation procedures on the the gradient of a differentiable loss function (pseudo-residuals). In the case of Generalized Linear Models (GLMs), this loss is usually chosen

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

as the negative of the corresponding log-likelihood. For each iteration, the model fit is updated in a component-wise fashion, i.e. every single covariate is fitted separately and the algorithm identifies and updates the covariate whose update reduces the loss function the most. The boosting procedure is run for a pre-specified number of iterations. Calculating prediction criteria based on the coefficient vector of each iteration allows to select the best performing coefficient vector afterwards. If some covariates did not enter the model up to the selected iteration, they are excluded from the final model. Furthermore, the iteration selection enables coefficient shrinkage. Boosting combines further advantages like fast computation, model choice as well as the applicability in high-dimensional settings, where the number of covariates exceeds the number of observations and offers a flexible framework, that can be used for various applications.
There exist several modifications of the original boosting algorithm that aim to optimize its performance even further, e.g. to lower the false positive rate (FPR) such that only true influential variables enter the final model (for an overview see Mayr et al., 2017). The majority of approaches mainly focuses on different stopping criteria, leaving the actual variable selection mechanism untouched. Since boosting is a greedy algorithm, this variable selection step is of major importance. Thus, we investigate the variable selection step in order to increase the sparsity of boosted GLMs without sacrificing the good predictive performance. Therefore, two different prediction-based variable selection mechanisms are presented and evaluated in the following. Instead of minimizing a loss function, the new approach focuses on minimizing various prediction criteria directly.

2 Methods

Investigating the variable selection step of boosting is motivated by its direct impact on the variable selection process compared to a more downstream impact of the boosting iteration selection. We examine two different variable selection criteria for boosting of GLMs. Pulling the former iteration-selection criteria, namely CV and the AIC, into the variable selection loop of the algorithm may prevent the inclusion of false positives already during the fitting process.
Both prediction-based variable selection mechanisms do not need further time-consuming computations to determine the optimal stopping iteration, since they incorporate a prediction-based measure, which is evaluated for each covariate in every iteration. However, a minimum improvement threshold for the prediction criterion and the coefficient vector of 10^{-8} is specified.

3 Simulation

In order to test the performance of the two new prediction-based variable selection mechanisms, a simulation study with a normally distributed

outcome has been conducted. We varied the number of covariates k, the noise-to-signal ratio (NSR), the amount of true positive covariates (INF) and the correlation structures of the covariates. In terms of the FPR, the

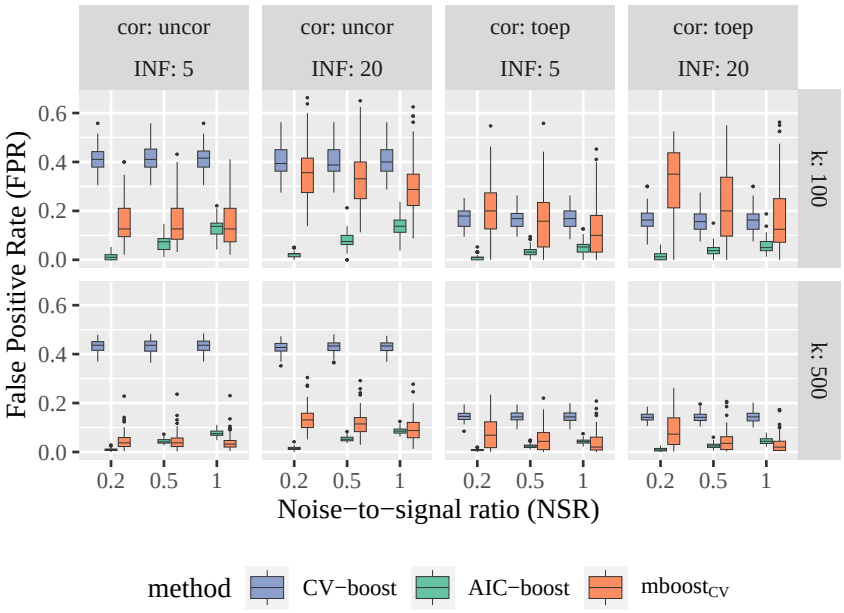


FIGURE 1. False Positive Rate of algorithms by simulation setting.

simulation study clearly recommends using the new AIC-boost since it outperforms `mboost` with 10-fold CV (mboost$_{CV}$) (Hothorn et al., 2010) in 19 of the 24 settings in terms of the median FPR (Figure 1). Furthermore, the variability of the FPR is drastically reduced in every tested simulation setup. The most influential parameter appears to be the NSR. In situations with strong signal, AIC-boost is more likely to outperform mboost$_{CV}$ than in weak signal situations. Enlarging the number of true informative covariates and keeping the other parameters constant results in a higher FPR of mboost$_{CV}$, whereas the two modifications perform more robust and reveal very similar FPRs across the differing numbers of informative covariates. In the majority of settings, CV-boost results in higher FPRs. However, the much lower FPR for AIC-boost is accompanied by a slightly lower TPR in some setups (not displayed here).

Since prediction and sparsity were seen as two opposing aims in model selection processes, the Mean Squared Prediction Error (MSPE) is analysed as well. Despite the fact that AIC-boost often results in sparser models, which increases the interpretability, it still keeps up with the benchmark model in terms of prediction accuracy (see Figure 2). CV-boost produces slightly

(considerably) higher median MSPE in low-dimensional (high-dimensional) settings with uncorrelated covariates. Using a Toeplitz covariance structure, differences between the three methods are less pronounced. Thus, applying the AIC-boost approach results in sparser models without sacrificing the predictive performance of the model in the vast majority of scenarios.

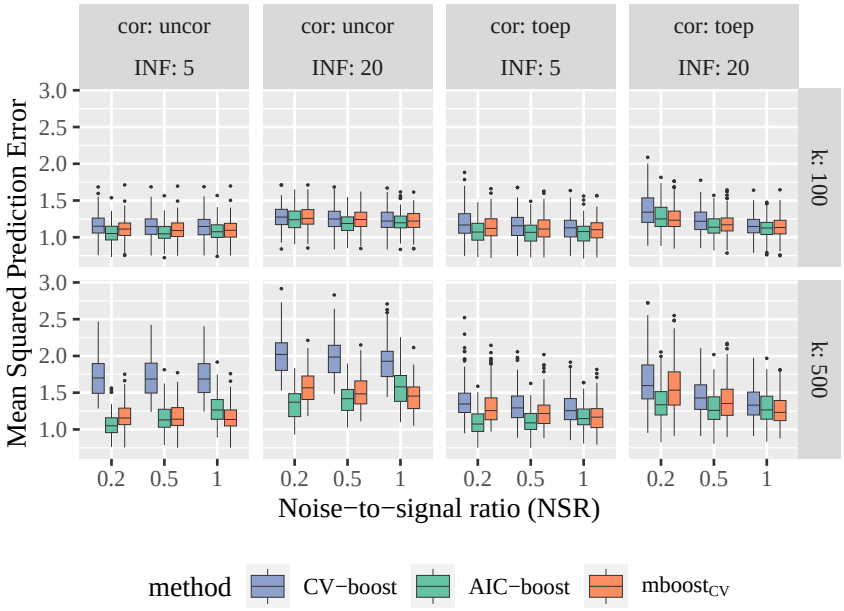


FIGURE 2. Mean Squared Prediction Error of algorithms by simulation setting.

4 Data Application on Covid-19 data

To evaluate the two modified algorithms in a realistic setting, they were applied to a real world data set. Therefore, a subset of the data base from Doblhammer et al. (2022) is used. The authors investigated the relationship of county-scale variables on the county-specific age-standardized Covid-19 incidence rates in Germany in order to uncover possible social disparities. The used subset of their data contains 163 variables measured on 401 counties in Germany. The variables cover socioeconomic characteristics, like demography, social economic and settlement structure, health care, poverty, unemployment as well as interrelationship with other regions.
By looking at the correlation structure of the possible covariates, one can observe highly correlated blocks of variables corresponding to thematically-related covariates. All variables are either metric or dummy-coded. Metric

covariates are standardized. The outcome of interest is the age-standardized incidence rate on the county-level for the first lockdown period (March 16, 2020 – March 31, 2020). It follows a log-normal distribution.
For reasons of comparability, the baseline model and the new CV-boost are trained using the same folds. Since these two algorithms are highly dependent on the data split, they are run five times with different data splits and median values are reported. The calculation of the MSPE is based on 100 randomly chosen data points. The high sparsity of $\text{mboost}_{\text{CV}}$ comes along

TABLE 1. Performance of algorithms on Covid-19 data set.

method	stopping iteration	no. of covariates	MSE	MSPE
$\text{mboost}_{\text{CV}}$ †	20	8	0.232	0.365
$\text{mboost}_{\text{AIC}}$	2039	85	2.553	2.801
LASSO-AIC	/	10	0.211	0.342
AIC-boost	164	21	0.154	0.296
CV-boost †	447	53	0.140	0.281

† model averaging performed, median values are reported.

with a poorer predictive performance regarding the MSE and MSPE. Both tested algorithms outperform $\text{mboost}_{\text{CV}}$ regarding the predictive performance but include (many) more variables than the baseline model. They also outperform LASSO with an AIC stopping criterion. Comparing the two new algorithms, the predictive performance is very similar by differing numbers of included covariates.

5 Summary

In summary, the findings suggest that the AIC modification can improve variable selection properties in component-wise gradient boosting by bridging sparsity and predictive performance. The purely loss-function based approaches CV-boost does not exhibit a lower FPR and further result in worse predictions. In the application on Covid-19 data, AIC-boost combines the sparsest model with a comparable prediction accuracy and thus would be the preferred model.
These results, however, have certain limitations, e.g. the modifications are restricted to the OLS base learner; however, other base learners, such as splines or tree-based ones, are worth investigating. This is also important in order to preserve the flexibility of the boosting framework. Despite the fact that the simulation study used a common model selection strategy as benchmark, further research comparing AIC-boost to other boosting modifications (e.g. probing (Thomas et. al 2017) or deselection (Strömer et. al 2022)) will provide a more complete picture of its performance. Since the

simulation study only addresses one type of outcome-distribution, a simpler simulation study with Poisson distributed values has been performed to overcome this limitation. Contrary to the expectations, the FPR is not reduced in a majority of the settings when applying AIC-boost. In most settings AIC-boost results in higher prediction errors. With regard to the TPR, $\text{mboost}_{\text{CV}}$ does not perform well, especially in the case of a high NSR, and it becomes obvious, that AIC-boost often results in a higher TPR. In this simulation, the higher sparsity of $\text{mboost}_{\text{CV}}$ comes with the drawback of a low TPR which is undesirable.
Some first assumptions on the underlying reasons for the differing results of the two simulations include the approximation of the hat matrix, which may be inaccurate in this setup. Since the advantages of using AIC-boost in terms of sparsity and prediction accuracy diminish, the preliminary results warrants further research to test the properties of the modifications for other types of GLMs, e.g. Binomial distributed outcomes.

Acknowledgments: Special Thanks to Gabriele Doblhammer and Daniel Kreft for aggregating and sharing the data. This work was supported by the DFG (Number 426493614) and the Volkswagen Foundation.

References

Bühlmann, P. and Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, **22(4)**, 477 – 505.

Doblhammer, G., Reinke, C., and Kreft, D. (2022). Social disparities in the first wave of COVID-19 incidence rates in Germany: a county-scale explainable machine learning approach. *BMJ Open*, **12(2)**, 1 – 11.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based Boosting 2.0. *Journal of Machine Learning Research*, **11(71)**, 2109 – 2113.

Mayr, A., Hofner, B., Waldmann, E., Hepp, T., Meyer, S., and Gefeller, O. (2017). An Update on Statistical Boosting in Biomedicine. *Computational and Mathematical Methods in Medicine*, **2017(6083072)**.

Strömer, A., Staerk, C., Klein, N., Weinhold, L., and Titze, S. (2022). Deselection of base-learners for statistical boosting - with an application to distributional regression. *Statistical Methods in Medical Research*, **31(2)**, 207 – 224.

Thomas, J., Hepp, T., Mayr, A., Bischl, B., and Zhao, Y. (2017). Probing for Sparse and Fast Variable Selection with Model-Based Boosting. *Computational and Mathematical Methods in Medicine*, **2017**, 1 – 8.

Computationally efficient ranking of groundwater monitoring locations

Peter Radvanyi[1], Claire Miller[1], Craig Alexander[1], Marnie Low[1], Wayne R. Jones[2], Luc Rock[3]

[1] University of Glasgow, School of Mathematics and Statistics, United Kingdom
[2] Shell Research Ltd, United Kingdom
[3] Shell Global Solutions Canada Inc, Canada

E-mail for correspondence: Peter.Radvanyi@glasgow.ac.uk

Abstract: Sampling groundwater quality monitoring wells is a costly and time intensive process that incurs health and safety risks. Reducing the number of wells whilst minimising information loss can greatly increase the sustainability of long-term monitoring. Wells that provide redundant information can be identified by assessing their observations' influence on statistical model estimates. Well-based cross-validation (WBCV) could be used to obtain such a measure of influence for each well, however, the associated computational cost renders this option unfavourable. In this paper, we propose a method based on influence statistics of regression-based, groundwater solute concentration models, as a computationally efficient, approximate alternative. The method, named well influence analysis (WIA), approximated WBCV results in a simulation study and real groundwater contaminant observations with an average 77% and 73% accuracy respectively. WIA will be implemented in the "well redundancy analysis" feature of GWSDAT, an open-source software for the spatiotemporal modelling of groundwater monitoring observations.

Keywords: Groundwater Monitoring; Groundwater Contamination; Statistical Modelling; Spatiotemporal; Influence Statistics.

1 Introduction

The aim of groundwater quality monitoring during the remediation of contaminated sites is to understand the behaviour of the solutes of concern by observing changes in their concentration levels at fixed sampling locations called wells. Spatiotemporal statistical models can be used to estimate contaminant concentrations over spatial domains of interest using these

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

observations. However, collecting and analysing samples from groundwater monitoring wells is costly, time intensive and incurs health and safety risks. Reducing sampling intensity whilst minimising the loss of information can greatly increase the efficiency and sustainability of long-term groundwater quality monitoring. Sampling intensity can be decreased by reducing the number of sampling locations. In many cases, fewer wells can be sufficient for supporting robust statistical models, provided they adequately capture the spatiotemporal heterogeneity in solute concentrations. Therefore, the choice of monitoring wells to omit from sampling is crucial, and should be based on qualitative and quantitative analyses. A possible quantitative approach using statistical models is assessing sampling wells based on their observations' impact on solute concentration estimates. Wells whose data provide redundant information to the model, could be considered for omission from future sampling campaigns. Feedback from users of the open-source, spatiotemporal groundwater quality modelling software, GWSDAT (Jones et al. 2014), highlighted the need for a tool to facilitate this well redundancy analysis. Ranking wells by influence prior to testing the impact of omitting one, aims to reduce the need for a trial-and-error approach. Assessing well influence can be done iteratively, using well-based cross-validation (WBCV). However, the computational cost associated with re-fitting the model in each iteration makes this approach unfavorable. In this work, we aim to show that for regression-based groundwater contamination models, well influence analysis (WIA) could be a computationally efficient, approximate alternative to the cross-validation-based method. WIA provides a suggested sequence for omitting wells, by ranking them using influence statistics commonly used in regression analysis. The proposed method was tested in a simulation study and on real groundwater monitoring data.

2 Simulation Study

A simulation study (Radvanyi, 2023) was designed to analyse how closely WIA approximated the cross-validation-based well influence rankings in different scenarios, and to compare different influence statistics that could be used for WIA. The simulation study was conducted using synthetic data sets.

2.1 Synthetic Data

The synthetic data (McLean et al. 2019) contained coordinates, sampling times and solute concentrations for three hypothetical contaminant plumes of increasing complexity simulated using process based models (Figure 1). 15 % multiplicative random noise was applied to the data to mimic the measurement errors of real groundwater observations. Samples were then drawn at select times and coordinates to mimic sampling from monitoring

wells. Nine monitoring network designs were created for each plume using 6, 12 and 24 wells with 3 well placement strategies. These strategies were random, grid and expert, the latter implying knowledge of plume characteristics, such as origin and groundwater flow direction. Each scenario ran for 100 iterations.

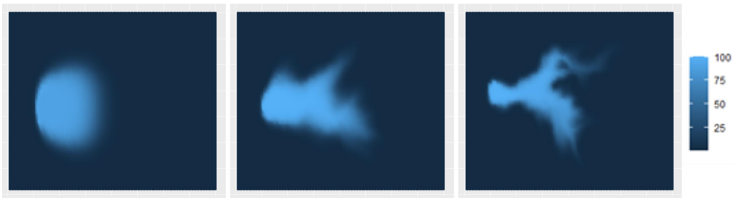


FIGURE 1. Hypothetical plumes: simple (l), moderate (c) and complex (r).

2.2 Modelling Approach

Concentration estimates over the full spatial domain were obtained using P-splines models, also used in GWSDAT (Evers et al. 2015). P-splines (Eilers & Marx, 1996) are regression splines fitted by least-squares with a roughness penalty. The P-splines model can be written as

$$y_i = \sum_{j=1}^{m} b_j(x_i)\alpha_j + \epsilon_i,$$

where y_i, $i = 1, 2, ...n$, are the natural logarithm of the solute concentrations, x_i are the corresponding coordinates and sampling times, b_j, $j = 1, 2, ...m$, are B-spline basis functions, α_j are the basis coefficients and ϵ_i are errors, assumed to be independent with $N(0, \sigma^2)$.

2.3 Well-Based Cross-Validation

The baseline ranking of well influence on estimated solute concentrations was computed via well-based cross-validation (WBCV; Evers et al. 2015). WBCV is a form of leave-one-out cross-validation, where each well (and hence associated observations) was removed sequentially and used as the test set for a model trained on the remaining data. The well ranking was given by the numerical order of corresponding root-mean-square errors (RMSE) calculated by:

$$RMSE_k = \sqrt{\frac{\sum_{l=1}^{n_k} (y_{kl} - \hat{y}_{kl})^2}{n_k}},$$

where $k = 1, .., w$ and w is the total number of wells, y_{kl} is the l-th observation from the k-th well, $\hat{y}_{kl}$ is the l-th fitted value and n_k is the number of observations.

2.4 Well Influence Analysis

Different influence metrics were compared for approximating the WBCV rankings. Cook's distance (CD; Cook, 1977), which is a measure of the sum of changes in regression estimates if an observation is deleted, produced the most informative results. It can be expressed using leverages, which are the diagonal elements of the projection matrix from the P-splines model:

$$CD_i = \frac{1}{p}(r_i^s)^2 \frac{h_{ii}}{1 - h_{ii}},$$

where p is the effective degrees of freedom, r_i^s is the standardised residual and h_{ii} is the leverage of the i-th observation. The rankings were given by the numerical orders of the median CD values for each well. The studied influence metrics were originally derived for linear regression. Their application in this case is supported by the fact that P-splines are analogous to linear regression.

2.5 Assessing Performance

The performance of WIA was quantified by calculating the normalised difference score D_n, which indicated the total difference in well ranks between WIA and WBCV. D_n is bounded by $0 \leq D_n \leq 1$ with 0 meaning the rankings were equivalent. D_n was calculated by

$$D_n = \frac{\sum_{i=1}^{w} |o_i^{wbcv} - o_i^{ia}|}{D_{max}},$$

where o_k^{wbcv} is the rank of the k-th well based on WBCV and o_k^{ia} is its rank based on WIA. The maximum difference between the two rankings, D_{max} is a function of the number of monitoring wells such that $D_{max} = \frac{w^2}{2}$.

2.6 Results

The mean D_n for CD-based WIA was 0.23, which means that on average, it approximated the baseline (WBCV) rankings with 77% accuracy. Figure 2 shows the results in the form of a boxplot categorised by scenario design features. Mean D_n values increased with plume complexity from 0.20 to 0.27. The complex plume is also associated with the highest variance. The monitoring well network design also seemed to play a role in the outcome of the analysis. The results show that WIA has better performance if well placement is done based on site characteristics as opposed to randomly or in a grid pattern. In terms of the number of monitoring wells, the smallest mean D_n results were obtained with 6 wells. However, this is most likely an artifact related to fewer possible differences in well ranks between WIA and the baseline. This effect also seems to disappear given a sufficient number of wells, since there is little difference in results between 12 and 24 wells.

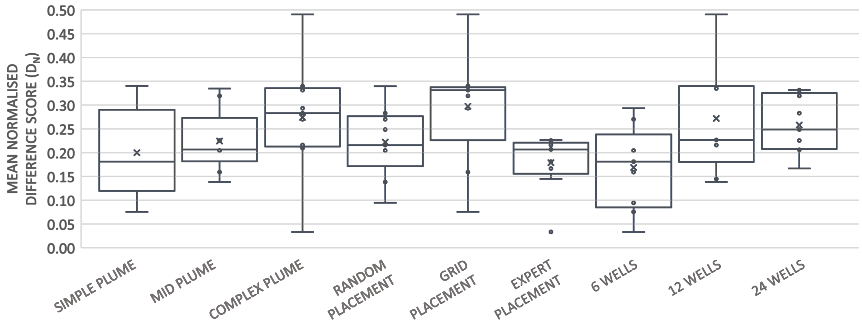


FIGURE 2. Breakdown of mean normalised difference scores (D_n; $0 \leq D_n \leq 1$) by design attributes plume complexity, well placement and the number of wells. A smaller D_n indicates a more accurate estimation of WBCV rankings by WIA.

3 Real Data Application

The comparison of WIA and WBCV was also performed on real groundwater contamination data from an undisclosed monitoring site. The data set contained the concentrations of five contaminants in groundwater samples from 32 monitoring wells collected over a 4 year period. The contaminants were modelled separately. Table 1 shows the results of the analysis by contaminant.

TABLE 1. Breakdown of normalised difference scores (D_n; $0 \leq D_n \leq 1$) by contaminant from the groundwater monitoring data. A smaller value indicates a more accurate estimation of WBCV ranking by WIA.

Contaminant	$\mathbf{D_n}$
Ethylbenzene	0.36
Toluene	0.25
Nitrate	0.18
Sulphate	0.23
TPH	0.32
Mean	**0.27**

The mean D_n was 0.27, which translated to an average of 73% accuracy in comparison to WBCV. Just as in the simulation study, most of the deviation in the WIA ranking compared to WBCV was due to an aggregation of minor rank differences. This means that wells generally occupied similar positions in both rankings.

4 Conclusions

In conclusion, empirical evidence was presented to support the application of influence statistics in the proposed context. WIA estimated WBCV results with an average 77% and 73% accuracy in the simulation study and real data examples respectively. These results were positive given the aim and the approximate nature of the analysis. The simulation study also showed that the monitoring network design and contaminant plume characteristics also affect the accuracy of WIA. WBCV would be the preferred ranking method, but it is computationally unfavorable because it requires fitting w models for each well that is considered for omission from future sampling campaigns. In contrast, WIA only requires a single model before each omission, which makes it a more efficient alternative to WBCV for ranking wells by influence on solute concentration estimates. In other words, there is trade-off between accuracy and computational efficiency, but the results indicate that in this case, the gain in efficiency is greater than the loss in accuracy. WIA is easy to implement in software built around regression-based groundwater quality models, such as GWSDAT, and it can help determine the sequence in which wells should be omitted during well redundancy analysis.

References

Cook, R.D. (1977). Detection of Influential Observations in Linear Regression. *Technometrics*, **19**, 15 – 18.

Eilers, P.H.C., Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11.2**, 89 – 121

Evers, L., Molinari, D.A., Bowman, A.W., Jones, W.R., Spence, M.J. (2015). Efficient and automatic methods for flexible regression on spatiotemporal data, with applications to groundwater monitoring, *Environmetrics*, **26.6**, 431 – 441.

Jones, W.R., Spence, M.J., Bowman, A.W., Evers, L., Molinari, D.A. (2014). A software tool for the spatiotemporal analysis and reporting of groundwater monitoring data. *Environmental Modelling and Software*, **55**, 242 – 249.

McLean, M.I., Evers, L., Bonte, M., Bowman, A.W., Jones, W.R. (2019). Statistical modelling of groundwater contamination monitoring data: A comparison of spatial and spatiotemporal methods. *Science of the Total Environment*, **652**, 1339 – 1346.

Radvanyi, P. (2023). Well Influence Analysis. https://github.com/peterradv/Well-Influence-Analysis

A distributional regression approach for Gaussian process responses

Hannes Riebl[1], Nadja Klein[2], Thomas Kneib[1]

[1] Chair of Statistics, Georg-August-Universität Göttingen, Germany
[2] Chair of Statistics and Data Science, Humboldt-Universität zu Berlin, Germany

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

Abstract: Measurements of high-resolution tree circumference dendrometers are the result of two distinct processes: the irreversible growth of the tree stem and reversible shrinking and swelling. We propose a novel statistical method that allows us to decompose these measurements into a permanent and a temporary component, while explaining differences between the trees and years by covariates. Our model embeds Gaussian processes (GPs) with parametric mean and covariance functions as response structures in a distributional regression framework with structured additive predictors. We present different mean and covariance functions, connections with other model classes, and demonstrate the efficiency of our Markov chain Monte Carlo sampling scheme in applications and simulations.

Keywords: Generalized additive model for location, scale, and shape; growth curve model; Matérn covariance function; spatio-temporal regression.

1 Introduction

Tree growth, and the growth of tree stems in particular, is a process that is of strong ecological and economic interest. Unfortunately, it is difficult to measure the formation of new wood and bark cells in the cambium resulting in permanent stem growth, and while electronic dendrometers can record the variation of the stem circumference on small time scales, these measurements also capture the reversible shrinking and swelling of the stem due to changes in its water content.
We propose a novel statistical method for the analysis of high-resolution dendrometer measurements that permits us to decompose the dendrometer measurements into a permanent and a temporary component through stochastic assumptions and explanatory variables. Figure 1 shows a subsample of the recorded growth curves, each of which is assumed to be a realization of a Gaussian process (GP). We observe that the colored ash grows primarily between mid-April and mid-July, while the colored beech grows later and more during the vegetation period.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

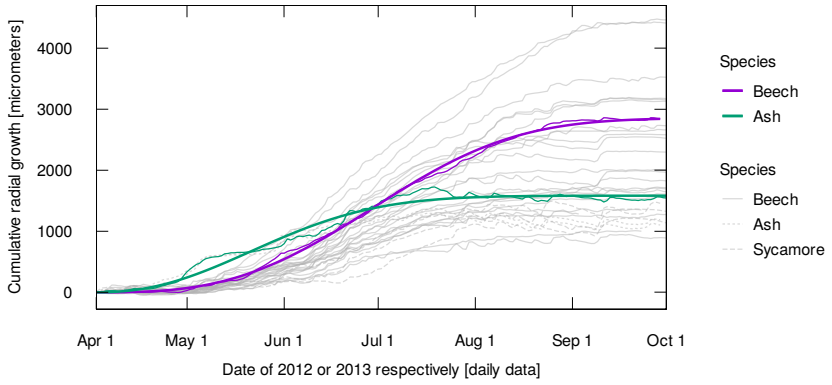


FIGURE 1. Cumulative radial growth of a subsample of the trees from our dataset. Colored lines represent two exemplary trees, one beech and one ash, while the gray lines illustrate the diversity of the growth patterns in the dataset.

The fact that we link multiple properties of the mean and covariance functions of the GPs to explanatory variables puts our model in the domain of so-called distributional regression models. Standard distributional regression models use univariate or low-dimensional multivariate response variables. Following this line of thought, we show that the distributional regression approach also works for more general, continuous response structures such as GPs.

2 Gaussian Process Responses

We consider GPs $\{Y_i(t); t \in T\}$ as response structures in structured additive distributional regression, where the observation index i runs from 1 to N and the index set T is a metric space that can represent time, space, or space-time. The GPs are assumed to be conditionally independent given the covariate vectors x_i,

$$\{Y_i(t); t \in T\} \mid x_i \overset{\text{ind.}}{\sim} \mathcal{GP}(m(t; x_i), c(t, t'; x_i)), \tag{1}$$

where $t, t' \in T$. As a specific feature of distributional regression, the mean function $m(\cdot; x_i)$ and the covariance function $c(\cdot, \cdot; x_i)$ both depend on the covariates x_i, which differ between the observations 1 to N but are constant within the index set T.
More precisely, the mean and the covariance function are linked to the covariates x_i via their respective parameter vectors θ^m and θ^c,

$$m(t; x_i) = m(t; \theta^m(x_i)) \qquad \text{and} \qquad c(t, t'; x_i) = c(t, t'; \theta^c(x_i)).$$

Let $\theta_i = [\theta^m(x_i)^\top, \theta^c(x_i)^\top]^\top$ be the vector of all parameters of the GPs and K its dimension. Each parameter θ_{ki} is then linked to a predictor η_{ki} via a

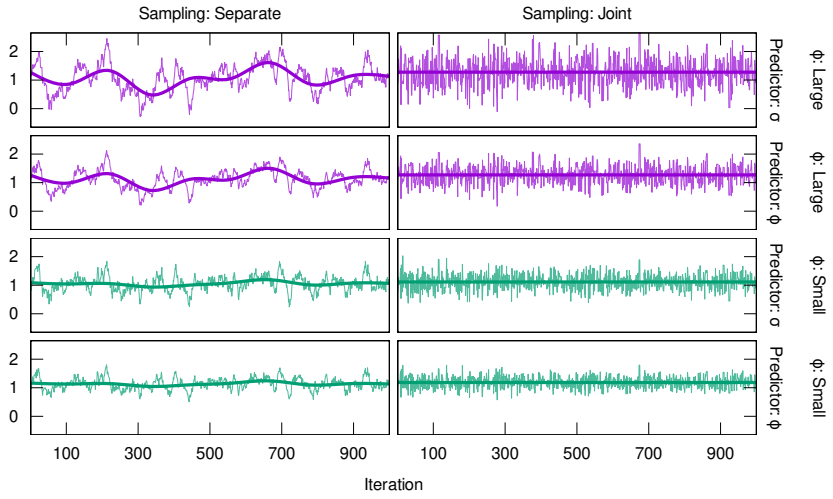


FIGURE 2. Trace plots for a sampler with separate blocks for the regression coefficients for the covariance parameters (left) and a sampler with one joint block for the coefficients (right).

strictly monotonic link function h_k, for $k = 1, \ldots, K$, i.e. $h_k(\theta_{ki}) = \eta_{ki}$ or $\theta_{ki} = h_k^{-1}(\eta_{ki})$.

In practice, each GP $\{Y_i(t)\}$ can only be observed at a finite number of points $t_j \in T$, for $j = 1, \ldots, n_i$. The collection of random variables at these points has a multivariate normal distribution,

$$[Y_i(t_1), \ldots, Y_i(t_{n_i})]^\top \mid z_i(t_1), \ldots, z_i(t_{n_i}), x_i \overset{\text{ind.}}{\sim} \mathcal{N}_{n_i}(\mu_i, \Sigma_i), \tag{2}$$

where the elements of the mean vector μ_i and the covariance matrix Σ_i are the evaluations of the mean function m and the covariance function c at the observed points

$$\mu_i = [\mu_{i,j}] = m(z_i(t_j); \theta_i^m) \text{ and } \Sigma_i = [\sigma_{i,j,j'}] = c(z_i(t_j), z_i(t_{j'}); \theta_i^c) \tag{3}$$

for $j, j' = 1, \ldots, n_i$.

From these distributional assumptions, we can derive the likelihood (as well as the score function and Fisher information) which serve as essential parts of our Bayesian treatment of GP response regression. We perform fully Bayesian inference with an MCMC sampler which uses inverse gamma priors and Gibbs updates for each scalar element of the smoothing variance, and Metropolis-Hastings updates with locally adaptive IWLS proposals for the regression coefficients. As the IWLS proposals involve the observed or expected Fisher information matrix, the regression coefficients are sampled in blocks for numerical stability and efficiency. Typically, one block consists of the regression coefficients of one smooth term, and the blocks are

sampled in a nested loop over the distributional parameters first and the smooth terms second. As discussed in the next section, we sample the parameters of certain smooth terms in one joint block, which can reduce the autocorrelation of the MCMC chains substantially as illustrated in Figure 2.

3 Application and Simulations

In our analysis of stem growth, we rely on a Weibull growth curve

$$m^w\left(z(t) = t; \theta^m = [l, a, b]^\top\right) = l \times \left[1 - \exp\left(-\left(\frac{t}{b}\right)^a\right)\right]$$

as a mean function and a scaled Matérn covariance function

$$c^m\left(t, t'; \theta^c = [\sigma, \phi]^\top\right) = \sigma^2 \times \rho\left(\frac{d(t, t')}{\phi}; \nu\right),$$

where ρ is the Matérn correlation function with smoothness parameter ν, and $d(t, t')$ is a distance function. Consequently, we have the following five distributional parameters: the limit l, the shape a, and the scale b of the Weibull growth curve, and the standard deviation σ and the range ϕ of the covariance function. Each of these parameters is related to regression effects of different complexity with the predictors and inverse link functions being defined as

$$\begin{aligned}
l_i &= \exp(\beta_{l0} + (\text{Tree} * \text{Year})_i \times \beta_{l1}),\\
a_i &= \exp(\beta_{a0} + \text{Species}_i \times \beta_{a1} + \text{DBH}_i \times \beta_{a2} + (\text{Site} * \text{Year})_i \times \beta_{a3}),\\
b_i &= \exp(\beta_{b0} + \text{Species}_i \times \beta_{b1} + \text{DBH}_i \times \beta_{b2} + (\text{Site} * \text{Year})_i \times \beta_{b3}),\\
\sigma_i &= \exp(\beta_{\sigma0} + \text{Species}_i \times \beta_{\sigma1} + \text{DBH}_i \times \beta_{\sigma2} + f_{\text{Year}_i}(x_i, y_i; \beta_{\sigma3})),\\
\phi_i &= \exp(\beta_{\phi0} + \text{Species}_i \times \beta_{\phi1} + \text{DBH}_i \times \beta_{\phi2} + (\text{Site} * \text{Year})_i \times \beta_{\phi3}),
\end{aligned}$$

where $\beta_{\bullet,0}$ and $\beta_{\bullet,2}$ are scalar regression coefficients, while $\beta_{\bullet,1}$ and $\beta_{\bullet,3}$ are vectors of regression coefficients, and Species_i denotes the entries of the design matrix for the dummy variable for the species of the tree where the i-th growth curve was recorded,

$$\text{Species}_i = \begin{cases} [0,0] & \text{if the } i\text{-th growth curve is of a beech,}\\ [1,0] & \text{if it is of a ash, and}\\ [0,1] & \text{if it is of a sycamore.}\end{cases}$$

Similarly, $(\text{Tree} * \text{Year})_i$ and $(\text{Site} * \text{Year})_i$ are the entries of the design matrix for the interaction of two dummy variables: in the first case, of the individual tree and the year, and in the second case, of the field site and the year. Finally, f_{Year_i} denotes a year-specific spatial kriging smooth. Figure 3

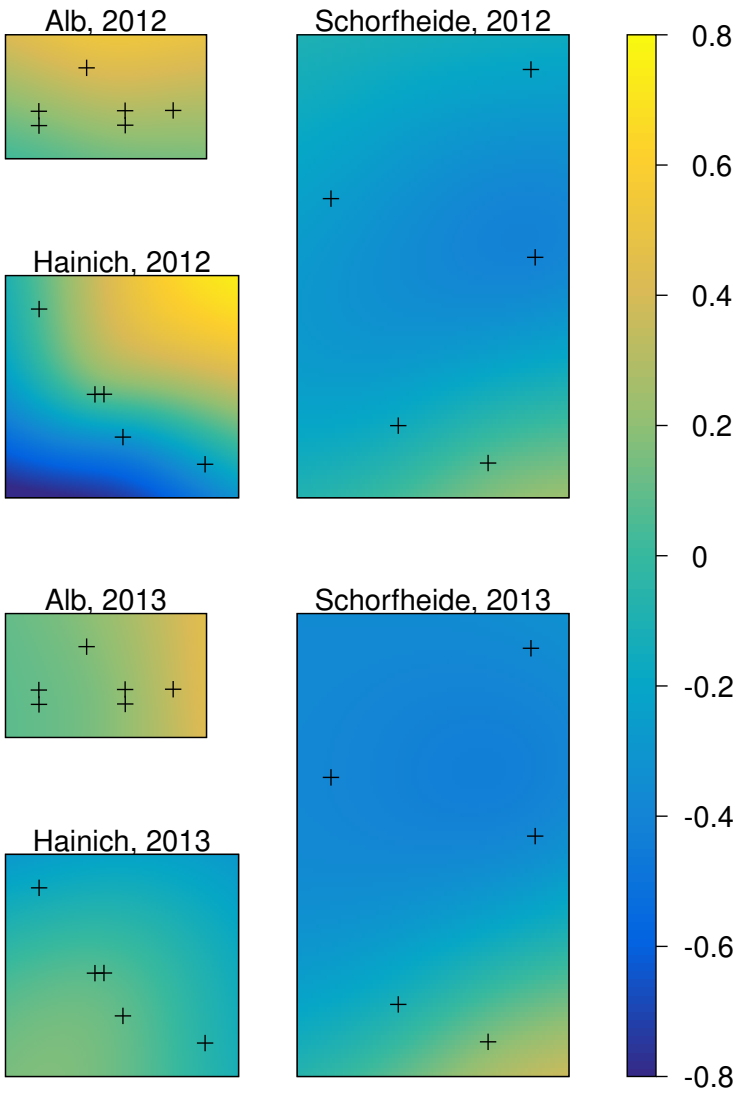


FIGURE 3. Posterior mean of a spatial kriging smooth in the predictor for the standard deviation. The field sites are marked with small crosses. Lighter colors indicate a higher standard deviation.

exemplarily shows the posterior mean estimate of a spatial smooth in the predictor for the standard deviation.

Figure 4 illustrates one of our simulation scenarios where we explore the extension of GP regression to processes on a sphere to highlight that higher-dimensional Euclidean space or even non-Euclidean metric spaces can be

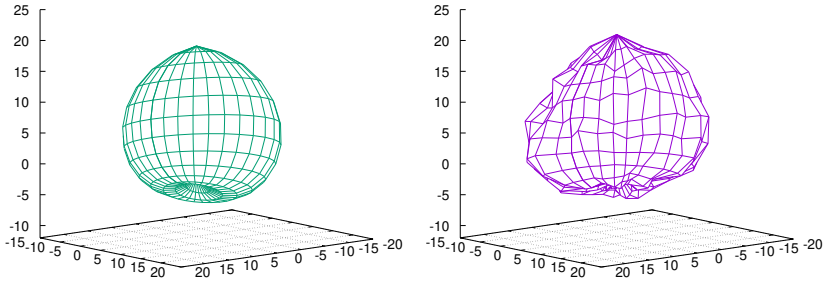


FIGURE 4. Mean function (left) and corresponding realization (right) of a GP with an exponential covariance function resembling a stylized crown shape.

treated when employing appropriate distances. The processes in this specific scenario are defined on a sphere, resembling shapes of tree crowns, and we use the great circle distance for quantifying distances. In an application, the tree species or the light availability could be used as covariates to explain the properties of the mean and the covariance function of the crown shapes. The mean properties are, among others, the average radius and the vertical elongation, while the covariance properties are the size and the persistence of the deviations from the mean.

References

Riebl, H., Klein, N., and Kneib, T. (2023). Modeling Intra-Annual Tree Stem Growth with a Distributional Regression Approach for Gaussian Process Responses. Journal of the Royal Statistical Society Series C (Applied Statistics), to appear.

Multi-state models for double transitions associated with parasitism in biological control

Idemauro Antonio Rodrigues de Lara[1], Gabriel Rodrigues Palma[3], Victor José Bon[2], Carolina Reigada[2] , Rafael de Andrade Moral[3]

[1] University of São Paulo, Brazil
[2] Federal University of São Carlos, Brazil
[3] Maynooth University, Ireland

E-mail for correspondence: idemauro@usp.br

Abstract: The motivation for this work was an experiment was developed to evaluate the effects of previous parasitism on the parasitism rate of the species *Trissolcus basalis* and *Telenomus podisi*. The statistical problem in this study was to model, successively, the choice of eggs (with four possibilities with parasitised eggs or not) and the conditional behaviour given the choice (marking, ovipositing or drumming on the chosen egg). We consider state-space models in two successive steps to calculate double transition probabilities. We found statistically significant differences regarding the choice of parasitised eggs, with *T. podisi* being more likely to choose healthy eggs than the competing species.

Keywords: stochastic process, likelihood procedure, entomological data.

1 Introduction

Longitudinal studies with categorical data are very common in the Entomology and extensions of Generalized Linear Models can be used, such as marginal, mixed effects and transition models (Diggle et al., 2002). The focus of this work is on a continuous time transition model motivated by a biological problem, associated with the parasitoid wasps *Telenomus posidi* and *Trissolcus basalis*, which useful for pest control in soybean (Bon, 2021). Since the parasitoids can make two successive choices regarding the type of egg (non-parasitised, parasitised by their own species or by the oppos-

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

ing species) and the behaviour after choosing an egg (marking, ovipositing or drumming), the subjects have double transitions. In this context, the multi-state model is very useful to describe the transitions from one state to next, and also to assess the effect of experimental design conditions (Meira-Machado et. al., 2009, Lara et al., 2020). The main goal this work is to present an extension to the multi-state models associated with successive transitions of the parasitoids.

2 Material and Methods

2.1 Motivational study

The data comes from an experiment carried out at the Department of Ecology and Evolutionary Biology, Federal University of São Carlos, Brazil, in 2021. Interactions between parasitoid females of the species *Telenomus posidi* or *Trissolcus basalis* with eggs of the stinkbug *Euschistus heros* took place in experimental arenas, represented by Petri dishes (15×2 cm). A total of 12 eggs were made available to a female parasitoid, divided into 3 groups: 4 eggs previously parasitised by females of *T. podisi*; 4 eggs parasitised by *T. basalis* and 4 healthy eggs (not parasitised by any species) (Figure 1). For the observations, the following behaviours were defined and quantified: a) walking; b) drumming; c) ovipositing and d) marking. Each female was observed for 35 minutes. Ten replicates were performed for each parasitoid species.

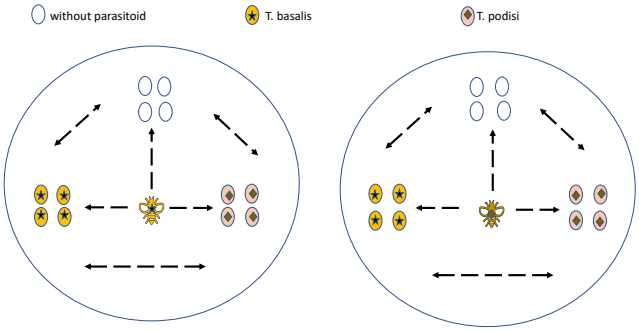


FIGURE 1. Experimental scenarios for quantifying the success of parasitism in the presence of bug eggs previously parasitised by *Trissolcus basalis*, by *Telenomus podisi* and not parasitised in the absence of competition (Adapted from Bin et al., 1993)

2.2 Methods

The methodological procedures are centred on continuous stochastic processes and Generelized Linear Models. We consider two random variables, namely: $\{Y_1(t) \in S_1, t \in \tau\}$, $S_1 = \{1, 2, \ldots, k = 4\}$ for the egg choice, where "4" represents walking or no choice, and $\{[Y_2 \mid Y_1](t) \in S_2, t \in \tau\}$ $[Y_2 \mid Y_1](t) \in S_2, t \in \tau$ $S_2 = \{1, 2, \ldots, k = 4\}$ for choice, where "4" represents return to set S_1, hereafter named "other". In both state sets, we assume the Markov propriety with a finite number of jumps at each time interval and an exponential distribution, i.e., $F_a(t) = 1 - \exp(-q_a t)$ if $t \geq 0$. Then, assuming stationarity of the processes, by means of the Markovian Cox-model, we obtain, through maximum likelihood, the intensities and transition probabilities matrices, in this stochastic double random walk of the parasitoids. We denote these matrices by $\mathbf{P_1(t)}$, $\mathbf{Q_1(t)}$, and $\mathbf{P_2(t)}$, $\mathbf{Q_2(t)}$. The likelihood-ratio test was used to assess the treatment effect. The analyses were carried out using packages `survival` and `msm` available for `R` software (R Core Team, 2022).

3 Results

We begin with a brief initial exploratory data analysis using contingency tables: (1) treatment versus egg choice and (2) treatment versus behaviour. According to the χ^2 test, there is no homogeneity of treatment (parasitoid species) in relation to egg choice ($p < 0.05$), but it is homogeneous with respect to the behaviours ($p = 0.6022$). A total of 684 transitions were observed, both for choosing eggs and for behaviours, without taking into account the treatment structure.

Next, we model egg choice, i.e., the the Y_1 process, using the Cox model. There was a significant effect of treatment (species) in the process of choosing the eggs to be parasitised ($p < 0.05$). The transition probabilities are shown in Figure 2, observing that the species *T. basalis* is less selective in the process of oviposition, with higher probabilities of transition to self-parasitisation.

Finally, we modelled the behaviours given choice, i.e. the $Y_2 \mid Y_1$ processes. Given that the initial choice is the healthy egg, there were a total of 334 behaviour transitions, with 48 transitions to the initial choice. Also, we found no significant effect of the difference in behaviour between species in this condition of initial choice ($p = 0.9438$). Similarly, there was no difference between behaviours when the initial choice was eggs parasitised by *T. basalis* ($p = 0.092$) Nonetheless, if the initial choice was for eggs parasitized by *T.podisi*, there were a total of 141 behaviour transitions, with 29 transitions for new egg choices or walking. The *T.basalis* species was the one that made the most transitions between behaviours and returns to phase 1 of choice, demonstrating their behavioural fragility. For this species, we found significant differences between the behaviors ($p < 0.05$).

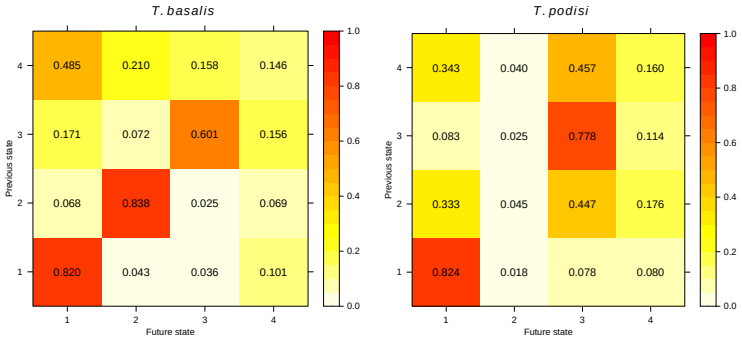


FIGURE 2. Transition probabilities in the process of egg choice by the two species, where 1: healthy egg, 2: egg parasitised by *T.podisi*, 3: egg parasitised by *T.basalis* and 4: walking or no egg choice.

Therefore, in practical terms, through the use of multi-state models, this study reinforces the greater efficiency of *T.podisi* as a biological control agent.

Acknowledgments: Special thanks to Brazilian Fundation, CAPES, process number 88887.716582/2022-00. This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049.

References

Bin, F., Vinson, S.B., Strand, M.R., Colazza, S., Jones, W.A. (1993). Source of an egg kairomone for Trissolcus basalis a parasitoid of Nezara viridula. *JPhysiological Entomology*, **18**, 7 – 15.

Bon, V.J. (2021). *Efeito da interação de Trissolcus basalis e Telenomus podisi (Hymenoptera: Scelionidae) na efetividade do controle biológico de Euschistus heros (Hemiptera: Pentatomidae*. Federal University of São Carlos, Master dissertation.

Diggle, P.J., Heagerty, P.J.;, Liang, K.Y., Zeger, S.L (2002). *Analysis of Longitudinal Data*. New York: Oxford University Press.

Lara, I. A. R.; Moral, R. A.; Taconeli, C. A.; Reigada, C.; Hinde, J. (2020). A generalized transition model for grouped longitudinal categorical data. *Biometrical Journal*, **x**, 1 – 12. DOI: 10.1002/*bimj*.201900394.

Meira-Machado, L., Uña Alvarez, J. de, Cadarso-Suárez, C., Andersen, P.K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, **18.2**, 195 – 222.

R Core Team (2022). `R`: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. URL <http://www.R-project.org>

Bias reduced predictions for black-box models

Philipp Sterzinger[1], Ioannis Kosmidis[1,2]

[1] Department of Statistics, University of Warwick, Coventry, UK
[2] The Alan Turing Institute, London, UK

E-mail for correspondence: philipp.sterzinger@warwick.ac.uk

Abstract: Prediction is a core task in statistics, machine learning and related disciplines. Although predictive models have become increasingly more powerful, they typically exhibit statistical bias – a systematic misrepresentation of the actual prediction point. While such models vary in approach and complexity, ranging from simple linear models, to complex, nonlinear models such as neural networks, the estimation problem that underlies the learning phase of the predictive model can oftentimes be framed as a M-estimation problem. By leveraging the rich statistical literature on M-estimation, we develop a novel approach to improved, first-order unbiased prediction for black-box models that satisfy standard M-estimation regularity conditions. Amongst others, this methodology encompasses the large class of predictive models where training is conducted through optimisation of some loss function (e.g. maximum likelihood estimation or fitting of neural networks). The method's improved predictive performance is illustrated in a simulation study for predicting success probabilities in logistic regression.

Keywords: M-estimation; estimating equations; adjusted score; logistic regression

1 Introduction

Consider the common prediction task where, upon observing a sequence of variables of interest, $y_1, \ldots, y_k$, $y_i \in \mathcal{Y} \subseteq \Re$ and covariates,$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)^{\mathrm{T}}$, $\boldsymbol{x}_i \in \mathcal{X} \subseteq \Re^p$ $(i = 1, \ldots, k)$ and given a new data point $\boldsymbol{x}$, one wishes to make a prediction about some function of the corresponding unobserved data point y. The predictive task is captured by a known parametric function $g(\boldsymbol{x}; \boldsymbol{\theta})$, which takes as inputs the prediction point $\boldsymbol{x}$ and a parameter $\boldsymbol{\theta} \in \Theta \subseteq \Re^p$. Let $\boldsymbol{\theta}_0$ be the unknown parameter that identifies the prediction of interest $g(\boldsymbol{x}; \boldsymbol{\theta}_0)$. Typically, $g(\boldsymbol{x}; \boldsymbol{\theta}_0)$ reflects

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

some parameter of the conditional distribution of an unobserved response y given $\boldsymbol{x}$, but this is immaterial for the development of the bias reduced prediction (BRP) theory. The parameter estimation of $\boldsymbol{\theta}_0$, or learning of the predictive model $g(\boldsymbol{x};\boldsymbol{\theta}_0)$, can oftentimes be framed as a M-estimation problem, that is, the estimator $\hat{\boldsymbol{\theta}}$ is characterized as the root of a system of equations

$$\sum_{i=1}^{k} \psi^i(\boldsymbol{\theta}) = \mathbf{0}_p \,, \tag{1}$$

where $\mathbf{0}_p$ is a p-vector of zeroes, $\psi^i(\boldsymbol{\theta}) = (\psi^i_1(\boldsymbol{\theta}), \ldots, \psi^i_p(\boldsymbol{\theta}))^{\mathrm{T}}$, and $\psi^i(\boldsymbol{\theta}) = \psi(\boldsymbol{\theta}, y_i, \boldsymbol{x}_i)$, $\psi^i_r(\boldsymbol{\theta}) = \psi_r(\boldsymbol{\theta}, y_i, \boldsymbol{x}_i)$, are estimating functions that depend on $\boldsymbol{\theta}$ and the observables $y_i, \boldsymbol{x}_i$ $(i = 1, \ldots, k)$.

2 Prediction as M-estimation

We use the idea of stacking redundant estimators, which themselves do not constitute M-estimators, to define a new M-estimation problem (see Stefanski and Boos, 2002), to reformulate the 2-step prediction as a M-estimation problem by parametrizing the pointwise prediction $\pi_0 = g(\boldsymbol{x};\boldsymbol{\theta}_0)$. Adding the redundant estimating equation $\pi - g(\boldsymbol{x};\boldsymbol{\theta})$ at the observation level to (1) yields the simultaneous estimation problem

$$\begin{aligned} \sum_{i=1}^{k} \psi^i(\boldsymbol{\theta}) &= \mathbf{0}_p \\ k\,\{\pi - g(\boldsymbol{x};\boldsymbol{\theta})\} &= 0 \end{aligned} \,, \tag{2}$$

for $\boldsymbol{\vartheta} = (\boldsymbol{\theta}^{\mathrm{T}}, \pi)^{\mathrm{T}} \in \Re^{p+1}$ with the solution $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\theta}}^{\mathrm{T}}, g(\boldsymbol{x};\hat{\boldsymbol{\theta}}))^{\mathrm{T}}$. Hence, (2) recovers the plug-in prediction for $g(\boldsymbol{x};\hat{\boldsymbol{\theta}})$. Writing $\varphi^i(\boldsymbol{\vartheta}) = (\psi_i(\boldsymbol{\theta})^{\mathrm{T}}, \pi - g(\boldsymbol{x};\boldsymbol{\theta}))^{\mathrm{T}}$ defines a new M-estimation problem that coincides with (2).

3 Bias Reduced prediction

Letting $A(\boldsymbol{\vartheta})$ be the bias-reducing adjustment function of Kosmidis and Lunardon (2022), for (2), one gets the BRP equations

$$\sum_{i=1}^{k} \varphi^i(\boldsymbol{\vartheta}) + A(\boldsymbol{\vartheta}) = \mathbf{0}_{p+1} \,. \tag{3}$$

The solution to (3) yields first order bias unbiased for $g(\boldsymbol{x};\boldsymbol{\theta}_0)$ under the regularity conditions Kosmidis and Lunardon (2022) for (1) and differentiability and continuity conditions on the prediction function $g(\boldsymbol{x};\cdot)$. One can reduce the BRP approach to a 2-step estimation procedure, where in

a first step, the original model tuning parameters $\boldsymbol{\theta}$ are estimated by reduced bias M-estimation (RBM, see Kosmidis and Lunardon, 2022). The solution is then plugged in the BRP function which is defined as the explicit solution of the $(p+1)$th equation of (3) for π given $\boldsymbol{\theta}$. If the first two Bartlett identities are satisfied for the original M-estimation problem, as would for example be the case in a fully specified model where the training phase of (2) is conducted via maximum likelihood estimation, the reduced bias M-estimator of $\boldsymbol{\theta}$ reduces to the adjusted score equations estimator of Firth (1993), which can be of use if the original estimator does not exist (see for example Kosmidis and Firth, 2021).

4 Predicting success probabilities in logistic regression

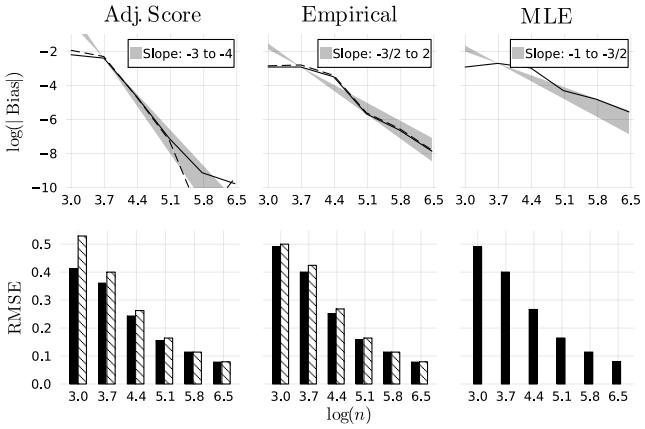


FIGURE 1. Top panels show the log of the absolute Monte Carlo estimates of the prediction bias for BRP methods using adjusted score and empirical adjustment approaches with full (solid line) and refined (dashed line) adjustment functions and the MLE. Grey shaded areas are spanned by lines with slopes of $(-3,-2)$, $(-2,-3/2)$ and $(-3/2,1)$ for adjusted score, empirical and MLE predictions. Bottom panels show the estimated root mean squared prediction error.

We predict success probabilities in logistic regression and compare it to standard maximum likelihood estimation using the simulation setup of Puhr et al. (2017). For $p = 11$, $n = 10 \times 2^i, i = 1, \ldots, 6$, the design matrix $\boldsymbol{X}$ and the prediction point $\boldsymbol{x}$ are held fixed. The parameter vector $\boldsymbol{\beta}_0$ is chosen as in Puhr et al. (2017) to give expected predicted probabilities of $1/2$ for their data generating process. $\boldsymbol{x}$ is chosen to achieve a predicted probability of approximately $1/2$. Finally, given $\boldsymbol{X}, \boldsymbol{x}, \boldsymbol{\beta}_0$ and for each n, about 6.7×10^6 (to enusre that the Monte-Carlo error is two orders of magnitude smaller than the estimated bias with high probability), independent

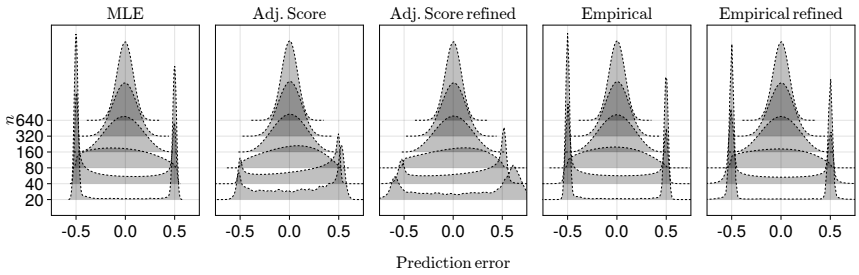


FIGURE 2. Estimated distribution of the centred prediction error $\hat{\pi} - \mu(\beta_0)$ for MLE and BRP based predictions and various sample sizes.

copies of y are drawn and predictions were obtained using four variants of the BRP method as well as the MLE plug-in predictions. Figure 1 shows the decay of the log of the absolute bias for each prediction method as well as the root mean squared error. The BRP methods decay at rates faster than the $\mathcal{O}(n^{-3/2})$ rate guaranteed by theory, whereas the MLE predictor decays at a rate that lies somewhere between n^{-1} and $n^{3/2}$. All BRP methods have substantially lower bias than the MLE while keeping the RMSE at a comparable level. The estimated density plots of the prediction error $\hat{\pi} - \mu(\boldsymbol{\beta}_0)$ distribution in Figure 2, where $\hat{\pi}$ denotes the predicted probability, illustrate that the asymptotic normality of the MLE based predictor is preserved for the bias reduced predictions. The peaks at $-1/2$ and $1/2$ for the MLE and empirical adjustment functions based predictions come from separated datasets, in which the infinite MLE and the RBM-estimates give rise to predictions that are essentially zero or one (see Puhr et al., 2017).

References

Firth, D. Bias Reduction of Maximum Likelihood Estimates. (1993) *Biometrika*, **80(1)**, 27 – 38.

Kosmidis, I. and Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, **108(1)**,71 – 82.

Kosmidis, I. and Lunardon, N. (2022). Empirical bias-reducing adjustments to estimating functions. *arXiv preprint*, 2001.03786.

Puhr, R., Heinze, G., Nold, M., Lusa, L., and Geroldinger, A. (2017). Firth's logistic regression with rare events: accurate effect estimates and predictions?. *Statistics in Medicine*, **36(14)**, 2302 – 2317.

Stefanski, L. A. and Boos, D. D. (2002). The Calculus of M-Estimation. *The American Statistician*, **56(1)**, 29 – 38.

Autoregressive hidden Markov models for high-resolution animal movement data

Ferdinand V. Stoye[1], Roland Langrock[1]

[1] Bielefeld University, Germany

E-mail for correspondence: roland.langrock@uni-bielefeld.de

Abstract: New types of high-resolution (e.g. 1 Hz) animal movement data allow for increasingly comprehensive biological inference, but method development to meet the statistical challenges associated with such data is lagging behind. In this contribution, we develop a new class of hidden Markov models specifically tailored to address the requirements posed by high-resolution movement data, in particular accounting for the very strong serial correlation. The models feature autoregressive components of general order in both the step length and the turning angle variable, with lasso-based automated order selection.

Keywords: circular statistics; lasso penalty; time series

1 Introduction

High-resolution movement data, e.g. with sampling at 1 Hz, holds vast potentials for ecological inference: behavioural modes and highly agile manoeuvres such as foraging attempts can be more accurately identified, social interactions and predator-prey encounters can be measured, and the effects of environmental stimuli can reliably be estimated (Nathan et al., 2022). Figure 1 displays one such time series, showing step lengths derived from the track of a tern species, a surface foraging seabird, observed at 30 Hz (Lieber et al., 2023). Short-lived foraging manoeuvres like hovering or shallow plunge-dives, as indicated by troughs in the time series, can be identified and further analysed using HMMs, assuming that each observation is associated with an underlying behavioural mode assumed to be generated from an N–state Markov chain (McClintock et al., 2020).
However, a basic HMM would assume the observations within states to be *uncorrelated*, which is not realistic for the data shown in Figure 1, where the individual does indeed seem to switch between two (or more) behavioural modes, but where within a behavioural mode, the step lengths are clearly not independent over time, as they vary gradually. Basic HMM formulations

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

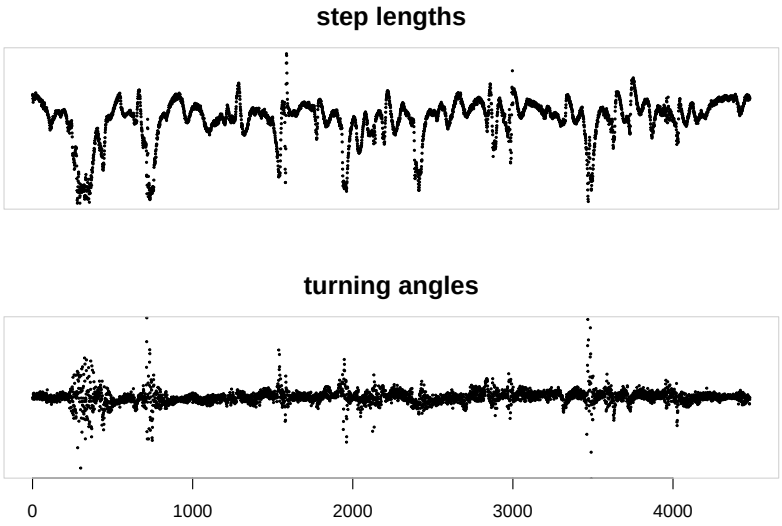


FIGURE 1. Bivariate time series of step lengths and turning angles of a tracked tern (seabird) using aerial drone observations at 30 Hz (units are omitted).

do not acknowledge such strong within-state serial correlation, which can have undesirable consequences for the associated statistical inference.
In this contribution, we develop a class of HMMs incorporating within-state autoregressions specifically designed to meet the requirements of high-resolution animal movement data. Our work extends a similar endeavour by Lawler et al. (2019) in four ways: i) not only the step lengths, but also the turning angles are modelled using autoregression; ii) the step length distribution is modelled assuming a constant coefficient of variation; iii) the autoregressive component is modelled using general lag p; iv) lasso regularisation is used to automatically select the order of the autoregression.

2 Model formulation

In movement ecology, the observed sequence of movement metrics — in many cases the step lengths and turning angles between successive locations — is commonly modelled conditional on an underlying, non-observable sequence of states (typically interpreted as proxies of behavioural modes such as resting, foraging, or travelling). The corresponding class of HMMs involves i) an N–state Markov chain $\{S_t\}_{t=1,\dots,T}$, here for simplicity assumed to be homogeneous, defined by its initial state distribution $\boldsymbol{\delta} = \big(\Pr(S_1 = 1, \dots, S_1 = N)\big)$ and transition probability matrix $\boldsymbol{\Gamma} = (\gamma_{ij})$,

and ii) suitable state-dependent distributions for the observed movement metrics (Zucchini et al., 2016).
We consider the bivariate time series $\{\mathbf{X}_t\}_{t=1,\ldots,T}$, $\mathbf{X}_t = (X_t^{\text{step}}, X_t^{\text{turn}})$, with X_t^{step} the step length and X_t^{turn} the turning angle at time t. We assume the step lengths to follow a state-dependent gamma distribution,

$$X_t^{\text{step}} \mid S_t = j \sim \Gamma\left(\mu_{t,j}, \sigma_{t,j}\right), \tag{1}$$

where the state-dependent mean $\mu_{t,j}$ fluctuates around a steady-state mean μ_j according to an AR(p_j) process:

$$\mu_{t,j} = \sum_{k=1}^{p_j} \phi_{j,k}^{\text{step}} x_{t-k}^{\text{step}} + \Big(1 - \sum_{k=1}^{p_j} \phi_{j,k}^{\text{step}}\Big)\mu_j.$$

The state-dependent standard deviation $\sigma_{t,j}$ is calculated as $\sigma_{t,j} = \omega_j \mu_{t,j}$, with the constant coefficient of variation ω_j a parameter to be estimated. For the turning angles, the circular nature of the variable needs to be taken into account. This is achieved assuming a von Mises state-dependent distribution,

$$X_t^{\text{turn}} \mid S_t = j \sim \text{von Mises}\left(\mu_{t,j}, \kappa_j\right), \tag{2}$$

formulating an AR(p_j) process on the mappings of the turning angles to their corresponding values on the unit circle:

$$\mu_{t,j} = \text{Arg}\Bigg(\sum_{k=1}^{p_j} \phi_{j,k}^{\text{turn}} \exp\left(i\, x_{t-k}^{\text{turn}}\right) + \Big(1 - \sum_{k=1}^{p_j} \phi_{j,k}^{\text{turn}}\Big) \exp\left(i\, \mu_j\right)\Bigg).$$

For both the step lengths and the turning angles, $\phi_{j,k}$ are the within-state autoregressive parameters for state j and time lag k.
Parameter estimation is conducted by optimising the conditional likelihood, for each state j conditioning on the first p_j observations. In order to automate the choice of the state-dependent autoregressive order p_j, we additionally include a lasso penalty on the autoregressive coefficients, resulting in the partially penalised conditional likelihood

$$\begin{aligned}\mathcal{L} = {} & \boldsymbol{\delta}\mathbf{P}(x_1^{\text{step}}, x_1^{\text{turn}})\boldsymbol{\Gamma}\mathbf{P}(x_2^{\text{step}}, x_2^{\text{turn}}) \cdot \ldots \cdot \boldsymbol{\Gamma}\mathbf{P}(x_T^{\text{step}}, x_T^{\text{turn}})\mathbf{1}^t \\ & - \lambda\Bigg(\sum_{j=1}^{N}\sum_{k=1}^{p_j} |\phi_{j,k}^{\text{step}}| + \sum_{j=1}^{N}\sum_{k=1}^{p_j} |\phi_{j,k}^{\text{turn}}|\Bigg),\end{aligned} \tag{3}$$

with a complexity penalty $\lambda \geq 0$. The diagonal matrices $\mathbf{P}(x_t^{\text{step}}, x_t^{\text{turn}})$ comprise the products of the state-dependent gamma and von Mises densities as implied by (1) and (2) on the diagonal, for each state j replacing the initial p_j state-dependent densities by ones. Following Oelker and Tutz (2017), to obtain a differentiable objective function the L_1 norm $\|\cdot\|$ in the penalty is approximated by $\sqrt{(\cdot + \epsilon)^2}$, with a small positive ϵ.

3 Case study

We consider the tern movement data already shown in Figure 1, downsampled to 1 Hz to avoid numerical instability. All used data is published in Lieber et al. (2022). In our analyses we use the bird with ID *Tern-h2-51*. The model described above, with $N = 2$ states, was fitted using maximisation of the partially penalised conditional likelihood (3), allowing for a maximum autoregression lag of 5 in each state. Figure 2 displays the trajectories of the autoregression coefficients $\phi_{j,k}$ for increasing complexity penalty λ, indicating successful lasso-type variable selection. While the AIC selects a model featuring autoregressive terms for each state and variable, the BIC favours a model with autoregressive terms in state 2 only. The latter makes intuitive sense as state 2 is associated with foraging manoeuvres and high tortuosity, which requires the autoregressive component for capturing the current curvature.

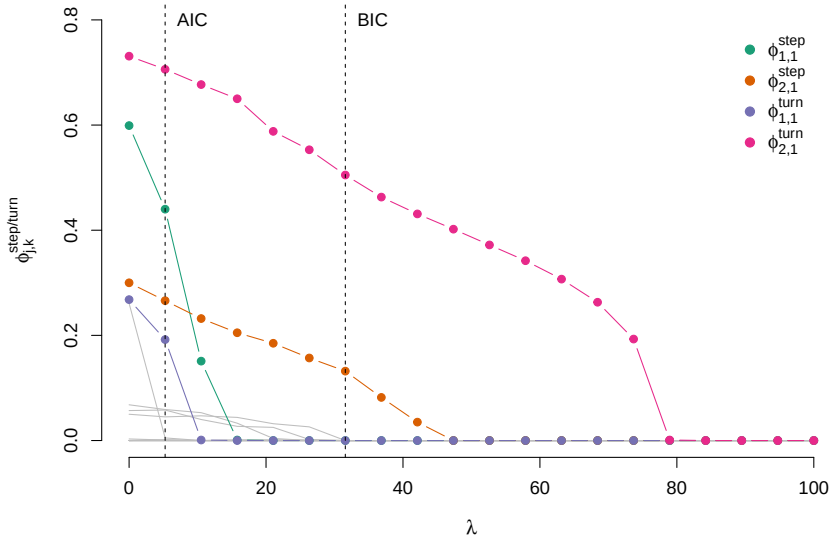


FIGURE 2. Trajectories of the autoregression coefficients for increasing complexity penalty $\lambda > 0$. For clarity, the legend comprises only those coefficients that based on information criteria should be included in the model and differ substantially from zero.

To further illustrate model adequacy, Figure 3 shows the original movement track (top) as well as simulated tracks from a basic HMM (bottom left) and the autoregressive HMM selected by the BIC (bottom right). The latter seems to be able to produce more pronounced circles than a standard model formulation, leading to a visually better model fit when compared to the real data.

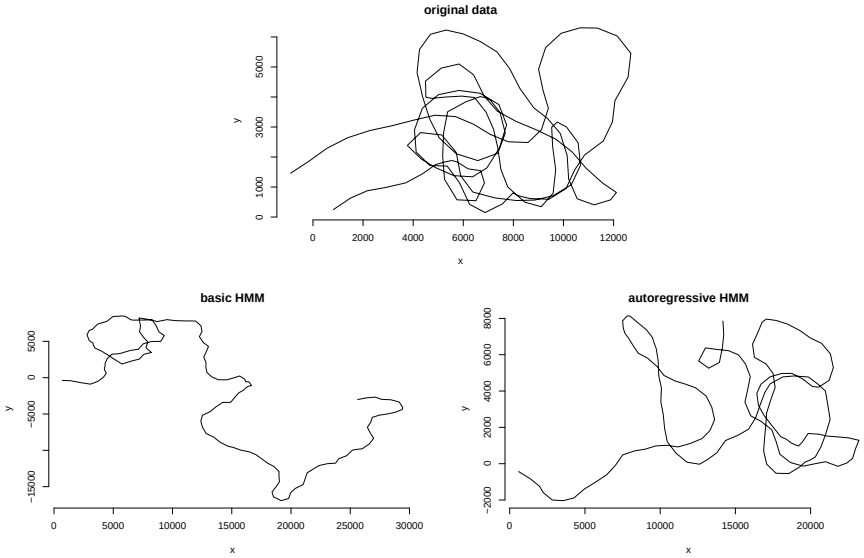


FIGURE 3. Comparison of real data (top) to data simulated from a basic HMM fitted to the data (bottom left) and an autoregressive HMM (bottom right). In the latter case, the choice of $\lambda \approx 31.6$ corresponds to the best-scoring model regarding BIC (cf. Figure 2).

Acknowledgments: The authors are grateful to Dr. Lilian Lieber for her helpful comments on an earlier version of this paper.

References

Lawler, E. et al. (2019). The conditionally autoregressive hidden Markov model (carHMM): Inferring behavioural states from animal tracking data exhibiting conditional autocorrelation. *Journal of Agricultural, Biological and Environmental Statistics*, **24**, 651 – 668.

Lieber, L. et al. (2022). Data from: Selective foraging behavior of seabirds in small-scale slicks. DOI: 10.6084/m9.figshare.20496957.v1.

Lieber, L. et al. (2023). Selective foraging behavior of seabirds in small-scale slicks. *Limnology and Oceanography Letters*, **8.2**, 286 – 294.

McClintock, B.T. et al. (2020). Uncovering ecological state dynamics with hidden Markov models. *Ecology Letters*, **23**, 1878 – 1903.

Nathan, R. et al. (2022). Big-data approaches enable increased understanding of animal movement ecology. *Science*, **375**, eabg1780.

Oelker M.-R. and Tutz G. (2017). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification*, **11**, 97 – 120.

Zucchini, W., MacDonald, I.L. and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*, Chapman & Hall/CRC Press.

Complexity reduction via deselection for boosting distributional copula regression

Annika Strömer[1], Nadja Klein[2], Christian Staerk[1], Hannah Klinkhammer[1], 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: Boosting distributional copula regression is a flexible tool to jointly model multivariate outcomes, in which all parameters of the joint distribution can be related to covariates via additive predictors. Estimation via model-based boosting allows to fit these complex models also to high-dimensional data ($p > n$). Additionally, boosting can incorporate data-driven variable selection simultaneously for all parameters of the marginal distributions as well as for the association parameter of the copula. However, as known from univariate (distributional) regression models, the boosting algorithm tends to select too many variables, particularly for low-dimensional settings ($p < n$). To counteract this behavior, we adapt a recent deselection approach for statistical boosting to multivariate copula regression models to deselect base-learners with only a negligible impact on the overall performance of the model. We illustrate our approach by jointly modelling LDL and HDL cholesterol based on large UK Biobank genotype data.

Keywords: Model-based boosting; Variable selection; GAMLSS; Copula regression.

1 Introduction

In distributional copula regression, potentially different marginal response distributions can be combined by an appropriate copula function that defines the dependency structure between the outcomes for multivariate modelling. Within the framework of generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005), all parameters

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

of the distributional copula regression model (i.e. the distribution parameters of the marginals and the dependency parameter) are modelled by an additive predictor incorporating different effect types for the covariates (Klein and Kneib, 2016). In combination with component-wise gradient boosting, we can incorporate data-driven variable selection for potentially high-dimensional data, which is controlled by the number of boosting iterations (Mayr et al., 2012). However, despite these advantages, the boosting algorithm still tends to select too many variables (including ones which are non-informative or have a very low signal), which occurs particularly for low-dimensional settings. In these situations, we can observe a slow overfitting behavior, which results in a later stopping of the algorithm and therefore a larger set of included base-learners that might have only minor importance. As a result, we are faced with an unnecessary large model, that might be performing good for prediction but is difficult to interpret.

2 Deselection of base-learners

We address this issue by adapting the deselection approach by Strömer et al. (2022) for boosting distributional copula regression. The pragmatic and simple idea is to start with a classical boosted model tuned by cross-validation or resampling techniques to determine the optimal stopping iteration m_{stop} to achieve high prediction accuracy. Then, the base-learners and variables that were selected but only have a minor impact on the model are identified and are deselected. Afterwards, the model is boosted again only with the remaining ones. The importance of a base-learner j in the deselection approach is measured via the risk reduction after m_{stop} iterations:

$$R_j = \sum_{m=1}^{m_{\text{stop}}} I(j = j^{*[m]})(r^{[m-1]} - r^{[m]}), \quad j = 1, \ldots, \sum p_k,$$

where I denotes the indicator function and $j^{*[m]}$ is the selected base-learner in iteration m. Furthermore, $r^{[m-1]} - r^{[m]}$ represents the risk reduction in iteration m, for risks $r^{[m]}$ and $r^{[m-1]}$ at iterations m and $m-1$. Note that in the case of distributional copula regression, all distribution parameters are considered together and each parameter $\theta_k, k = 1, \ldots, K$ may depend on a different number of variables p_k. For a given threshold $\tau \in (0, 1)$, we deselect base-learner j if

$$R_j < \tau \cdot (r^{[0]} - r^{[m_{\text{stop}}]}),$$

where $r^{[0]} - r^{[m_{\text{stop}}]}$ represents the total risk reduction and R_j denotes the attributable risk reduction of base-learner j. In other words, only base-learners which contribution R_j to the total risk reduction is larger than the relative τ threshold (e.g., 1%, Strömer et al, 2022) will remain in the model after the deselection step.

3 Simulations for comparison with competitors

We conducted a simulation study (based on a similar set-up as in Hans et al., 2023) to investigate and compare the variable selection properties, the predictive performance and the computation time of the classical boosting algorithm with the adapted deselection approach. As additional competitors, we also considered stability selection (Meinshausen and Bühlmann, 2010) and probing (Thomas et al., 2017) to benchmark our results. For a low-dimensional setting, the classical boosted model selected many non-informative variables for every distribution parameter. All approaches effectively reduced the number of false positives. Probing and stability selection did not select all informative variables in each simulation run, whereas the deselection approach maintained all informative variables in the model. In a high-dimensional setting, fewer non-informative variables were included in the boosted models. The approaches performed similar as in the low-dimensional setting and reduced the number of selected non-informative variables almost completely.
A comparison of the negative log-likelihood for the low-dimensional and high-dimensional setting showed that stability selection and deselection resulted in a slightly better predictive performance than the classical boosted model. Probing, on the other hand, led to a lower predictive performance. In terms of computation time, probing is the fastest and stability selection takes much more computational resources than the classical boosted model or the deselection approach.

4 Joint modelling of LDL and HDL cholesterol

We illustrate our deselection approach on high-dimensional genomic cohort data from the UK Biobank, modelling the joint genetic predisposition for two continuous phenotypes, LDL and HDL cholesterol, in dependence of different genetic variants. For both phenotypes, the 1000 variants (typically single nucleotide polymorphisms) with the largest marginal associations with each of the two phenotypes were selected in a pre-screening process. Overall, the data set includes 20,000 sampled observations and 1,179 variants (803 variants selected for both phenotypes). The log-logistic distribution was considered as marginal distribution for both phenotypes and the Gumbel copula was used for modelling the dependency structure based on the comparison of the predictive risk. All variants were included with simple linear models as base-learners.
Figure 1 illustrates the resulting estimated absolute coefficients for every distribution parameter (similar to Manhattan plots). The classical boosted model selected several variants for each distribution parameter. After the deselection approach with $\tau = 0.01$, only some variants for μ_1 and μ_2 are left. This means that with the deselection approach we can not only reduce

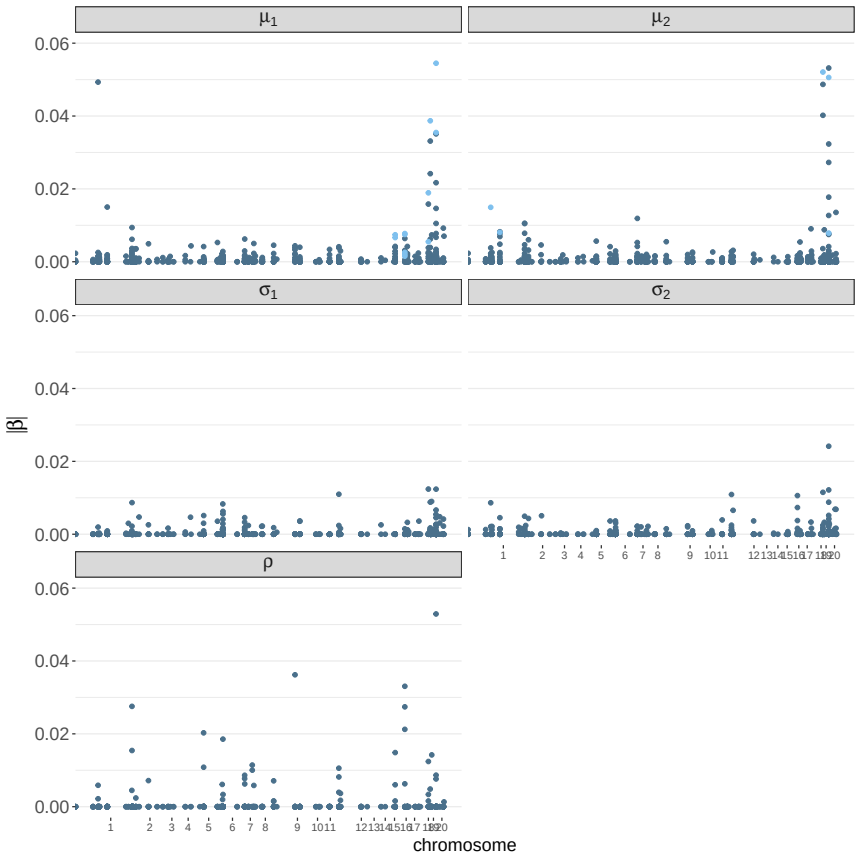


FIGURE 1. Manhattan-type plots (chromosomes on x-axis) for the absolute coefficients of boosted copula regression for the joint analysis of LDL and HDL cholesterol. The dark blue dots are the selected variants by classical boosting, the lighter blue points are the remaining variables after the deselection approach.

the included variables and obtain a much sparser model with a potentially simpler interpretation: In this case the approach also further reduces the overall complexity by completely deselecting all variants of distribution parameters resulting in two simple univariate models for both phenotypes. Furthermore, the deselection leads to a comparable predictive performance on test data as the classical boosted model.

5 Conclusion

We presented a pragmatic deselection approach for boosting multivariate distributional copula regression models. The new deselection approach re-

sults in much sparser models and can even lead to more simple univariate regression models, reducing the complexity of the overall analysis. The prediction accuracy usually does not improve but can lead to comparable accuracy as the classical boosted model with less predictors. Consequently, the interpretability of resulting prediction models is improved.
The presented deselection procedure is controlled via a threshold value τ, which represents the minimum amount of total risk reduction which should be attributed to a corresponding base-learner in order to avoid deselection. This can be interpreted as a threshold-value for the importance of the particular predictor variable. In the simulation study, a threshold of $\tau = 0.01$ (i.e. 1% of total risk reduction) was considered to be appropriate. However, depending on the research question and the context of the problem, the choice of τ is a trade-off between more complex models with the highest prediction accuracy and a sparser, more interpretable model with potentially reduced prediction accuracy.

Acknowledgments: The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number 428239776).

References

Hans, N., Klein, N., Faschingbauer, F., Schneider, M. and Mayr, A. (2022). Boosting distributional copula regression. *Biometrics*, **00**: 1 – 13.

Klein, N. and Kneib, T. (2016). Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Statistics and Computing*, **26** (4), 841 – 860.

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** (3): 403 – 427.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4): 417 – 473.

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

Strömer A., Staerk C., Klein N., Weinhold L., Titze S. and Mayr A. (2022). Deselection of base-learners for statistical boosting — with an application to distributional regression. *Statistical Methods in Medical Research*, **31**(2): 207 – 224.

Thomas, J., Hepp, T., Mayr, A. and Bischl, B. (2017). Probing for sparse and fast variable selection with model-based boosting. *Computational and Mathematical Methods in Medicine*, **2017** 1 – 8.

Bayesian nowcasting with Laplacian-P-splines

Bryan Sumalinab[1,2], Oswaldo Gressani[1], Niel Hens[1,3], Christel Faes[1]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Data Science Institute (DSI), Hasselt University, Hasselt, Belgium
[2] Department of Mathematics and Statistics, College of Science and Mathematics, Mindanao State University - Iligan Institute of Technology, Iligan City, Philippines
[3] Centre for Health Economic Research and Modelling Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute, Antwerp University, Antwerp, Belgium

E-mail for correspondence: bryan.sumalinab@uhasselt.be

Abstract: During an epidemic, the daily number of reported infected cases or deaths is often lower than the actual number due to reporting delays. Nowcasting aims to estimate the cases that have not yet been reported and combine it with the already reported cases to obtain an estimate of the daily cases. In this paper, we present a fast and flexible Bayesian approach to nowcasting combining P-splines and Laplace approximations. The main benefit of Laplacian-P-splines (LPS) is the flexibility and faster computation time compared to Markov chain Monte Carlo (MCMC) algorithms that are often used for Bayesian inference. In addition, it is natural to quantify the prediction uncertainty with LPS in the Bayesian framework, and hence prediction intervals are easily obtained. Model performance is assessed through simulations, and the method is applied to the COVID-19 mortality and incidence cases in Belgium.

Keywords: Nowcasting; Laplacian-P-splines; Epidemic; COVID-19; Reporting delay

1 Introduction

Nowcasting is a term used for estimating the occurred-but-not-yet-reported-events (Donker et al. (2011); van de Kassteele et al. (2019)). In epidemiology, real-time updates of new symptomatic/infected individuals are helpful to assess the present situation and provide recommendations

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

for rapid planning and for implementing essential measures to contain an epidemic outbreak. The exact number of new daily cases is frequently subject to reporting delays, resulting in underreporting of the real number of infected individuals for that day. The main goal of nowcasting is to estimate the actual number of new cases by combining the (predicted) not-yet-reported-cases with the already reported cases. In 2019, van de Kassteele et al. proposed a nowcasting model in which the number of cases are structured in a two-dimensional table (with calendar time as the first dimension and delay time as the second dimension), yielding the data matrix used as an input in the model. The reporting intensity is assumed to be a smooth surface and is modelled using two-dimensional P-splines.
In this paper, we work on the method of van de Kassteele et al. (2019) by proposing a new nowcasting methodology based on Laplacian-P-splines (LPS) in a fully Bayesian framework. We build on the work of Gressani and Lambert (2021) to extend the LPS methodology to nowcasting. This approach combines the flexibility of P-splines (Eilers and Marx (1996)) and faster computational time (compared to MCMC approaches) induced by Laplace approximations. Therefore, given this computational benefit, it can be a helpful tool in the daily monitoring of new cases during an epidemic period. To evaluate the (predictive) performance of our method, a simulation study is implemented, and performance measures are reported such as the mean absolute percentage error (MAPE) and prediction interval coverage. Finally, we apply our method to the COVID-19 incidence cases and mortality data in Belgium.

2 Methodology

Let $y_{t,d}$ denote the number of cases (infections or deaths) that occurred at time $t = 1, 2, ..., T$ (corresponding to the calendar day) and reported with a delay of $d = 0, 1, 2..., D$ days. The information on cases can be summarized in matrix form (denoted by Y) with rows as the time dimension and columns corresponding to the delay. The not-yet-reported cases correspond to (t, d) combinations satisfying $t > T - d$. The main objective is to predict the total number of cases, $y_t = \sum_{d=0}^{D} y_{t,d}$, for $t = T - (D - 1), \ldots, T$ for which the nowcasted and already reported cases can be combined.
Let $\mathcal{D} := \boldsymbol{y} = (y_1, y_2 \ldots, y_n)'$ denote the vector of the observed number of cases by stacking the columns of matrix Y for the reported cases, where each entry corresponds to each (t, d) combination of reported cases $y_{t,d}$. The model assumes that the number of cases either follows a Poisson or negative binomial (NB) distribution with mean $\mu_{t,d} > 0$. Following van de Kassteele et al. (2019), the (log) mean number of cases is modeled with two dimensional B-splines:

$$\log(\mu_{t,d}) = \beta_0 + \sum_{j=1}^{K_T} \sum_{k=1}^{K_D} \theta_{j,k} b_j(t) b_k(d) + \sum_{l=1}^{p} \beta_l z_l(t, d),$$

where β_0 is the intercept; $b_j(\cdot)$ and $b_k(\cdot)$ are univariate B-splines basis functions specified in the time and delay dimensions, respectively; and $z_l(t,d)$ represents additional covariates with regression coefficients β_l. In matrix notation:

$$\log(\boldsymbol{\mu}) = B\boldsymbol{\theta} + Z\boldsymbol{\beta}, \tag{1}$$

where $B = B_D \otimes B_T$ is the Kronecker product of B-splines matrices B_T (time) and B_D (delay), Z is the design matrix (including intercept) for additional covariates, and vectors $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are the associated parameters to be estimated. Let $D_t = D_t^m$ and $D_d = D_d^m$ denote the mth order row-wise and column-wise difference matrix. Define the penalty matrices $P_t = D_t^{'} D_t + \delta I_{K_T}$ and $P_d = D_d^{'} D_d + \delta I_{K_D}$, where δ is a small number (say $\delta = 10^{-6}$), to ensure that P_t and P_d are full rank and invertible. To formulate the Bayesian model, we assumed Gaussian priors on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ (Lang and Brezger, 2004), that is, $\boldsymbol{\beta} \sim N(\mathbf{0}, V_{\boldsymbol{\beta}}^{-1})$ and $(\boldsymbol{\theta}|\boldsymbol{\lambda}) \sim N(\mathbf{0}, \mathcal{P}^{-1}(\boldsymbol{\lambda}))$ with $V_{\boldsymbol{\beta}} = \zeta I_{p+1}$ ($\text{small}\,\zeta, \text{e.g.}\,\zeta = 10^{-5}$), $\boldsymbol{\lambda} = (\lambda_t, \lambda_d)'$ is the penalty vector that controls the roughness of the fit and $\mathcal{P}(\boldsymbol{\lambda}) = \lambda_t(I_{K_D} \otimes P_t) + \lambda_d(P_d \otimes I_{K_T})$ the global penalty matrix. Furthermore, denote by $X = (B, Z)$ the global design matrix, $\boldsymbol{\xi} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}$ the latent parameter vector and $Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}} = \begin{bmatrix} V_{\boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \mathcal{P}(\boldsymbol{\lambda}) \end{bmatrix}$ the precision matrix for $\boldsymbol{\xi}$. The full Bayesian (negative binomial) model is then summarized as follows:

$$\begin{aligned}
&(y_i|\boldsymbol{\xi}) \sim \text{NB}(\mu_i, \phi) \;\text{ with }\; \log(\boldsymbol{\mu}) = X\boldsymbol{\xi}, \\
&(\boldsymbol{\xi}|\boldsymbol{\lambda}) \sim \mathcal{N}_{dim(\boldsymbol{\xi})}(\mathbf{0}, (Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}), \\
&(\lambda_t|\delta_t) \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu\delta_t}{2}\right), \\
&(\lambda_d|\delta_d) \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu\delta_d}{2}\right), \\
&\delta_t \sim \mathcal{G}(a_\delta, b_\delta), \\
&\delta_d \sim \mathcal{G}(a_\delta, b_\delta), \\
&\phi \sim \mathcal{G}(a_\phi, b_\phi),
\end{aligned}$$

where ϕ is an overdispersion parameter and $\mathcal{G}(\cdot)$ denotes the Gamma density. The posterior density of $\boldsymbol{\xi}$ conditional on the penalty vector $\boldsymbol{\lambda}$ is approximated by a Gaussian density, denoted by $\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda}, \mathcal{D}) = \mathcal{N}(\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}, \widehat{\Sigma}_{\boldsymbol{\lambda}})$, using a Newton-Raphson algorithm. Let $\boldsymbol{\eta} = (\lambda_t, \lambda_d, \delta_t, \delta_d)^{\mathrm{T}}$ denote the vector of hyperparameters. Following Rue et al. (2009), the marginal posterior of $\boldsymbol{\eta}$ can be approximated as $\widetilde{p}(\boldsymbol{\eta}|\mathcal{D}) \propto \left.\frac{\mathcal{L}(\boldsymbol{\xi};\mathcal{D})p(\boldsymbol{\xi}|\boldsymbol{\eta})p(\boldsymbol{\eta})}{\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\eta},\mathcal{D})}\right|_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}}$ where $\mathcal{L}(\boldsymbol{\xi};\mathcal{D})$ denotes the likelihood function. The posterior mode (obtained via Newton-Raphson) is then used as a point estimate for the penalty vector.

3 Results

A simulation study is implemented in order to evaluate the predictive performance of the proposed method and several nowcast dates are considered. The results of the simulation for the negative binomial model show that the mean absolute percentage error (MAPE) on the nowcast day typically falls within the range of 30% to 40%. This is reasonable considering that we do not have data available on the nowcast day because all cases have a delay of at least one day. Furthermore, the prediction interval coverage ranges from 90% to 94%, which is close to the 95% nominal level. In the case of the Poisson model, the MAPE results are comparable to those of the negative binomial model. However, the prediction interval coverage is lower, with an accompanying narrower interval width, ranging approximately from 20% to 60%. This is because the Poisson model tends to underestimate the variability when there is overdispersion in the data.
Moreover, we apply our method to COVID-19 mortality data in Belgium. The nowcast predictions are fairly close to the observed cases for both Poisson and negative binomial models. In addition, all the observed cases fall within the prediction interval. The prediction interval is wider for the negative binomial model. Figure 1 shows the nowcast plot for the mortality data with different nowcast dates using the negative binomial model. For the COVID-19 incidence data, which involves a higher number of cases, the Poisson assumption failed to contain the observed incidence within the prediction interval, as opposed to the negative binomial model. This aligns with our simulation results, demonstrating that the Poisson assumption indeed yields a narrower interval.

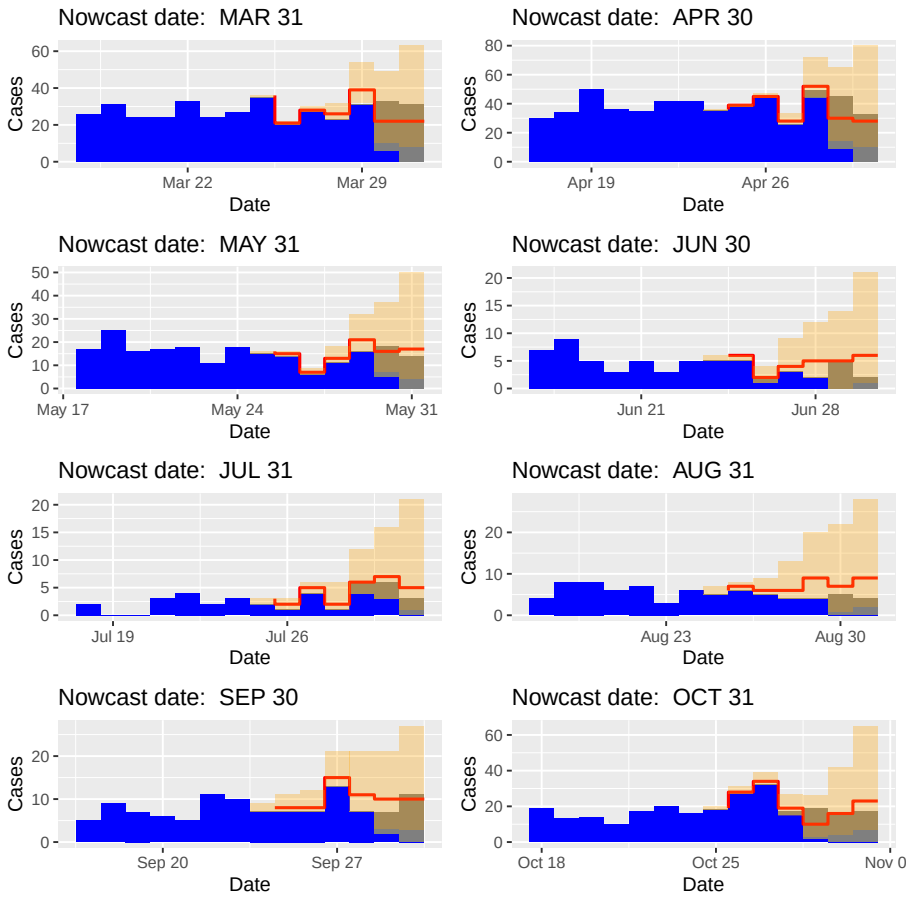


FIGURE 1. Nowcast plot for mortality data with different nowcast dates using negative binomial model. Blue - reported cases ; Gray - Not-yet-reported cases; Orange - nowcast with prediction interval.

Acknowledgments: The authors would like to thank the European Union Research and Innovation Action under the H2020 work programme EpiPose (grant number 101003688).

References

Donker, T., van Boven, M., van Ballegooijen, W. M., van't Klooster, T. M., Wielders, C. C., and Wallinga, J. (2011). Nowcasting pandemic influenza A/H1N1 2009 hospitalizations in the Netherlands. *European Journal of Epidemiology*, **26(3)**, 195–201.

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

Gressani, O. and Lambert, P. (2021). Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines. *Computational Statistics & Data Analysis*, **154**, 107088.

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

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

Sumalinab, B., Gressani, O., Hens, N. and Faes, C. (2022). Bayesian nowcasting with Laplacian-P-splines. *medRxiv*, 2022-08.

van de Kassteele, J., Eilers, P. H. C., and Wallinga, J. (2019). Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing. *Epidemiology*, **30(5)**, 737.

Boosting distributional soft regression trees

Nikolaus Umlauf[1], Johannes Seiler[1],
Mattias Wetscher[1], Nadja Klein[2]

[1] Universität Innsbruck, Austria
[2] Research Center Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technische Universität Dortmund), Germany

E-mail for correspondence: Nikolaus.Umlauf@uibk.ac.at

Abstract: Distributional soft trees offer a flexible and effective way to model full probabilistic regression models. On the one hand, unlike classical regression trees and forests, which use hard splits to partition data, soft trees provide smooth estimates through soft splits, leading to improved performance in many cases due to reduced approximation error. On the other hand, compared to structured additive distributional regression, distributional soft trees allow for more complex interactions of possibly high-dimensional feature vectors. In this article, we introduce a boosted version of a distributional adaptive soft regression tree that can be applied to very large datasets while performing variable selection on the fly. We demonstrate the strong predictive capabilities of this method through a complex regression problem involving the spatial mapping of recent child anaemia risk data in sub-Saharan Africa. Our results further highlight the potential of the proposed boosting method in large-scale complex regression problems.

Keywords: Boosting; GAMLSS; soft trees; variable selection.

1 Introduction

Distributional regression involves modeling the entire distribution of a response variable, rather than just its mean or median. This can provide a more comprehensive understanding of the relationship between covariates and the response, as well as enable more accurate probabilistic predictions beyond point forecasts. While there are various methods for obtaining a distributional model, this paper focuses on the class of structured additive distributional regression, also known as generalized additive models for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos, 2005). GAMLSS can model every parameter of an arbitrary parametric target

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

distribution through input features, resulting in a probabilistic prediction model.

Distributional (adaptive) soft regression trees (DAdaSoRT; Umlauf and Klein, 2022) offer a flexible and effective way to model full probabilistic regression models, and have recently been shown to be a promising alternative to structured distributional methods. One key advantage of these DAdaSoRT, which embed classical soft trees into the distributional framework of GAMLSS, is the smoothness of their estimates on respective distributional parameters, which is achieved through the use of soft splits rather than hard splits. This smoothness can reduce approximation error and improve performance in many cases. In fact, DAdaSoRT have been shown to outperform both classical GAMLSS and full probabilistic distributional forests (DF; Schlosser et al., 2019) in certain situations. Figure 1 shows a simple

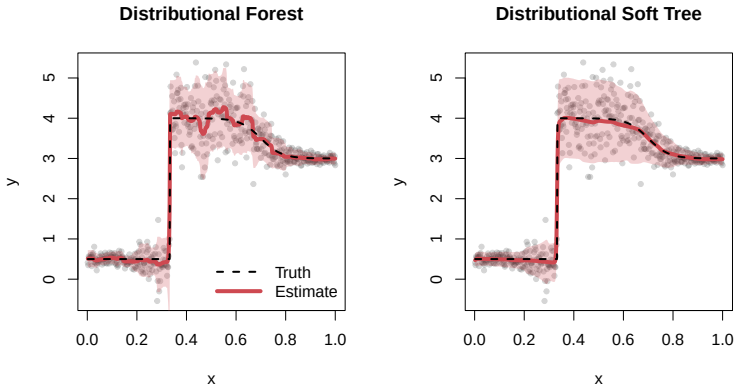


FIGURE 1. Simulated 2D data: Shown are the estimates for $E(Y|x)$ of a DF using 2000 trees (left) and a DAdaSoRT (right). The solid red lines represent the mean estimates, and the red shaded areas depict the 5% and 95% estimated quantiles of $E(Y|x)$. The dashed black lines show the true mean function.

2D regression example with a classical DF compared to a DAdaSoRT. Although DF is estimated with 2000 trees, the resulting estimate is quite wiggly and tends to overfit the data compared to the DAdaSoRT in the right panel of Figure 1. As mentioned before, the reason for this is mainly the hard splitting rule of classical trees and forests, which favors an approximation error that can even be amplified when modeling high-dimensional covariate interactions. The example also illustrates that DAdaSoRT can represent both smooth transitions and abrupt jumps of a function.

This article presents a new boosting algorithm designed to further improve the flexibility of distributional modeling using soft trees. Compared to the estimation method of Umlauf and Klein (2022), the boosting algorithm needs far less tuning, can be applied to very large data sets and selects the

most relevant features in the data on the fly . The latter capability is particularly useful as it helps to reduce the complexity of the modeling process, favours sparse and thus often better interpretable models and improves the accuracy of the results.

2 Model and Boosting Algorithm

DAdaSoRTs are introduced in Umlauf and Klein (2022), and we follow their notation for simplicity. Now, suppose there is data $\mathbf{y} = (y_1, \ldots, y_n)^\top$, such that for each output y_i, $i = 1, \ldots, n$ there is a q-dimensional feature vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^\top$ available and $\mathbf{X} = (\mathbf{x}_1 | \ldots | \mathbf{x}_n)^\top$ is the $n \times q$ feature matrix. Assume $y \sim D_y\left(h_1(\theta_1) = \eta_1,\ h_2(\theta_2) = \eta_2, \ldots, h_K(\theta_K) = \eta_K\right)$, where D_y denotes a parametric distribution for the response variable y and $h_k(\cdot)$ are monotonic and twice differentiable link functions mapping to the distributional predictors η_k which are modeled by soft trees. Following Umlauf and Klein (2022), we use adaptive soft tree structures given by $\boldsymbol{\eta}_k \equiv f_k(\mathbf{X}) = \beta_{k,0} + \sum_{j=1}^{J_k} P_{k,j}(\mathbf{X}, \boldsymbol{\Omega}_{(k,j)})\beta_{k,j}$,where for $k = 1, \ldots, K$, $P_{k,j}(\cdot)$ represent the path probabilities from a soft splitting rule, $\boldsymbol{\Omega}_{(k,j)}$, $\beta_{k,0}$ and $\beta_{k,j}$ are weights that need to be estimated and J_k is the number of "basis functions" of $f_k(\cdot)$ obtained from left and right soft splitting.
In contrast to the multivariate soft splitting of Umlauf and Klein (2022), we use an univariate soft split for $P_{k,j}(\cdot)$, to automatically incorporate variable selection in the final DAdaSoRT by selecting only the best performing feature $\mathbf{x}_q$ according to the current log-likelihood contribution. Specifically, to set up a boosting type algorithm, we specify at each iteration $t = 0, \ldots, T$ and for each distributional predictor $\boldsymbol{\eta}_k$ the updating equation

$$\boldsymbol{\eta}_k^{[t+1]} = \boldsymbol{\eta}_k^{[t]} + \nu \cdot f_k^{[t]}(\mathbf{X}), \tag{1}$$

where ν is a step length parameter (e.g., $\nu = 0.1$). Therefore, predictors are improved slowly while each tree is estimated with maximum likelihood using offsets $\boldsymbol{\eta}_k^{[t]}$. In addition, the depth of the trees is kept small, which is a tuning parameter, so that a single $f_k^{[t]}(\cdot)$ only contributes a small amount to the overall model fit, similar to Bayesian additive regression trees (BART; Chipman et al. 2010). Moreover, instead of using all observations n for fitting a single tree in iteration t we only use a randomly selected subset $\boldsymbol{s}^{[t]} \subset \{1, \ldots, n\}$ of the data, i.e., each tree is build using (possibly) different data batches $\mathbf{X}_{\boldsymbol{s}^{[t]}}$. This leads to a regularization such that convergence of the algorithm is achieved when the log-likelihood evaluated on the batches becomes stationary around a certain level, i.e., in most applications, only enough boosting iterations T need to be provided without further tuning. In addition, it can be applied to very large data sets since the batchwise updating requires only a relatively small computational cost. We call this novel method batchwise boosting DAdaSoRT. An implementation is provided in the R package **softtrees** (Umlauf, 2023), see `help("BB-DAdaSoRT")`.

3 Child Anaemia Risk in Sub-Saharan Africa

Anaemia is a major health issue in low- and middle-income countries, particularly in sub-Saharan Africa, where over 50% of children under five are affected. We analyze haemoglobin (HgB) in a yet unexplored large-scale dataset with $> 340k$ observations from Demographic and Health Surveys. The data include climate, environmental and geospatial data. To perform model calibration checks, we split the data randomly into training and testing sets, with 80% of the data allocated to training and 20% to testing. We then benchmark the performance of a classical Bayesian additive model for location, scale, and shape (BAMLSS, Umlauf et al., 2018) with our proposed DAdaSoRT model. Notably, DAdaSoRT exhibited a considerably faster runtime, requiring approximately 6.5 hours to process 200 batches of 10000 data points, compared to approximately 65 hours for the Bayesian GAMLSS with 8000 MCMC iterations. Ultimately, we found that a model with skew exponential power type 3 distribution as implemented in the **gamlss.dist** package (Stasinopoulos and Rigby, 2022), achieved the best performance based on the out-of-sample continuous rank probability score (CRPS). Without further tuning, the skill score of this model, compared to a simple Gaussian intercept-only model, demonstrated an 11.13% improvement, while the skill score of the best-fitting BAMLSS model yielded an 11.05% improvement, indicating a marginal enhancement. However, this outcome is particularly promising as the identification of interactions is automated in our proposed DAdaSoRT model, unlike in BAMLSS. Additionally, compared to conventional distributional trees and forests, estimation with our approach is significantly more efficient and currently not feasible with available implementations of distributional trees or forests.

Figure 2 displays the out-of-sample quantile residuals of the final model. The histogram indicates approximately normally distributed residuals, while the worm plot reveals slight, yet statistically significant deviations from a zero mean for higher estimated quantiles. Overall, the model appears to be well calibrated, even on the test data.

The left panel of Figure 3 presents the log-likelihood contributions for each of the selected variables, indicating that land type and the age of the child in months are the two most influential factors. In the right panel, we depict the estimated spatial risk for $Pr(\text{HgB} < 110\,\text{g/L})$, illustrating substantial variations across the continent. Figure 4 showcases the marginal effects on $Pr(\text{HgB} < 110\,\text{g/L})$. All the figures related to $Pr(\text{HgB} < 110\,\text{g/L})$ demonstrate the exceptional ability of the proposed DAdaSoRT model to accurately approximate both sharp and smooth transitions. For instance, we observe sharp regional changes in the map of Figure 3 in contrast to the very smooth estimated effects in Figure 4.

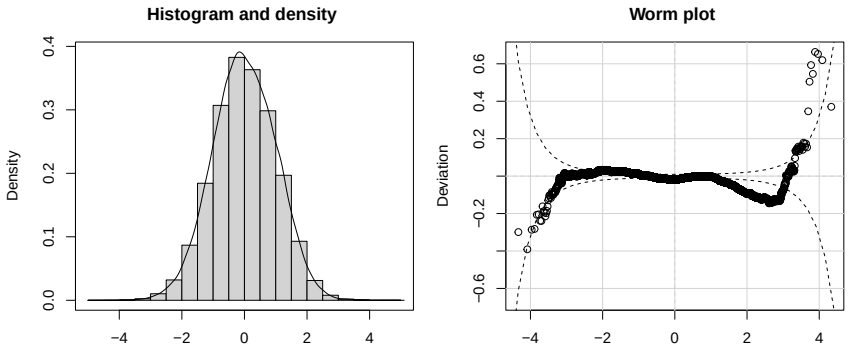


FIGURE 2. Model calibration plots using out-of-sample quantile residuals.

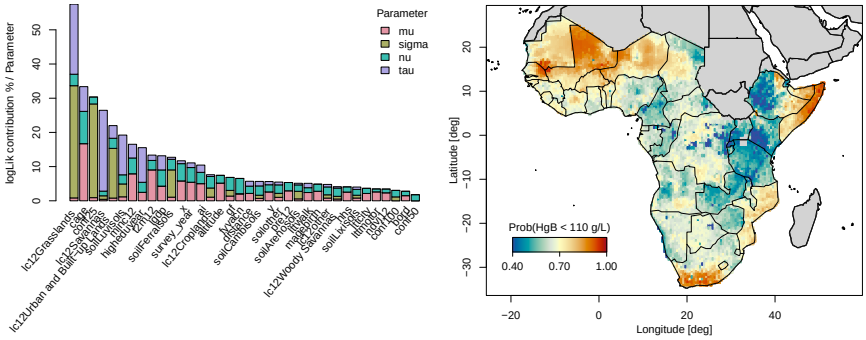


FIGURE 3. Variable log-likelihood contributions (left) and estimated anaemia risk for female infants at 30 months of age in 2020 (right).

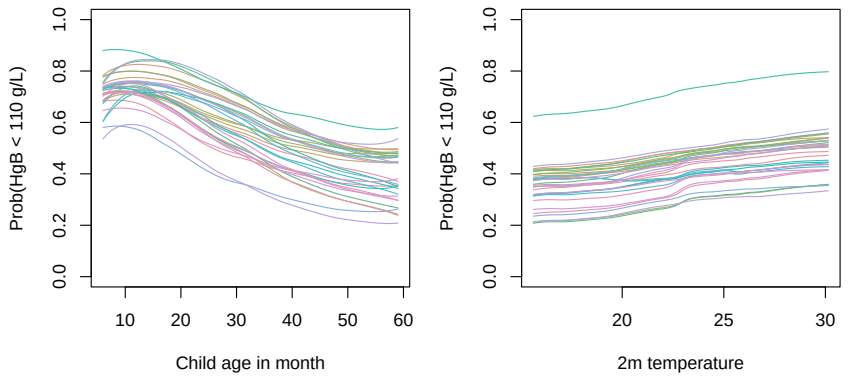


FIGURE 4. Estimated marginal effects on $Pr(\text{HgB} < 110\,\text{g/L})$ for each country in the data set.

Acknowledgments: This project was funded by the FWF grant #33941 and the DFG through the Emmy Noether grant KL 3037/1-1.

References

Rigby, R.A. and Stasinopoulos, D.M. (2005) Generalized Additive Models for Location, Scale and Shape. *Appl. Stat.*, **54**(3), 507 – 554.

Chipman, H.A., George, E.I, and McCulloch, R.E. (2010). BART: Bayesian Additive Regression Trees. *Ann. Appl. Stat.*, **4**(1), 266 – 298.

Umlauf, N., Klein, N, and Zeileis, A. (2018) BAMLSS: Bayesian additive models for location, scale and shape (and beyond). *J. Comput. Graph. Stat.*, **27**(3), 612 – 627.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. (2019) Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain. *Ann. Appl. Stat.*, **13**(3), 1564 – 1589.

Umlauf, N. and Klein, N. (2022). Distributional Adaptive Soft Regression Trees. *arXiv*, URL: `https://arxiv.org/abs/2210.10389`

Stasinopoulos, M. and Rigby, R. (2022) **gamlss.dist**: Distributions for Generalized Additive Models for Location, Scale and Shape. R package version 6.0-5, URL: `https://CRAN.R-project.org/package=gamlss.dist`

Umlauf, N (2023). **softtrees**: Soft Distributional Regression Trees and Forests. R package version 1.1, URL: `https://github.com/freezenik/softtrees`

A one-step spatial+ approach to mitigate spatial confounding in multivariate spatial areal models

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] Institute of Health Research, IdisNA, Spain.

E-mail for correspondence: arantxa.urdangarin@unavarra.es

Abstract: Multivariate spatial models for areal count data offer advantages over univariate counterparts as they reduce estimation error and unveil underlying correlations between the phenomena under study. However, assessing relationships between the responses and covariates of interest suffers from the challenging problem of spatial confounding, that is, the difficulty in disentangling the effects of the observed covariates and the spatial random effects. Though there is now a corpus of research about this problem, no definitive solution has been reached. In this work, we propose a modification of the so called spatial+ method in the multivariate framework. In particular, we use M-models and extend and modify the spatial+ method to the multivariate setting to estimate the linear relationship between the responses and some covariates. We use the proposal to analyse two form of crimes against women in Uttar Pradesh, India, and their relationship with some socio-demographic covariates.

Keywords: Crimes against women; Spatial confounding; M-models; Spatial+.

1 Introduction

Violence against women is a major problem in many countries where cultural traditions favour gender inequality. This is the case of India, one of the most populated countries in the world, where crimes against women are on the rise.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

Univariate spatial disease mapping models are crucial to visualize the spatial patterns of crimes against women, but a multivariate approach allows establishing relationships between them. The coregionalization framework derived by Martínez-Beneito (2013) covers many of the multivariate proposals in the literature. However, it may be computational prohibitive and an alternative reformulation known as M-models (Botella-Rocamora et al., 2015) has been developed.

When covariates enter in the model, an important challenge appears. Namely, the impossibility of separating the fixed effects from the spatial random effects. This is known as "spatial confounding" and it has been considered as a multicollinearity problem (Reich et al., 2006) resulting in biased estimates of the fixed effects. Various procedures have been proposed to alleviate spatial confounding. Here we focus on the spatial+ method (Dupont et al., 2022), consisting on removing spatial dependence of the covariates. More precisely, we consider M-models incorporating covariates and we modify the spatial+ approach to remove the spatial structure of the covariates and fit the multivariate models in a one-step procedure. We use the proposal to analyse two form of crimes against women, rapes and dowry deaths, in Uttar Pradesh (Vicente et al., 2020) and to asses their relationship with some sociodemographic covariates. Model fitting and inference is carried out using integrated nested Laplace approximations (Rue et al., 2009).

2 M-models

Let Y_{ij} and E_{ij} denote the number of observed and expected cases, respectively, in the ith small area ($i = 1, \ldots, I$) for jth crime ($j = 1, \ldots, J$). Conditional on the relative risk R_{ij}, Y_{ij} is assumed to follow a Poisson distribution

$$Y_{ij}|R_{ij} \sim Poisson(\mu_{ij} = E_{ij}R_{ij}) \quad \text{and} \quad \log \mu_{ij} = \log E_{ij} + \log R_{ij}.$$

The log-risk is modelled as

$$\log R_{ij} = \alpha_j + \beta_j x_i + \theta_{ij}, \tag{1}$$

where α_j is the intercept of jth crime, β_j is a crime-specific regression coefficient related to the covariate of interest $\boldsymbol{X} = (x_1, \ldots, x_I)'$, and θ_{ij} is the spatial effect of area i and crime j. To understand how M-models incorporate spatial dependence within each crime and induce correlation between different crimes, we rearrange the spatial effects in a matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(J)}) = \{\theta_{ij} : i = 1, \ldots, I;\ j = 1, \ldots, J\}$ where $\boldsymbol{\theta}^{(j)} = (\theta_{1j}, \ldots, \theta_{Ij})'$. Then, $\boldsymbol{\Theta}$ is expressed as

$$\boldsymbol{\Theta} = \boldsymbol{\Phi}\boldsymbol{M},$$

where $\boldsymbol{\Phi}$ is a matrix whose columns are independent and follow spatially correlated distributions, and $\boldsymbol{M}$ is a nonsingular and arbitrary $J \times J$ matrix that induces dependence between the columns of $\boldsymbol{\Theta}$. Entries of $\boldsymbol{M}$ can be interpreted as coefficients in the regression of log-relative risks on the columns of $\boldsymbol{\Phi}$. Hence, they can be interpreted as fixed effects and a $N(0, \sigma^2)$ prior with large fixed variance σ^2 is a sensible option for them. This is equivalent to assign $\boldsymbol{M}^{'}\boldsymbol{M} \sim Wishart(J, \sigma^2 \boldsymbol{I}_J)$ (see Botella-Rocamora et al., 2015).

3 One step spatial+ model

Spatial+ (Dupont et al., 2022) is a two step procedure designed to reduce bias in spatial models by eliminating the spatial dependence of the covariates. The first step consists in removing the spatial dependence of the covariate through a model. Then, in the second step, the spatial model (1) is fitted replacing the covariate by the residuals obtained in the first step. Here, we modify the procedure to remove the spatial dependence of the covariate and fit the spatial model in one single step. In more detail, we express the covariate $\boldsymbol{X}$ as a linear combination of the eigenvectors $\boldsymbol{U}^{(i)}$, $i = 1, \ldots, I$ of the random effects precision matrix. That is

$$\boldsymbol{X} = \delta_1 \boldsymbol{U}^{(1)} + \cdots + \delta_I \boldsymbol{U}^{(I)}.$$

Given that the eigenvectors corresponding to the lowest non-null eigenvalues are responsible for the collinearity between the fixed and random effects, we split the covariate into two parts $\boldsymbol{X} = \boldsymbol{Z} + \boldsymbol{Z}^{'}$,where $\boldsymbol{Z}^{'}$ comprises large-scale eigenvectors and $\boldsymbol{Z}$ contains the rest. More precisely, $\boldsymbol{Z}^{'}$ is formed by at least 5% and at most 20% of the large-scale eigenvectors (Urdangarin et al., 2022). Finally, the M-model is fitted replacing the covariate $\boldsymbol{X}$ in (1) by its spatially decorrelated part $\boldsymbol{Z}$ as

$$\log R_{ij} = \alpha_j + \beta_j z_i + \theta_{ij}. \tag{2}$$

4 Results

We fit models (1) and (2) to study rapes and dowry deaths in Uttar Pradesh in 2011. Specifically, our interest relies on the association between both crimes and the socio-demographic covariate sex ratio (number of females per 1000 males). Here we consider an intrinsic (ICAR) and the BYM2 prior for the spatial random effects, and six large-scale eigenvectors in $\boldsymbol{Z}^{'}$. Table 1 shows the fixed effect estimates. It can be observed that the estimates are different with the multivariate spatial model (1) and the multivariate spatial+ model (2). In addition, the posterior standard errors are smaller with the spatial+ approach. The association between sex ratio and dowry deaths is significant, while it is not significant for rapes.

TABLE 1. Posterior means, posterior standard deviations and 95% credible intervals of the sex ratio coefficient for rapes (β_{rape}) and dowry deaths (β_{dowry}).

	Φ	**Model**	**mean**	**sd**	**95% CI**	
β_{rape}	ICAR	(1)	-0.1560	0.1050	-0.3640	0.0510
		(2)	-0.0750	0.0680	-0.2090	0.0590
	BYM2	(1)	-0.1800	0.0990	-0.3720	0.0150
		(2)	-0.0680	0.0670	-0.2000	0.0650
β_{dowry}	ICAR	(1)	-0.1920	0.0600	-0.3080	-0.0720
		(2)	-0.0940	0.0410	-0.1740	-0.0130
	BYM2	(1)	-0.2500	0.0590	-0.3620	-0.1310
		(2)	-0.1100	0.0430	-0.1940	-0.0260

Acknowledgments: This work has been supported by Project PID2020-113125RB-I00/MCIN/AEI/10.13039/501100011033.

References

Botella-Rocamora, P., Martínez-Beneito, M.A., and Banerjee, S. (2015). A unifying modeling framework for highly multivariate disease mapping. *Statistics in Medicine*. **34**, 1548 – 1559.

Dupont, E., Wood, S.N., and Augustin, N.H. (2022). Spatial+: A novel approach to spatial confounding. *Biometrics*. **78**, 1279 – 1290.

Martínez-Beneito, M.A. (2013). A general modelling framework for multivariate disease mapping *Biometrika*. **100**, 539 – 553.

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., and Ugarte, M.D. (2022). Evaluating recent methods to overcome spatial confounding. *Revista Matemática Complutense*. https://doi.org/10.1007/s13163-022-00449-8.

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.

Extending central statistical monitoring to electronic patient-reported outcomes in clinical trials

Lawson Wang[1], Sebastiaan Höppner[1], Laura Trotta[1]

[1] CluePoints S.A., Louvain-la-Neuve, Belgium

E-mail for correspondence: zhendong.wang@cluepoints.com

Abstract: An increasing number of clinical trials are adapting the use of electronic patient-reported outcomes (ePRO) in their study protocol and conduct. Compared to the traditional data from multicenter clinical trials, ePRO data has no visit label and are more consistent as longitudinal time series. To extend central statistical monitoring in traditional clinical studies, we propose a new methodology for testing the anomalies using time series concepts and a mixed-effects model for continuous outcomes. The methods are divided into two modules that address two aspects of continuous outcomes: the first and second moment of the variable.

Keywords: ePRO data; Multicenter clinical trials; Mixed-effects model.

1 Introduction

A classic method of central statistical monitoring in longitudinal clinical studies is to use statistical tests to generate p-values, indicating potential outlying centers with abnormal data patterns. Examples of such methods are proposed by Desmet et al. (2014).
However, traditional clinical trials have scheduled visits and thus have limited time points. With electronic patient-reported outcomes (ePRO), patient data can be collected daily or even multiple times per day. The time evolution of endpoint variables behaves more like a time series. Therefore, the goal of this article is to extend the method to adapt to the large number of time points. Although scarce in the field of pharmaceutical monitoring, similar concepts with time series data can be seen in fields such as econometrics, following the work of Rousseeuw et al. (2019) and Olsen et al. (2021).

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Test on the mean and variability of a continuous ePRO variable

Consider a randomized clinical trial with I centers and J patients. Suppose patients are required to report a continuous outcome x (e.g. heart rate) electronically every day starting at their baseline visit until the end of their treatment. The expected duration is denoted as T. Let x_{ijt} be the outcome of patient j in center i at time t.
The measurements x_{ijt} can be seen as a time series for patient j if the patient reports the outcome x repeatedly over time. If the patient's treatment during the clinical trial has an impact on the measured continuous outcome over time, then we would expect a general trend over time T. Otherwise, we can reasonably assume the time series x_{ijt} is approximately stationary. A conventional choice for modeling longitudinal data is using a mixed-effects model:

$$x_{ijt} = \mu(x,t) + \gamma_i + \epsilon_{ijt}$$

where $\mu(x,t)$ is a function that depends on the time t and outcome variable x, γ_i which follows $N(0,\rho_c^2)$ is the random effect of center i, and ϵ_{ijt} which follows $N(0,\rho_r^2)$ is the residual error for each outcome record of patient j at time t.
Based on the fitted model, $\mu(x,t)$ can take three possible forms:

$$\mu(x,t) = \mu + bt + cx_{ij(t-1)} \quad \text{or} \quad \mu(x,t) = \mu + bt \quad \text{or} \quad \mu(x,t) = \mu$$

If $\mu(x,t)$ follows the first form, then it is assumed that the endpoint variable has a time trend and it is autocorrelated with the lag-1 term ($x_{ij(t-1)}$). Otherwise, the endpoint variable may have only a general time trend or no time effects at all. These possible models are selected using a combination index of AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). After selecting the best model, the original endpoint variable is modified by subtracting the fixed effect and the mean time trend:

$$y_{ijt} = x_{ijt} - \mu(x,t)$$

2.1 Estimate p-values for test on the means

The sample mean of the modified variable $y_i^* = \frac{1}{N_i}\sum_{j,t}^{N_i} y_{ijt}$ in each center i, follows a normal distribution $N\left(0, \sqrt{\rho_c^2 + \frac{\rho_r^2}{N_i}}\right)$ with N_i as the total number of patients for center i. We use this normal distribution to test for centers whose sample mean y_i^* is significantly different from 0 through the p-values of the test: $H_0 : y_i^* = 0$ vs. $H_1 : y_i^* \neq 0$ for each center i. Using all data points, the p-value for center i is computed as:

$$p(y_j^*) = \begin{cases} \min(2P(y \leq y_i^*), 1) & \text{if } y_i^* \leq 0 \\ \min(2P(y > y_i^*), 1) & \text{if } y_i^* > 0 \end{cases}$$

2.2 Estimate p-values for test on the variances

The sample variance of the modifed variable y_{ijt} is computed for each center i as:

$$s_i^2 = \frac{\sum_{i,k} N_{T_i}(y_{ijt} - \bar{y}_{ij*})^2}{(\sum_{j=1}^{N_i} N_{ij}) - N_i}$$

where $\bar{y}_{ij*}$ is the mean of all modified records of patient j in center i; N_{ij} is the number of records for patient j in center i; N_i is the number of patients in center i; and N_{T_i} is the total number of records in center j across all patients. We assume that the sample variance follows a gamma distribution, $s_j^2 \sim \Gamma(\kappa, \theta)$. For p-values, we need to first introduce $N = \left(\sum_{j=1}^{N_i} N_{ij}\right) - N_i$ and then compare the sum of squares in each center $ss_i = Ns_i^2$ to their expected value.

3 Simulation study

To assess the properties of the proposed tests, we conducted a simulation study. Table 1 shows how the clinical trials were simulated. The performance of the proposed tests is measured by their specificity and sensitivity.

TABLE 1. Setup of the simulation study.

Number of centers	75
Total number of patients	700
Number of patients per center	5, 10, 20, 40 or 100
Number of ePRO records per patient	100
Distribution of the endpoint variable	at $t = 1$: $x_{ij1} \sim N(3, 0.6)$
Additive time trend per patient	$0.01t + N(0, 0.5)$
Autoregressive coefficient	$0.9x_{t-1}$
Number of simulations	1000

First, we assess the specificity of the tests without any outlying centers. Figure 1 shows that the average specificity is above 0.98 for both the mean test and variance test, regardless of the center size (i.e. number of patients in the center). There is no visible effect of the center size on the estimated p-values when the data does not contain outlying centers. Next, we add contamination to a random center and check if the outlying center is detected by our new tests. We introduce a shift of $0.1 * mean(x_{ijt})$ to a random center for the mean test and a shift of $0.1 * sd(x_{ijt})$ to a random center for the variance test. We then calculate the sensitivity and specificity of the new tests per size of the contaminated center. Figure 2 shows that the specificity of the tests remains high, despite having a contaminated center

in the data., so the tests do not generate too many false positives. The sensitivity of both tests improves as the center size of the contaminated center increases, which is expected because smaller outlying centers are more difficult to detect and require more statistical power to be identified.

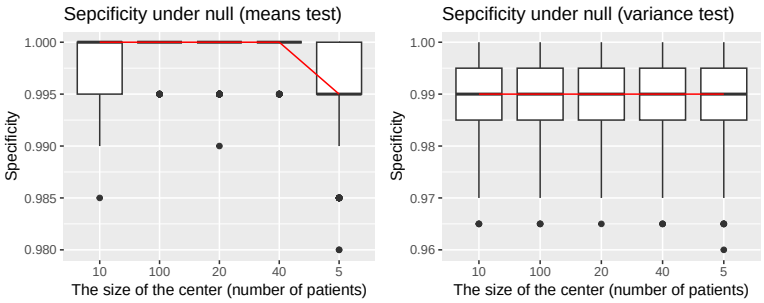


FIGURE 1. Specificity analysis of the mean test (left) and the variance test (right) when no centers are contaminated.

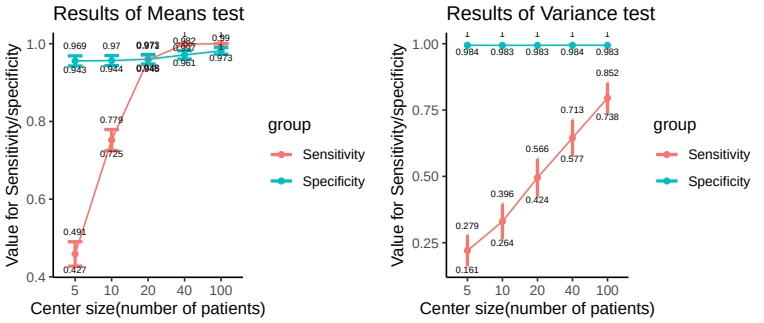


FIGURE 2. Sensitivity and specificity analysis of the mean test (left) and the variance test (right) when one center is contaminated.

References

Desmet, L., et al. (2014). Linear mixed-effects models for central statistical monitoring of multicenter clinical trials. *Statistics in medicine*, **33(30)**, 5265 – 5279.

Olsen, M.H., et al. (2021). Central data monitoring in the multicentre randomised SafeBoosC-III trial – a pragmatic approach. *BMC Medical Research Methodology*, **21**(1), 1 – 10.

Rousseeuw, P., et al. (2019). Robust monitoring of time series with application to fraud detection. *Econometrics and Statistics*, **9**, 108 – 121.

Ordinal compositional data and time series

Christian H. Weiß[1]

[1] Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany

E-mail for correspondence: weissc@hsu-hh.de

Abstract: For various areas of ordinal compositional data (CoDa), approaches to consider the natural order among the categories are proposed and recommended to complement existing CoDa methods. Their benefits are demonstrated for a descriptive data analysis, for statistical inference based on CoDa samples, for control charts to monitor a CoDa process, and for compositional time series analysis.

Keywords: Compositional data; Conditional regression models; Control charts; Ordinal categories; Time series.

1 Introduction

For a given set of categories, say $\mathcal{S} = \{s_0, \ldots, s_d\}$ with $d \in \mathbb{N} = \{1, 2, \ldots\}$, the vector $\boldsymbol{x} = (x_0, \ldots, x_d)^\top \in (0;1)^{d+1}$ is said to be a $(d+1)$-part composition iff its components sum up to one. Here, x_i is interpreted as the proportion of category s_i, and a data set consisting of such compositions is referred to as CoDa. The range of $\boldsymbol{x}$ is the $(d+1)$-part simplex

$$\mathbb{S} \;:=\; \big\{\boldsymbol{x} \in (0;1)^{d+1} \;\big|\; x_0 + \ldots + x_d = 1\big\}.$$

A CoDa vector $\boldsymbol{p} \in \mathbb{S}$ might serve as the probability mass function (PMF) of a categorical random variable (RV) Q with range $\mathcal{S}$, thus establishing a natural connection between CoDa and categorical data. In many applications, the categories behind CoDa are unordered, so $\mathcal{S}$ is a nominal range, and the notation "$s_0, \ldots, s_d$" uses a lexicographic order. Hence, "the conclusions of a compositional analysis should not depend on the order of the parts" (Pawlowsky-Glahn and Buccianti, 2011, p. 17). However, in some examples, the categories in $\mathcal{S}$ exhibit a natural order, namely $s_0 < \ldots < s_d$, so the categorical RV Q is indeed an ordinal RV. In this case, we refer to

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

$\boldsymbol{x} \in \mathbb{S}$ as an ordinal composition. While it is still justified to apply the well-established CoDa approaches to ordinal CoDa, it is shown that additional insights are gained if these are supplemented by new CoDa approaches that explicitly account for the order within $\mathcal{S}$. Such novel ordinal CoDa approaches are derived by adapting well-established concepts from ordinal data analysis. Their potential benefits are demonstrated for a descriptive analysis of ordinal CoDa, for statistical inference regarding ordinal CoDa RVs, for control charts to monitor ordinal CoDa processes, and for analyzing and modelling compositional time series (CoTS).

2 Applying Ordinal Statistics to CoDa

Let Q be an ordinal RV with range $\mathcal{S}$ and PMF $\boldsymbol{p} \in \mathbb{S}$. To account for the natural order in $\mathcal{S}$, one prefers the cumulative distribution function (CDF) given by the vector $\boldsymbol{f} = (f_0, \ldots, f_{d-1})^\top \in [0;1]^d$, where $f_j = P(Q \leq s_j)$. For a parametric modelling of ordinal data, one may use the latent-variable approach (Agresti, 2010, p. 11). If L is a (latent) real-valued RV with specified CDF $F_L(x)$, then one "reparametrizes" $\boldsymbol{f}$ in terms of threshold parameters $-\infty < \eta_0 < \ldots < \eta_{d-1} < +\infty$ such that

$$f_j \;=\; F_L(\eta_j) \qquad \text{for } j = 0, \ldots, d-1. \tag{1}$$

The most popular choice for F_L is the standard logistic distribution, leading to the cumulative logit model.
For an ordinal RV Q, the location is expressed by the median. A common ordinal dispersion measure is the index of ordinal variation,

$$\mathrm{IOV}(\boldsymbol{f}) \;=\; \frac{4}{d} \sum_{i=0}^{d-1} f_i(1 - f_i) \quad \in [0;1], \tag{2}$$

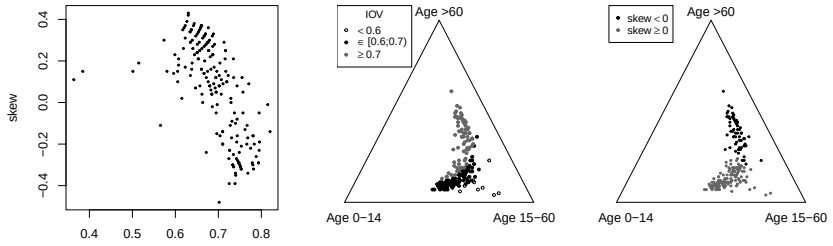
and an ordinal skewness measure is given by

$$\mathrm{skew}(\boldsymbol{f}) \;=\; \frac{2}{d} \sum_{i=0}^{d-1} f_i \; - 1 \quad \in [-1;1]; \tag{3}$$

see Weiß (2020) for a discussion. To apply the ordinal measures (2) and (3) to ordinal CoDa $\boldsymbol{x} \in \mathbb{S}$, we have to accumulate these vectors by $\boldsymbol{c} := \mathbf{T}\,\boldsymbol{x}$ with the matrix

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 1 & \cdots & 1 & 0 \end{pmatrix}.$$

Then, we evaluate the ordinal dispersion and skewness of $\boldsymbol{x}$ by applying formulae (2) and (3) to $\boldsymbol{c}$ instead of $\boldsymbol{f}$. Consider the data set `ageCatWorld` of the R-package `robCompositions` as an example, which provides 3-part



(a) (b) (c)

FIGURE 1. Age proportions in 195 countries from Section 2: plot of skew against IOV in (a), ternary diagrams in (b) and (c).

compositions ($d = 2$) $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ for the ordered categories of people with age < 15 ($x_{i,1}$), 15–60 ($x_{i,2}$), and > 60 ($x_{i,3}$) in $n = 195$ countries. According to Figure 1(a), the values of $\mathrm{IOV}(\boldsymbol{c}_i)$ vary between 0.4 and 0.8 (medium to strong dispersion). The few countries with IOV < 0.6 (empty circles in the ternary diagram in (b)) are those with a high proportion of people with age 15–60. Countries having an IOV ≥ 0.7, in turn, are plotted (by grey dots) relatively close to the axis between the lowest and largest age category. Part (a) also shows at most moderate skewness values (maximal absolute extent around 0.4), where countries with negative skewness tend to show larger dispersion. Part (c) shows that positive skewness values refer to the lower-left half of the triangle: right-skewed compositions are left leaning, so the lower categories dominate the upper ones.

3 Statistical Inference for Ordinal CoDa

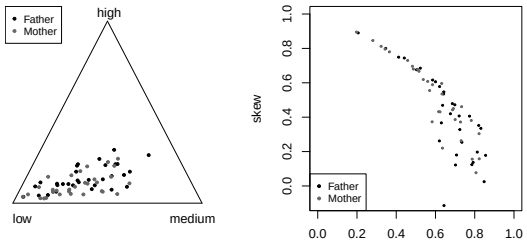
Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be an independent and identically distributed (i. i. d.) sample of ordinal CoDa RVs, define $\sigma_{ij} = Cov[X_i, X_j]$ for $i, j = 0, \ldots, d$. Let $\boldsymbol{C}_i := \mathbf{T}\,\boldsymbol{X}_i$ with mean $\boldsymbol{f}$, let $\overline{\boldsymbol{C}}$ denote the sample mean of $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n$, and let $\sigma'_{ij} = \sum_{r=0}^{i} \sum_{s=0}^{j} \sigma_{rs}$. Then, $\sqrt{n}\,\mathrm{IOV}(\overline{\boldsymbol{C}})$ according to (2) is asymptotically normally distributed with

$$\begin{aligned} E\big[\mathrm{IOV}(\overline{\boldsymbol{C}})\big] &\approx \mathrm{IOV}(\boldsymbol{f}) - \tfrac{1}{n}\tfrac{4}{d}\textstyle\sum_{i=0}^{d-1} \sigma'_{ii}, \\ V\big[\mathrm{IOV}(\overline{\boldsymbol{C}})\big] &\approx \tfrac{1}{n}\tfrac{16}{d^2}\textstyle\sum_{i,j=0}^{d-1}(1-2f_i)(1-2f_j)\,\sigma'_{ij}. \end{aligned} \tag{4}$$

For $\mathrm{skew}(\overline{\boldsymbol{C}})$ from (3), we have

$$E\big[\mathrm{skew}(\overline{\boldsymbol{C}})\big] = \mathrm{skew}(\boldsymbol{f}) \quad \text{and} \quad V\big[\mathrm{skew}(\overline{\boldsymbol{C}})\big] \approx \tfrac{1}{n}\tfrac{4}{d^2}\textstyle\sum_{i,j=0}^{d-1} \sigma'_{ij}. \tag{5}$$

The finite-sample performance of (4) and (5) is investigated by simulations. We apply (4) and (5) to the data set `educFM` of the R-package `robCompositions`, which provides the proportions of low, medium, and



(a) (b)

Father	Estim.	Bias	SE
IOV($\overline{\boldsymbol{C}}$)	0.731	-0.003	0.022
skew($\overline{\boldsymbol{C}}$)	0.414	0.000	0.043
Mother			
IOV($\overline{\boldsymbol{C}}$)	0.652	-0.002	0.033
skew($\overline{\boldsymbol{C}}$)	0.516	0.000	0.041

(c)

FIGURE 2. Proportions of education levels in 31 European countries from Section 3: ternary diagram in (a), plot of skew against IOV in (b), estimates in (c).

high education levels of father and mother, respectively, in $n = 31$ European countries. From the ternary plot in Figure 2(a) and the IOV-skew diagram in (b), no significant differences between the education proportions of father and mother (black vs. grey dots) can be recognized. But if looking at the point estimates for IOV($\overline{\boldsymbol{C}}$) and skew($\overline{\boldsymbol{C}}$) and their corresponding asymptotic standard errors (SEs) in (c) (the biases are negligible), we recognize a difference being larger than two times an SE. So in view of the asymptotic normality according to (4) and (5), we conclude that the IOV and skew estimates are significantly different between father and mother. In fact, the mothers' education proportions are closer to a one-point distribution in the category "low", possibly indicating unequal opportunities for education among males and females in the past.

4 Control Charts for Ordinal CoDa

If monitoring an ordinal CoDa process $(\boldsymbol{X}_t)$ being assumed to be i. i. d. under in-control conditions, then existing control charts such as in Vives-Mestres et al. (2014) do not make use of the natural order of the categories. We consider the accumulated process $(\boldsymbol{C}_t)$ with in-control mean $\boldsymbol{f}_0$ and apply an exponentially weighted moving-average (EWMA) approach with parameter $\lambda \in (0;1)$: $\boldsymbol{C}_{t,\lambda} = \lambda \boldsymbol{C}_t + (1-\lambda)\boldsymbol{C}_{t-1,\lambda}$ with $\boldsymbol{C}_{0,\lambda} = \boldsymbol{f}_0$.
Then, the EWMA IOV- and skew-charts are defined by plotting

$$\text{IOV}(\boldsymbol{C}_{t,\lambda}) \quad \text{and} \quad \text{skew}(\boldsymbol{C}_{t,\lambda}) \quad \text{for } t = 1, 2, \ldots \tag{6}$$

against appropriately chosen control limits. It is shown through simulations that these novel control charts allow for a targeted diagnosis of the actual out-of-control scenario and, thus, constitute a valuable complement of existing CoDa control charts. An illustrative data application about the manufacturing of grit (particle size proportions with categories "large", "medium", and "small", see Vives-Mestres et al. (2014) for details) is presented in Figure 3. While the skew-chart in (b) does not indicate a system-

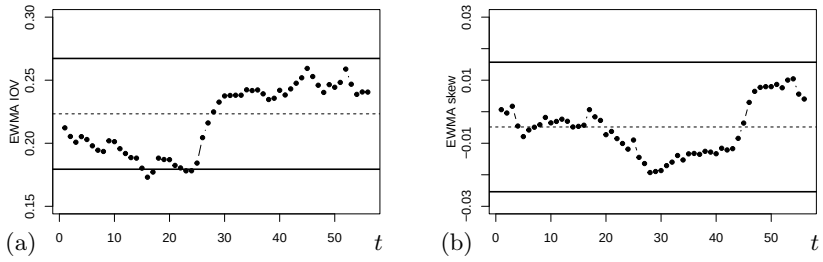


FIGURE 3. Particle size proportions from Section 4: EWMA IOV-chart in (a) and EWMA skew-chart in (b), where $\lambda = 0.1$.

atic skewness change, the IOV-chart in (a) uncovers an exceptionally low dispersion for $t \leq 16$ and $t \leq 23$.

5 Ordinal Compositional Time Series

For modelling an ordinal CoTS, an extension of the conditional regression model of Zheng and Chen (2017) is proposed that makes use of the cumulative logit approach (1). Define $\tilde{\boldsymbol{X}}_t = (X_{t,0}, \ldots, X_{t,d-1})^\top$ and $\tilde{\boldsymbol{p}}_t = (p_{t,0}, \ldots, p_{t,d-1})^\top$, and let $\boldsymbol{f}_t$ denote the conditional CDF vector given the information up to time $t-1$. Then,

$$f_{t,i} \;=\; F_L\Big(\eta_i + \textstyle\sum_{k=1}^{\mathrm{p}} \boldsymbol{\alpha}_k^\top\, \tilde{\boldsymbol{X}}_{t-k} + \sum_{l=1}^{\mathrm{q}} \boldsymbol{\beta}_l^\top\, \tilde{\boldsymbol{p}}_{t-l}\Big) \quad \text{for } i = 0, \ldots, d-1 \quad (7)$$

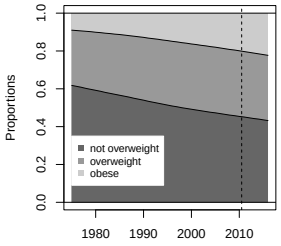
has a similar structure like an autoregressive moving-average model. Model (7) is adapted to account for covariate information $\boldsymbol{z}_t$ by adding the summand "$+\boldsymbol{\gamma}^\top \boldsymbol{z}_t$" within the parentheses.

The subsequent data example shows that (7) enables an efficient and well-interpretable modelling of ordinal CoTS. The CoTS $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{42}$ in Figure 4(a) shows the yearly proportions of three weight categories (so $d = 2$) in Germany for the period 1975–2016, which are determined based on the body mass index (BMI) as follows: "not overweight" (BMI < 25), "overweight" (BMI in $[25; 30)$), and "obese" (BMI ≥ 30). These (age-standardized) percentages for different BMI classes are provided by the World Health Organization. The vertical dashed line between the years 2010 and 2011 expresses the following partition: the data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{36}$ for 1975–2010 are used for model fitting, and $\boldsymbol{x}_{37}, \ldots, \boldsymbol{x}_{42}$ for 2011–2016 for out-of-sample forecasting. Because of the obvious trend in Figure 4(a), a first candidate model is a simple (logistic-)linear model:

$$f_{t,i} \;=\; F_{0,1}(\eta_i + \gamma_i\, t) \quad \text{for } i = 0, \ldots, d-1. \quad (8)$$

As a competitor, (8) is extended by an additional autoregressive component:

$$f_{t,i} \;=\; F_{0,1}(\eta_i + \gamma_i\, t + \boldsymbol{\alpha}_1^\top\, \tilde{\boldsymbol{X}}_{t-1}) \quad \text{for } i = 0, \ldots, d-1. \quad (9)$$



(a)

Model	i	η_i	γ_i	$\alpha_{1,i}$	CSS_{in}	CSS_{out}
(8)	1	0.480	-0.019		0.698	1.269
	2	2.350	-0.027			
(9)	1	0.469	-0.016	0.322	0.410	0.379
	2	2.354	-0.025	-0.651		

(b)

FIGURE 4. Proportions of weight categories from Section 5: (a) plot of proportions over time, and (b) table with model fits and corresponding CSS values.

Parameter estimation is done by a conditional least-squares (CLS) approach, i.e., by numerically minimizing the conditional sum of squares (CSS) defined by

$$\text{CSS}(\boldsymbol{\theta}) := \sum_t \|\boldsymbol{X}_t - \boldsymbol{p}_t\|^2 \ \to \min, \tag{10}$$

where $\boldsymbol{\theta}$ comprises all model parameters, and where $\|\cdot\|$ denotes the Euclidean norm. The results of CLS estimation are summarized in Figure 4(b). The linear coefficients for models (8) and (9) are negative in agreement with the decreasing curves in Figure 4(a). For performance analysis, two types of CSS are computed: "CSS_{in}" equals 10^3 times the CSS (10) computed for the in-sample data $t = 2, \ldots, 36$, and "CSS_{out}" the one for the out-of-sample data $t = 37, \ldots, 42$. It can be seen that the additional autoregressive term, given by "$+0.322\, X_{t-1,0} - 0.651\, X_{t-1,1}$", leads to an improved model performance.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. 2nd edition. Hoboken: John Wiley & Sons, Inc.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis — Theory and Practice*. Chichester: John Wiley & Sons, Ltd.

Vives-Mestres, M., Daunis-i-Estadella, J., and Martín-Fernández, J.-A. (2014) Individual T^2 control chart for compositional data. *Journal of Quality Technology*, **46**, 127 – 139.

Weiß, C.H. (2020). Distance-based analysis of ordinal data and ordinal time series. *Journal of the American Statistical Association*, **115**, 1189 – 1200.

Zheng, T., Chen, R. (2017) Dirichlet ARMA models for compositional time series. *Journal of Multivariate Analysis*, **158**, 31 – 46.

Stagewise boosting distributional regression

Mattias Wetscher[1], Johannes Seiler[1],
Reto Stauffer[1], Nikolaus Umlauf[1],

[1] Universität Innsbruck, Austria

E-mail for correspondence: Mattias.Wetscher@uibk.ac.at

Abstract: Forward stagewise regression is a simple algorithm for estimating regularized models. The updating rule slowly solves the optimization problem by adding a small constant to each regression coefficient in each iteration. This is similar to gradient boosting, but the step size is determined differently. Gradient boosting for distribution regression can lead to a vanishing gradient problem in a number of situations, resulting in suboptimal models. We propose a stagewise boosting type algorithm for complex distribution regression modelling with correlation filtering and best subset selection that can handle very large data problems. We demonstrate the effectiveness of our proposed approach using an example of lightning count data with over 9.1 million observations and 672 covariates.

Keywords: Stagewise regression; boosting; GAMLSS; variable selection, correlation filtering, batchwise updating.

1 Introduction

Modern regression models can provide full probabilistic predictions, crucial in numerous applications such as predicting severe weather (see, e.g., Simon et al., 2019). The generalized additive model for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos, 2005) is a well-known model class for such probabilistic settings. A common choice for stable estimation and variable selection for such models is gradient boosting (Mayr et al., 2012). However, the search for the optimal stopping iteration (e.g., in cross-validation) can lead to suboptimal models, especially when the updating scheme depends heavily on gradient information and when the vanishing gradient problem occurs for some predictors and updating basically stops after a few iterations, as illustrated on simulated negative binomial type 1 (`NBI`) data in Figure 1. Here the optimum $\beta_\sigma^\star$ is not reached with gradient boosting even after 1000 iterations, even with a large chosen step length

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

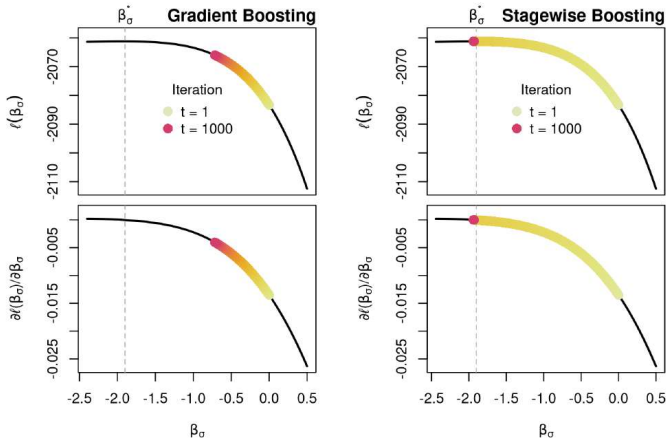


FIGURE 1. Boosted `NBI` model vanishing gradient problem. First row, (marginal) log-likelihood functions, second row, corresponding gradient information.

parameter, see the left column. To address the vanishing gradient problem in distributional regression models, we propose adapting the general stagewise algorithm (Thibshirani, 2015) to update coefficients at a small bounded (semi-constant) rate determined by the sign of the gradient. This approach allows for a more balanced selection of covariates, e.g., as illustrated in Figure 1, the optimum is reached in very few iterations with stagewise boosting. Moreover, our method improves on existing boosting approaches for GAMLSS by offering best subset selection of distributional parameters, a variable selection method for high-dimensional models, and a batchwise version for processing large datasets efficiently.

2 Model and Algorithm

Let y_i be the response and $\mathbf{x}_i$ covariate information for data with n observations indexed by $i = 1, \ldots, n$. We assume conditional independence of observations given covariates. Here, the response y_i has parametric density $Y_i|\mathbf{x}_i \sim \mathcal{D}(\theta_{i1}, \ldots, \theta_{iK})$, where the K parameters $\theta_{ik} \equiv \theta_k(\mathbf{x}_i)$, $k = 1, \ldots, K$, are linked to additive predictors η_k using known monotonic and twice differentiable functions by $h_k(\theta_{ik}) = \eta_k(\mathbf{x}_{ik}) = \eta_{ik} = \mathbf{x}_{ik}^\top \beta_k$. In this context, $\mathbf{x}_{ik}^\top$ is a row of the predictor specific design matrix $\mathbf{X}_k$ and $\beta_k = (\beta_{0k}, \ldots, \beta_{J_k k})^\top$ are regression coefficients that need to be estimated. Further, $\eta_k = \mathbf{X}_k \beta_k$ is the vector of the k-th linear predictor, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_K)$ is the full data matrix and $\beta = (\beta_1, \ldots, \beta_K)^\top$ the vector of all regression coefficients. In each iteration $t = 1, \ldots, T$, our stagewise boosting algorithm starts by preselecting the variables with the highest correlation (in absolute value) with the gradient vectors

$$\mathbf{g}_k = \left(\frac{\partial \log \mathcal{D}(y_i, \mathbf{x}_i, \beta^{[t-1]})}{\partial \eta_k} \right)_{i=1,\ldots,n}.$$

This leads to a set of K variables, one variable per distribution parameter. One novelty here, which deviates from the classical cyclic or non-cyclic boosting is that our algorithm selects the best subset of the potential variables for a succeeding update instead of an update of all (cyclic) or only the overall best performing variable (non-cyclic). Second, for determining the best performing subset of variables, consider a non-empty subset $S \subset \{1, \ldots, K\}$. We combine the derivatives of the averaged log-likelihood function $\partial \ell_{j_s^* s} = \frac{1}{n} \frac{\partial \ell(\beta^{[t-1]}; \mathbf{y}, \mathbf{X})}{\partial \beta_{j_s^* s}}$ determined from $\mathcal{D}(\cdot)$, with respect to the potential variables (indexed by j_s^*) into a gradient $\nabla \mathbf{L}_S = \left(\partial \ell_{j_s^* s}\right)_{s \in S}$ which will be used to define the updating step. To avoid an exploding or a vanishing gradient, the gradient $\nabla \mathbf{L}_S$ gets rescaled if its euclidean length exceeds a certain value ϵ and if any individual partial derivative $\partial \ell_{j_s^* s}$ fails to overcome a minimum threshold $\nu \cdot \epsilon$ (e.g., $\nu = 0.1, \epsilon = 0.01$) they get rescaled to this threshold value. For $s \in S$, this rescaling yields our semi-constant updating step length of the parameters ϵ_{Ss} and the corresponding updates

$$\beta_s^{[t]} = \beta_s^{[t-1]} + \epsilon_{Ss} \cdot \text{sign}\left((\mathbf{X}_s)_{\cdot j_s^*}^\top \mathbf{g}_s \right) \cdot \mathbf{e}_{j_s^*},$$

where $\mathbf{e}_{j_s^*}$ is a vector of zeros except at position j_s^* is a one. The subset S with the highest improvement in the log-likelihood gets choosen. By selecting variables through correlation with $\mathbf{g}_k$ and adding a threshold value κ (e.g., $\kappa = 0.15$), we can filter out variables with $r_{jk} \leq \kappa$ and consider only the remaining ones ($r_{jk} > \kappa$) for updating. If no variable remains in a distributional parameter with a sufficiently large correlation, no update is performed. As the updating continues, the correlation values decline until they no longer overcome the minimum requirement κ, indicating implicit early stopping. To estimate models with large datasets, we use stochastic approximations of correlations and updates. We perform preselection on a subset of data $\mathbf{i}^{[t]} \subset \{1, \ldots, n\}$ and choose which variables to update on the next batch $\mathbf{i}^{[t+1]}$, providing stability with quasi out-of-sample data. This described algorithm is our stagewise boosting variable selection step. Following this we refit the model with the selected variables and with the same algorithm but without the correlation filtering until convergence.

3 Lightning forecast in Austria

Lightning is a natural phenomenon that occurs during thunderstorms, when the electrical charge in the atmosphere becomes imbalanced. We use high-resolution data from the Austrian Lightning Detection and Information System (ALDIS; Schulz et al., 2005) and explain the lightning counts with

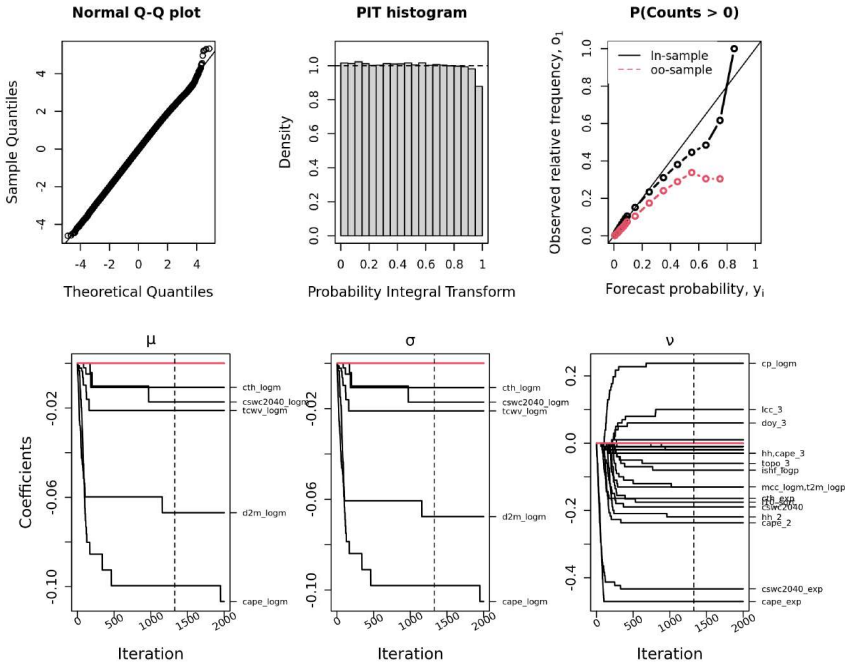


FIGURE 2. Out-of-sample diagnostic plots are shown in the top row. The bottom row shows the coefficient paths of selected variables in selection step of algorithm. The suffix _function corresponds to the transformation applied to the variable.

reanalysis data from ERA5, the fifth generation of ECMWF (European Centre for Medium-Range Weather Forecasts) atmospheric reanalyses of global climate. We use the data corresponding to 2010 up to 2018 (≈ 8.2 million observations) as training data and the year 2019 as validation data. We include 84 physical variables in our analysis, which we initially transform using the empirical distribution function to ensure that all variables fall within the interval $[0, 1]$. Following this initial transformation, we augment each variable with eight different transformations, resulting in each linear predictor having a pool of $84 \cdot 8 = 672$ variables to select from:

$$
\begin{aligned}
&x \mapsto x && x \mapsto \sqrt{x} \\
&x \mapsto x^2 && x \mapsto \operatorname{logp}(x) = \log(x + 0.01) \\
&x \mapsto x^3 && x \mapsto \operatorname{logm}(x) = \log(1 - x + 0.01) \\
&x \mapsto \exp(x) && x \mapsto \operatorname{logitc}(x) = \log\left(\frac{0.999 \cdot x + 0.001 \cdot 0.5}{1 - (0.999 \cdot x + 0.001 \cdot 0.5)}\right)
\end{aligned}
$$

To better capture the characteristics of rare positive lightning events, we aim to improve our model by subsampling the zero count data during the

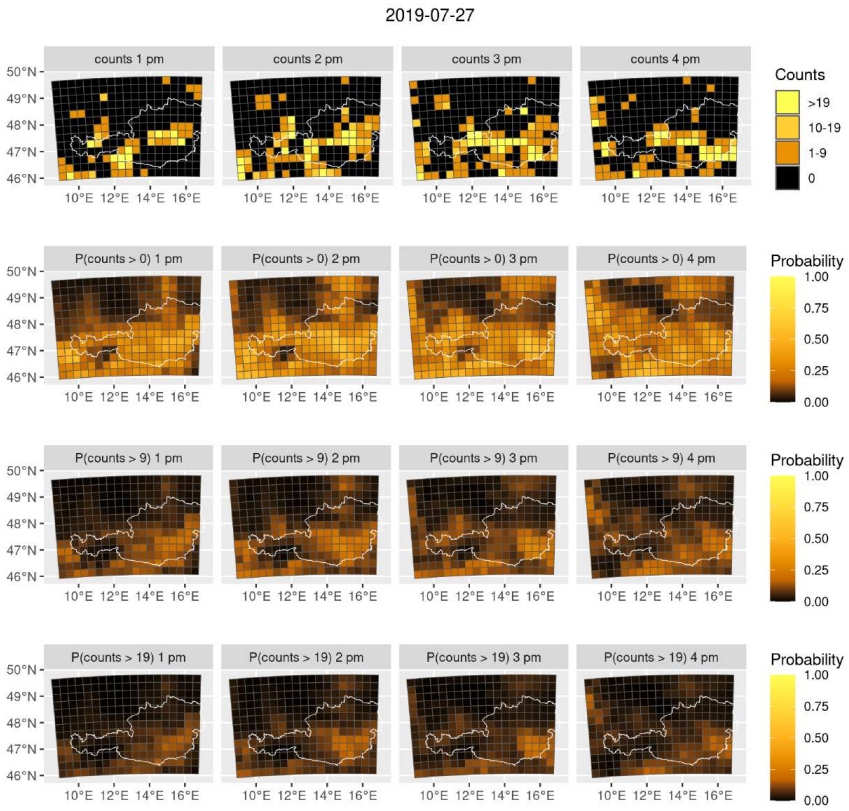


FIGURE 3. Refitted ZANBI lightning model. Compared is a (out of sample) forecast for the date 2019-07-27 with the observed counts. The time evolution from 1 pm up to 4 pm is depicted along the horizontal axis in the panel plot. The top row shows the observed number of lightning counts, the second row shows the probability for at least one count, the third row shows the probability for at least 10 counts and the fourth row shows the probability for at least 20 counts.

variable selection step. We achieve this by specifying batches $\mathbf{i}^{[t]}$ consisting of 10000 random samples each from the flash $\texttt{counts} = 0$ and $\texttt{counts} > 0$ observations. To account for subsampling, we apply an intercept adjustment in the logistic regression part of our model prior to the refitting step. The subsampled parameter $\boldsymbol{\beta}_{0\nu}^{\text{sub}}$ gets adjusted to

$$\boldsymbol{\beta}_{0\nu} = \boldsymbol{\beta}_{0\nu}^{\text{sub}} - \log\left(\frac{1-\tau_0}{\tau_0} \cdot \frac{t_0}{1-t_0}\right),$$

where τ_0 is the proportion of zeros in the dataset and t_0 is the proportion of zeros in the subsampled dataset. We have $t_0 = 0.5$ and $\tau_0 \approx 0.0265$. Diagnostic plots and the evolution of coefficients from the selection step are shown in Figure 2. On the top left entry a qq-plot based on randomized quantile residuals is shown, indicating a optimal fit for all but some extreme values corresponding to very high count observations. The PIT histogram also indicate a good calibration. The top right plot is a reliability diagram for the tail probability $\mathbb{P}(\text{counts} > 0)$. It shows that the in-sample fit is good and the out-of-sample fit experiences some overestimation for the high tail probabilities. Please note that the proportion of positive counts in the training data set is $\approx 2.65\%$ and for the out-of-sample data set $\approx 1.8\%$. An out-of-sample forecast for severe lightning counts (counts ≥ 10) is illustrated in Figure 3.

Acknowledgments: This project was partially funded by the Austrian Science Fund (FWF) grant number 33941. We are grateful for data support by Gerhard Diendorfer and Wolfgang Schulz from OVE-ALDIS.

References

Mayr A., Fenske N, Hofner B, Kneib T, Schmid M (2012). Generalized additive models for location, scale and shape for high-dimensional data - a flexible approach based on boosting. *J. R. Stat. Soc. Ser. C-Appl. Stat.*, **61**(3): 403 – 427.

Rigby, R.A. and Stasinopoulos, D.M. (2005) Generalized Additive Models for Location, Scale and Shape. *J. R. Stat. Soc. Ser. C-Appl. Stat.*, **54**(3), 507 – 554.

Schulz W, Cummins K, Diendorfer G, Dorninger M (2005). Cloud-to-Ground Lightning in Austria: A 10-Year Study Using Data from a Lightning Location System. *J. Geophys. Res.-Atmos.*, **110**(D9).

Simon T, Mayr GJ, Umlauf N, Zeileis A (2019). NWP-Based Lightning Prediction Using Flexible Count Data Regression. *Adv. Stat. Climatol. Meteorol. Oceanogr.*, **5**(1), 1 – 16.

Tibshirani RJ (2015). A General Framework for Fast Stagewise Algorithms. *J. Mach. Learn. Res.*, **16**(78), 2543 – 2588.

Gaussian process models: From astrophysics to industrial data

Jamie Wilson[1], Kevin Burke[1], Norma Bargary[1]

[1] University of Limerick, Ireland

E-mail for correspondence: jamie.wilson@ul.ie

Abstract: Gaussian processes offer a flexible approach to the statistical modelling of arbitrary functions and are particularly effective for time-series interpolation and prediction problems. We will demonstrate the versatility of Gaussian process models for time-series analysis through illustrative examples of exoplanet light curve modelling and applications to industrial and manufacturing data.

Keywords: Gaussian processes; exoplanet atmospheres; industrial data.

1 Introduction

Gaussian process (GP) models are a powerful and flexible tool for performing Bayesian inference and are used extensively for non-parametric regression and classification problems in the machine learning community (Rasmussen and Williams 2006, Bishop 2006). More recently, thanks largely to their versatility and robust uncertainty estimates, they have begun to be employed to solve problems in a diversity of fields such as engineering and astrophysical data analysis. For example, within the exoplanet community (an exoplanet being a planet which orbits a star other than the Sun), GPs are now routinely used to model and remove the effects of so-called "instrumental systematics" in transit light curves (Gibson, 2012). These are often encountered as a result of imperfect observing conditions and/or issues related to the light detectors and significantly hamper our ability to infer the transit parameters, which are required to accurately identify key atomic and molecular components in the planet's atmosphere. A particular strength of GPs is their ability to simultaneously model a potentially complex deterministic component (e.g. a transit function) alongside a stochastic component which describes the correlated noise structure. Furthermore, the versatility of GP models allows them to be successfully applied to a wide

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

range of problems commonly encountered in industry, such as statistical modelling of critical machine parameters, detecting outliers and predicting machine breakdown. In this presentation we aim to illustrate the applicability of GPs for solving challenges in a wide range of contexts, both astrophysical and industrial.

2 Gaussian Process Models for Regression

A Gaussian process is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution (Rasmussen and Williams 2006). In a typical regression problem we model our observed outputs $\mathbf{y}$ as

$$\mathbf{y} = f(\mathbf{x}, \boldsymbol{\phi}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{x}$ is an input variable (time in our applications), f is a mean function with parameters $\boldsymbol{\phi}$ and $\boldsymbol{\epsilon}$ is an independent and identically distributed Gaussian noise process. For a GP, we can write the joint probability distribution of $\mathbf{y}$ as:

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \mathcal{N}(f(\mathbf{x}, \boldsymbol{\phi}), \boldsymbol{\Sigma}). \tag{2}$$

Here, $\boldsymbol{\Sigma}$ is the covariance matrix and $\boldsymbol{\theta}$ are the parameters of a kernel function (often refereed to as the hyperparameters of the GP) and we can write the log marginal likelihood explicitly as:

$$\log \mathcal{L}(\mathbf{r}|\mathbf{x}, \boldsymbol{\phi}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{r}^T\boldsymbol{\Sigma}^{-1}\mathbf{r} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log(2\pi), \tag{3}$$

where $\mathbf{r} = \mathbf{y} - f(\mathbf{x})$ is the vector of residuals from the mean function. Each entry in the covariance matrix is populated by the kernel function which describes the correlations between nearby data points. Hence, a GP is a distribution over functions. We can form the joint probability distribution of our training data and some new test data y_*:

$$p\left(\begin{bmatrix}\mathbf{y}\\ y_*\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}(\mathbf{x})\\ f(x_*)\end{bmatrix}, \begin{bmatrix}\mathbf{K}(\mathbf{x},\mathbf{x}) & \mathbf{K}(\mathbf{x},x_*)\\ \mathbf{K}(x_*,\mathbf{x}) & k(x_*,x_*)\end{bmatrix}\right), \tag{4}$$

where $\mathbf{K}(\mathbf{x},\mathbf{x})$ is the covariance matrix of the training data, $\mathbf{K}(\mathbf{x},x_*)$ is the column vector formed from the elements $\mathrm{k}(\mathrm{x}_1,\mathrm{x}_*)$, …, $\mathrm{k}(\mathrm{x}_n,\mathrm{x}_*)$, and $\mathbf{K}(\mathrm{x}_*,\mathbf{x})$ is its transpose. $\mathrm{k}(\mathrm{x}_*,\mathrm{x}_*)$ is the scalar covariance of the test point with itself, i.e., the variance. From (4), and using the standard results for conditioning multivariate Gaussian distributions, we obtain the joint posterior. It is then straightforward to extend this to an arbitrary number of test points.

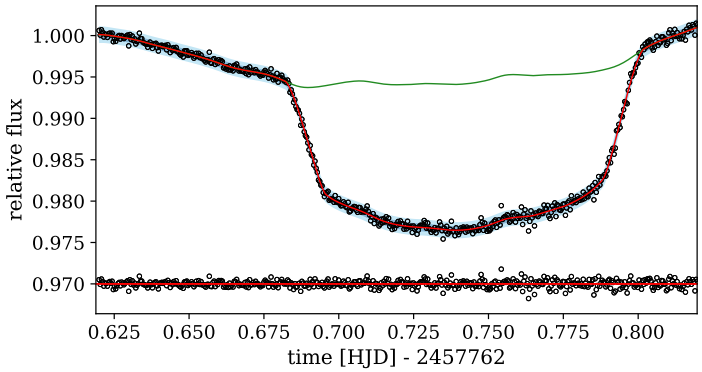


FIGURE 1. An example white light curve for a typical "Hot Jupiter" exoplanet. The red line shows the best fitting model and the green line shows the systematics model derived from the GP fit. Residuals are indicated below the light curve.

The marginal likelihood given by (3) may also be optimised to infer the hyperparameters of interest, which may include either the GP hyperparameters or the parameters of a mean function. It is also relatively straightforward to embed GPs within a Markov chain Monte Carlo (MCMC) framework in order to infer the full posterior distributions of the hyperparameters.

3 Applications to Transmission Spectroscopy

Transmission Spectroscopy involves measuring the wavelength-dependent absorption of starlight by a planet's atmosphere as it transits its host star. These are extremely challenging measurements, as the typical signal is usually dwarfed by systematic effects in the light curves, necessitating the use of sophisticated techniques to statistically model and remove them. Using a GP, we can simultaneously model the deterministic transit function whilst placing a distribution over possible functions to model the correlated noise. Hence, we can marginalise out our uncertainty in the GP hyperparameters and robustly infer the probability distributions of the transit function parameters (Wilson, 2021). Using this technique, we can begin to build up a picture of the measured transit depth as a function of wavelength (commonly referred to as the transmission spectrum). Such information allows us to infer the atomic and molecular composition of the planet's atmosphere and can also be used to infer the presence of clouds and scattering particles (Sing, 2016). Figure 1 shows an example of a transit light curve for a typical "Hot Jupiter" exoplanet fitted with a GP model.

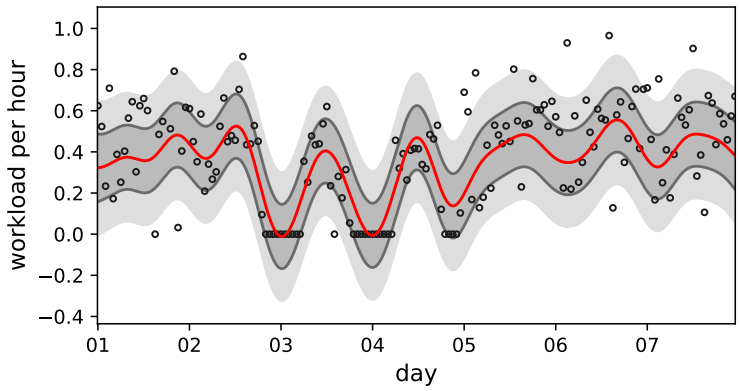


FIGURE 2. A GP model which has been fit to historical workload data in an industrial setting. The resulting model is then used as a template to predict and assess future performance.

4 Applications to Industrial Data

In an industrial and manufacturing setting, the versatility of GPs allows them to be applied to a multitude of problems involving either interpolation or prediction, particularly when we may not know the specific functional form of the process. In particular, data generated from machine sensor and Internet of Things (IoT) technologies can be used to gain insights into a wide range of processes and these insights are often of significant interest given their potential to increase efficiencies and optimise production. In our presentation we will demonstrate our use of GPs for modelling and predicting machine performance, monitoring critical machine parameters such as temperature, current or voltage and for predicting machine breakdown. For example, Figure 2 shows a GP model for hourly machine workload over a weekly period which has been trained on historical data. This template model can then be compared to new data in order to assess machine performance. Such models may also be used to create dashboard visualisations for real time feedback.

As a further example, in Figure 3 we show another typical industrial application - the statistical modelling of critical machine parameters. Here, we show a number of GP fits to simulated data which is inspired by a real industrial process having three input parameters (A, B, and C). These parameters may represent inputs common in such a setting such as temperature, vibration, current etc. GP models are well-suited to fitting such varied data and are capable of providing accurate predictions and robust uncertainty estimates. Such modelling may be used to gain useful insights into the behaviour of the parameters over time, or for extrapolating into the future for predictive tasks.

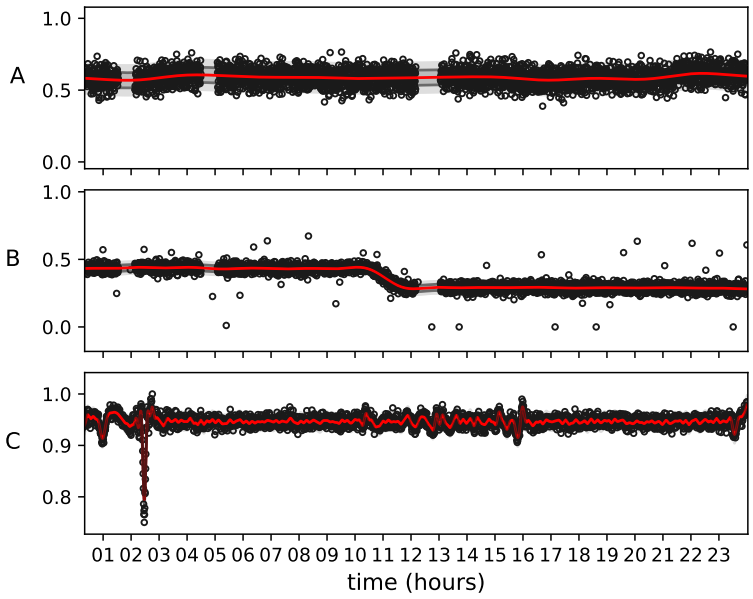


FIGURE 3. Example highlighting the ability of Gaussian processes to model many varied processes. The data in the individual panels are inspired by a real industrial process that has three input parameters A, B, and C. Black circles show the simulated data, whilst the red lines show the best-fitting GP models.

5 Discussion

Gaussian process models are both a powerful and flexible technique for Bayesian non-parametric regression and they provide an effective approach to statistically modelling many diverse phenomena from transiting exoplanets to a multitude of processes frequently encountered in an industrial and manufacturing context. Having briefly described the conceptual framework for Gaussian processes, we will describe some of the uses that we have found for GP models in our presentation, including illustrative examples of light curve modelling, modelling critical manufacturing process parameters, analysing and assessing machine performance and for identifying unusual behaviour and predicting rejected parts. These examples highlight the versatility of Gaussian processes for modelling diverse phenomena and their ability to enable data-driven decision-making in the presence of uncertainty. With a wide variety of kernel functions available to model various processes (e.g. periodic, quasi-periodic, long-term drifts etc.), GPs provide a very flexible and effective approach to statistical modelling. Such applications, typically incorporating data generated by machine sensor and Internet of Things (IoT) technologies, have the potential to increase process efficiencies, reduce manufacturing costs and optimise production.

Acknowledgments: This work was supported by the Confirm Smart Manufacturing Centre (https://confirm.ie/) funded by Science Foundation Ireland (Grant Number: 16/RC/3918).

References

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

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York City: Springer.

Gibson, N. P. et al. (2012). A Gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy. *Monthly Notices of the Royal Astronomical Society*, **419**, 2683 – 2694.

Wilson, J. et al. (2021). Gemini/GMOS optical transmission spectroscopy of WASP-121b: signs of variability in an ultra-hot Jupiter? *Monthly Notices of the Royal Astronomical Society*, **503**, 4787 – 4801.

Sing, D. K. et al. (2016). A continuum from clear to cloudy hot-Jupiter exoplanets without primordial water depletion *Nature*, **529**, 59 – 62.

A multilevel multivariate response model for data with latent structures

Yingjuan Zhang[1], Jochen Einbeck[2], Reza Drikvandi[1]

[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 propose a two-level extension of a previously introduced multivariate latent variable model, which allows incorporating covariates on both levels. The presented model accounts for correlations among the response variables through univariate random effects which are modelled using a mixture distribution. We estimate the model parameters via an EM algorithm and provide simulation results and a real data application.

Keywords: Mixture distribution; Multivariate response model; Posterior intercepts; Random effects.

1 Introduction

The use of multivariate response models is not very widespread in statistical practice. This may be related to the circumstance that ready-to-use implementations are either only accessible via specialized software (such as SAS), or are equivalent to fitting separate univariate response models (such as R function `lm`). However, accounting for the multivariate response character has several inferential benefits including potentially increased powers. Zhang and Einbeck (2022) introduced a versatile latent variable model for dimension reduction and simultaneous clustering of multivariate data. However, their model did not allow for the inclusion of covariates and could not deal with repeated measures. This paper aims to provide such extensions. We consider a scenario where multivariate data $x_{ij} \in \mathbb{R}^m$ has a two-level structure, with the upper level indexed by $i = 1, 2, ..., r$ and the lower level by $j = 1, 2, ..., n_i$. The proposed two-level model takes the form

$$x_{ij} = \alpha + \beta z_i + \Gamma v_{ij} + \varepsilon_{ij}, \tag{1}$$

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

where $\alpha, \beta \in \mathbb{R}^m$, $z_i \in \mathbb{R}$, $v_{ij} \in \mathbb{R}^p$ is the vector of covariates (which may include upper-level variates not depending on j), $\Gamma \in \mathbb{R}^{m \times p}$ is a matrix of coefficients, and ε_{ij} are independent Gaussian errors (if there is only one covariate, $v_{ij} \in \mathbb{R}$, we write $\Gamma = \gamma \in \mathbb{R}^m$). Under such a model, the data grouping process is carried out on the upper level, while the lower level units within the same upper level unit share a common random effect z_i. Model (1) does not require the normality of random effects so no concerns to check the random-effects distribution (e.g., Drikvandi et al 2017).
Figure 1 illustrates a data scenario corresponding to this concept. The data used here is simulated from model (1) in the case that the latent variable obeys a three-point mixture distribution. The grey straight line represents the one-dimensional latent space $\alpha + \beta z$, and the black triangles positioned along the straight line the mixture centres of each component. The coloured thinner lines are for illustration only and show the trend of lower-level units within each each upper level (which is to some extent a result of the random error and to some part driven by the covariate). The orange triangles are the fitted values: $x^*_{ij} = \hat{\alpha} + \hat{\beta} z^*_i + \hat{\gamma} v_{ij}$, where $z^*_i = \sum_{k=1}^{K} w_{ik} \hat{z}_k \in \mathbb{R}$ are obtained as the posterior random effects using posterior probabilities of component membership w_{ik} (Aitkin, 1996).

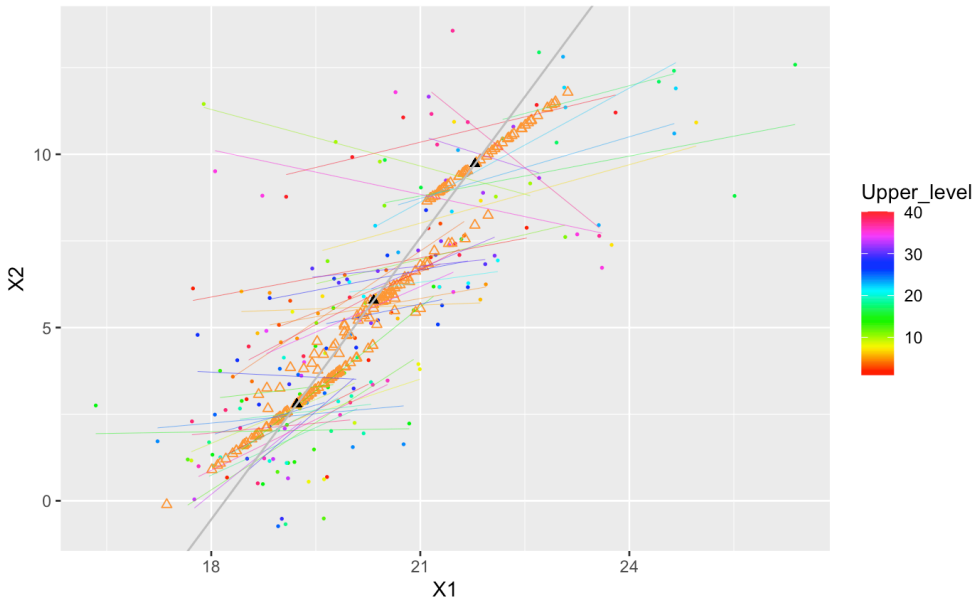


FIGURE 1. Simulated data with 40 upper level units, each with 5 lower level units, with $\alpha = (20, 10)$, $\beta = (1, 3)$, $\pi_k = (0.2, 0.3, 0.5)$, $z_k = (1.73, 0.29, -0.87)$, $\gamma = (0.5, 1)$. Observations are generated with component-specific diagonal variance matrices Σ_k. (We avoid the use of the term 'cluster' since this has a different connotation in the context of repeated measures.)

2 Methodology

We conduct the parameter estimation using maximum likelihood method. Since the component membership of each upper unit is unknown, we consider this as an 'incomplete data' problem, and apply the EM algorithm. The required complete data likelihood takes the shape

$$L_c = \prod_{i=1}^{r}\prod_{j=1}^{n_i}\prod_{k=1}^{K}(f_{ik}\pi_k)^{G_{ik}},$$

where G_{ik} is an indicator variable taking the value 1 if upper unit i belongs to component k. We specify a multivariate Gaussian model for the component-specific densities f_{ik} in model (1) as

$$f_{ik} = \frac{1}{(2\pi)^{m/2}}\frac{1}{|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})^T\Sigma_k^{-1}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})\right).$$

The expected complete log-likelihood is then given by

$$l = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}\log(\pi_k) + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}\log(|\Sigma_k|) + \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} -\frac{m}{2}\log(2\pi)w_{ik}$$
$$+\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} -\frac{1}{2}w_{ik}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij})^T\Sigma_k^{-1}(x_{ij}-\alpha-\beta z_k-\Gamma v_{ij}),$$

where Σ_k is a component-specific diagonal variance matrix, and $w_{ik} = \frac{\pi_k f_{ik}}{\sum_l \pi_l f_{il}}$ is the probability of upper unit i belonging to component k. The computation of w_{ik} is via the E-step. The parameters α, β, z_k, Σ_k, and Γ will be estimated through the M-step. The key parameter estimates are:

$$\hat{z}_k = \frac{\sum_{i=1}^{r} w_{ik}\sum_{j=1}^{n_i}\hat{\beta}^T\hat{\Sigma}_k^{-1}(x_{ij}-\hat{\alpha}-\hat{\Gamma}v_{ij})}{\sum_{i=1}^{r} n_i w_{ik}\hat{\beta}^T\hat{\Sigma}_k^{-1}\hat{\beta}},$$

and

$$\hat{\Gamma} = \left(\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{k=1}^{K} w_{ik}(x_{ij}-\hat{\alpha}-\hat{\beta}\hat{z}_k)v_{ij}^T\right)\left(\sum_{i=1}^{r}\sum_{j=1}^{n_i} v_{ij}v_{ij}^T\right)^{-1}.$$

3 Real data application

The real data used here is obtained from the International Adult Literacy Survey (IALS), collected in 13 countries on Prose, Document, and Quantitative scales between 1994 and 1995. The data are reported as the percentage of individuals who could not reach a basic level of literacy in each country. Based on the Prose scale only, Sofroniou et al (2008) used these data

TABLE 1. Posterior probabilities and intercepts for the IALS data. In the column 'mass points', the first two rows give estimated $\hat{\pi}_k$ and $\hat{z}_k$.

		Mass points			
		0.2308	0.5391	0.1532	0.0769
Country	posterior intercept	-1.1576	-0.0819	0.5904	2.8703
Sweden	-1.15760	1.0000	0.0000	0.0000	0.0000
Germany	-1.15756	1.0000	0.0000	0.0000	0.0000
Netherlands	-1.15754	0.9999	0.0001	0.0000	0.0000
Canada	-0.08188	0.0000	1.0000	0.0000	0.0000
Australia	-0.08188	0.0000	1.0000	0.0000	0.0000
Switzerland(French)	-0.08188	0.0000	1.0000	0.0000	0.0000
New Zealand	-0.08173	0.0000	0.9998	0.0002	0.0000
Belgium(Flanders)	-0.08163	0.0000	0.9996	0.0004	0.0000
Switzerland(German)	-0.08114	0.0000	0.9989	0.0011	0.0000
United States	-0.08036	0.0000	0.9977	0.0023	0.0000
Ireland	0.58386	0.0000	0.0098	0.9902	0.0000
United Kingdom	0.58912	0.0000	0.0019	0.9981	0.0000
Poland	2.87028	0.0000	0.0000	0.0000	1.0000

to rank countries according to their posterior intercepts $z_i^* = \sum_{k=1}^{K} \hat{z}_k w_{ik}$. We analyze the data considering the 3-variate response Prose, Document, and Quantitative, additionally including the lower-level covariate gender in the model; i.e. $m = 3$, $p = 1$ and $\Gamma = \gamma \in \mathbb{R}^3$.

The country-specific random effect z_i accounts for the correlation among the observations within upper-level units and the correlation among the three response dimensions of the model. We fit the model with $k = 4$ mass points and component-specific diagonal variances Σ_k, leading to an AIC value of 235.5 which does not drop significantly when increasing k further or with other variance parametrizations. Table 1 presents the joint ranking via the posterior random effect and classification of the countries. The table shows that Sweden, Germany, and the Netherlands are assigned to mass point 1 with the smallest number of people being illiterate. Poland is the only country that is assigned to the high illiteracy mass point 4. The US and Ireland have posterior probabilities that spread across two mass points but are assigned to different components. Using all three measurements as a multivariate response, the component allocation of each country is more decisive compared to the results using just Prose (Sofroniou et al, 2008).

4 Simulation study

We conduct a simulation study to examine the performance of our method. Another objective of this simulation is to investigate whether an increase in the number of upper- or lower-level units will effectively reduce the variance in the parameter estimates. We first consider a scenario with $r = 50$ upper

level units and $n_i = 5$ lower level units, for $i = 1, 2, \ldots, r$. This will be the baseline experiment. Then we keep $r = 50$ unchanged and increase the number of lower-level units to be $n_i = 10$, for $i = 1, 2, \ldots, r$. We consider another sample size with lower-level units $n_i = 5$ for $i = 1, 2, \ldots, r$ unchanged but increase the upper-level units to be $r = 100$. We generate 200 replicates from the model (1) with one lower level covariate in all three scenarios, with the covariate generated from a normal distribution with a mean of 0.3 and a standard deviation of 0.2. The results indicate that when we increase the upper-level units, the parameters' RMSE decreases stronger than when increasing the lower-level units. Then we further increase the upper level units to be $r = 200$ and keep the lower level units $n_i = 5$ for $i = 1, 2, \ldots, r$. The key results are shown in Figure 2, Table 2 and Table 3.

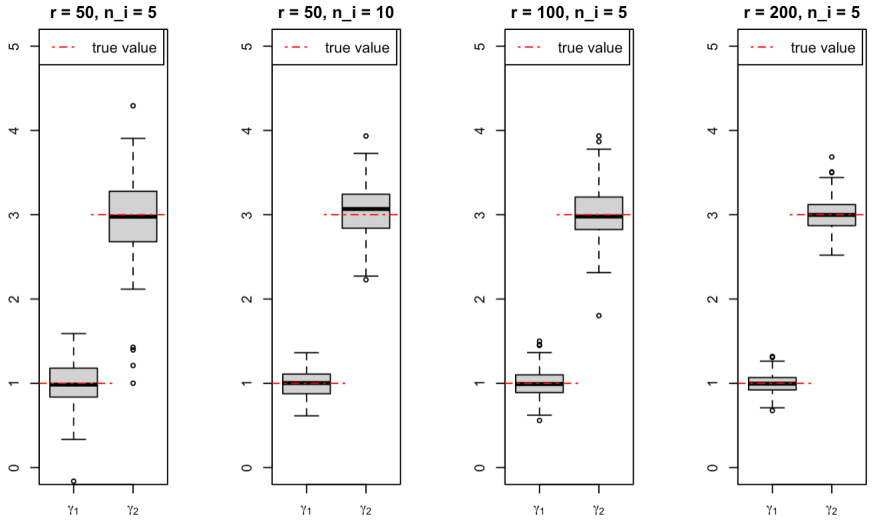


FIGURE 2. Estimates of key parameter γ with different number of upper-level and lower-level units.

5 Conclusion

This paper provides an extended random effect model that applies to two-level multivariate response data with latent structures. An EM algorithm is used for parameter estimation. In particular, a nonparametric maximum likelihood method (Aitkin 1999) is used for estimation of the random effect where the mass points z_k and their weights π_k, $k = 1, 2, \ldots, K$ are treated as unknown parameters to be estimated in the EM algorithm. An application of constructing a league table using the IALS data is provided. Another application is to fit multivariate response models.

TABLE 2. Estimates of key parameters γ, z_k and α with different upper-level and lower-level units.

		Average estimates			
	True	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
γ_1	1.000	0.989	0.993	0.991	0.995
γ_2	3.000	3.036	2.972	3.009	2.998
z_1	-0.816	-0.807	-0.814	-0.820	-0.809
z_2	1.225	1.268	1.258	1.234	1.246
α_1	2.000	2.037	2.039	2.034	1.991
α_2	10.000	10.020	10.007	10.019	10.002

TABLE 3. RMSE for key parameters γ, z_k and α with different upper-level and lower-level units.

	RMSE			
	$r = 50, n_i = 5$	$r = 50, n_i = 10$	$r = 100, n_i = 5$	$r = 200, n_i = 5$
γ_1	0.269	0.167	0.166	0.118
γ_2	0.474	0.297	0.296	0.194
z_1	0.124	0.124	0.084	0.068
z_2	0.198	0.207	0.133	0.132
α_1	0.464	0.447	0.302	0.232
α_2	0.172	0.161	0.120	0.088

References

Aitkin, M. (1996). Empirical bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In: *Proc's of the 11th IWSM*, Orvieto, Italy, 87 – 94.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55(1), 117 – 128.

Drikvandi, R., Verbeke, G. and Molenberghs, G. (2017). Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics*, 73(1), 63 – 71.

Sofroniou, N., Hoad, D., and Einbeck, J. (2008). League tables for literacy survey data based on random effect models. In: *Proc's of the 23rd IWSM*, Utrecht, Netherlands, 402 – 405.

Zhang, Y., and Einbeck, J. (2022). Simultaneous linear dimension reduction and clustering with flexible variance matrices. In: *Proc's of the 36th IWSM*, Trieste, Italy, 612 – 617.

Flexible modelling of time-varying training exposures on the risk of recurrent injuries in football

Lore Zumeta-Olaskoaga[1,2], Andreas Bender[3], Dae-Jin Lee[4]

[1] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain
[2] Departamento de Matemáticas, Universidad del País Vasco UPV/EHU, Spain
[3] Statistical Consulting Unit StaBLab, Ludwig-Maximilians Universität München, Munich, Germany
[4] School of Science and Technology, IE University, Madrid, Spain

E-mail for correspondence: lzumeta@bcamath.org

Abstract: Football players are repeatedly exposed to high competition demands, that in turn increase the sports burden applied to them, as well as exposure to the risk of injury. In this regard, we present a flexible modelling approach to estimate the effect of training load on the risk of injury. This model, in contrast to other modelling alternatives proposed in the sports injury research literature, considers that a player can sustain subsequent injuries, and that training load might vary over time, intensity and duration.

Keywords: survival analysis; piece-wise exponential additive mixed models; sports analytics; football injuries; recurrent events.

1 Motivation

In football, a large amount of data is now collected, including data related to external training load (e.g. training and competition time, power output, distance, sprints, speed etc.) that are tracked by Global Positioning System (GPS) devices. As such, the study of training load and its role in injury prevention is one of the hot topics in sports injury prevention research.
A good understanding of the training load is key to developing effective training plan strategies that will enhance players' performance while also lowering their risk of injury. And this requires establishing an etiologically plausible time-varying exposure model, which defines how previous training affects the hazard of injury. Further, the model must consider that

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

recurrent injuries may be associated within players. To take into account the latter, i.e. dependencies induced by injury recurrence, as well as the intensity, duration and timing of past exposures, we propose the use of a recurrent time-to-event flexible modelling approach with weighted cumulative exposure (WCE) effects (Sylvestre and Abrahamowicz 2009).

2 The model

We based on the piece-wise exponential mixed modelling (PAMMs) framework (Bender *et al.* 2018) and adapt its general formulation to build on a recurrent events PAMM model with time-dependent covariates as WCE-type cumulative effects. Then, the hazard rate of the i-th injury (event) of the l-th player, given the player's training exposure history $\boldsymbol{z}(t) = \{z(t_z) : t_z \leq t\}$, is expressed as:

$$\begin{aligned}\lambda_{l,i}(t|\boldsymbol{z}_{l,i}(t), b_l) &= \lambda_0(t)\exp\{g(\boldsymbol{z}, t) + b_l\} = \\ &= \exp\{\beta_0 + f_0(t_j) + g(\boldsymbol{z}, t) + b_l\} \end{aligned} \tag{1}$$

for all $t \in (\kappa_{j-1}, \kappa_j],\ t > 0$ (and e.g. $t_j := \kappa_j$), where $\kappa_j,\ j = 0, \ldots, J$, are the $J+1$ cut points defining J intervals that partition the study follow-up $(0, t_{\max}]$. In Eq.(1), we have:

- The log-baseline hazard, $\beta_0 + f_0(t_j)$, where $f_0(t_j)$ is expressed as a smooth term of the form $\sum_{m=1}^{M} \gamma_{0m} B_m(t_j)$.
- The term $g(\boldsymbol{z}, t)$ denotes that past exposure effects of $\boldsymbol{z}$ cumulate over time, over a relevant time-window $\tau(t)$, being the cumulative effect of $\boldsymbol{z}$ at time t the sum of all weighted effects of past observations.
- A Gaussian random effect associated to player l, $b_l \sim N(\boldsymbol{0}, \sigma_b^2)$.

For a WCE-type effect, in Eq. (1), we consider time-varying exposure effects weighted by latency $t - t_z$ and linear on $z(t_z)$. That is, the contribution of covariate z observed at time t_z with value $z(t_z)$ is defined by $h(t, t_z, z(t_z)) := h(t - t_z)z(t_z)$.
Let $\boldsymbol{z}(t) = \{z(t_z) : t_z \leq t\} = \{z(t_{z,1}), \ldots, z(t_{z,Q})\}$ be the set of all registered training load variables at time t. Then, $g(\boldsymbol{z}, t)$ is estimated with P-splines (Eilers and Marx, 1996) as follows:

$$\int_{\tau(t)} h(t - t_z)z(t_z)dt_z \approx \sum_{q=1}^{Q} \tilde{\Delta}_q \tilde{h}(t - t_{z,q}) = \sum_{q=1}^{Q} \tilde{\Delta}_q \sum_{m=1}^{M} \gamma_{1m} B_m(t - t_{z,q})$$

with $\tilde{\Delta}_q = z(t_z)(t_{z,q} - t_{z,q-1})$ if $t_{z,q} \in \tau(t)$ and 0 otherwise.

TABLE 1. Simulation results for $N_{\text{sim}} = 500$ replicates of the estimations of $h_{t,t_z,z(t_z)}$ and σ_b, in each scenario setting in terms of mean RMSE and mean coverage of 95% confidence intervals over each time-point $t_z = 1, \ldots, 40$.

Data generation mechanism		RMSE		Coverage
WCE shape	Heterogeneity	$h_{t,t_z,z(t_z)}$	σ_b	$h_{t,t_z,z(t_z)}$
Exponential decay	$\sigma_b = 0.05$	0.030	0.134	0.93
	$\sigma_b = 0.5$	0.032	0.166	0.92
	$\sigma_b = 1$	0.034	0.354	0.93
Bi-linear	$\sigma_b = 0.05$	0.032	0.138	0.91
	$\sigma_b = 0.5$	0.032	0.168	0.92
	$\sigma_b = 0.1$	0.034	0.346	0.92
Early peak	$\sigma_b = 0.05$	0.046	0.135	0.80
	$\sigma_b = 0.5$	0.047	0.176	0.78
	$\sigma_b = 1$	0.048	0.361	0.78

3 Simulation study

To evaluate the model performance, we simulated $N_{\text{sim}} = 500$ times a cohort of $L = 500$ individuals with exposures recorded $t_{z,1} = 1, t_{z,2} = 2, \ldots, t_{z,Q} = 40$ ($Q = 40$) days before the time at which we model the hazard, drawing survival times from the piece-wise exponential distribution, under three different true weight functions, each defined over a $[0, Q]$ interval, see Figure 1, and under three different levels of heterogeneity between recurrent events, $\sigma_b \in \{0.05, 0.5, 1\}$. We refer to Table 1 for the summary of the results.

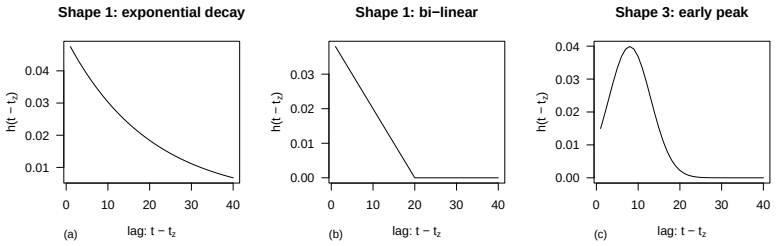


FIGURE 1. Each of the true weight functions **(a)-(c)** considered.

4 Application

We applied the proposed model to an observational injury data from an elite male football team participating in LaLiga during the 2017-2018 and 2018-2019 seasons. A total of 36 players were followed-up, over 150 variables

(external training load variables) were registered on a regular basis through tracking devices and 72 non-contact time-loss injuries occurred among 23 players. We were interested in modelling "how the cumulative stress placed on a player from multiple training sessions and matches, over a period of time, affects his risk of a (recurrent) football injury". In this regard, we defined the follow-up time (t) unit as well as the exposure time (t_z) unit as the "number of match and training sessions" (i.e. the n-th session). We analysed the effect of the "Average Metabolic Power" (z_{AvgMP}) training load variable, and set the lag-lead window to be $\tau(t) = \{t_z : t \geq t_z \wedge t < t_z + 11\}$, that means that all z_{AvgMP} that were observed in the last 10 sessions prior to t or at t can affect the hazard at time t. Figure 2 shows the estimated cumulative effect of z_{AvgMP} on the log-hazard scale.

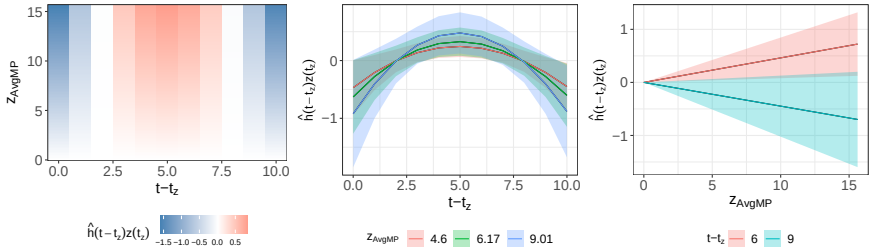


FIGURE 2. Visualization of the non-linearly varying cumulative effect of the variable "Average Metabolic Power" (z_{AvgMP}) on the log-hazard scale (left) and one-dimensional slices with respect to the covariate $z(t_z) \in \{4.6, 6.17, 9.01\}$ (middle) and the latency $t - t_z \in \{6, 9\}$ (right).

5 Conclusion and further work

The modelling framework presented provides a suitable way to flexibly model training load exposures and analyse their effect on recurrent football injuries, with respect to other alternative measures of training load exposures used in the literature. The validity of the approach was supported by the simulation study and applied to football injury data. In the case study, we got that the player-related random effect term was significant and that the "average metabolic power" recorded 5 sessions earlier had the greatest impact on the current hazard at time t. The lag-lead window used was defined based on experts' criteria, but it could be improved by selecting the optimal window with an additional penalty parameter.

Acknowledgments: We thank Medical Services of Athletic Club for data support. We acknowledge the support of the Basque Government through

the BERC 2022-2025 program; of the Ministry of Science, Innovation and Universities through BCAM Severo Ochoa accreditation and through SEV-2017-0718 PRE2018-084007 funding; of AEI/FEDER, UE through the PID2020-115882RB-I00 and acronym "S3M1P4R"; and of Provincial Council of Bizkaia through the 6/12/TT/2022/00006 and acronym "MATH4-SPORTS".

References

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

Bender, A., Scheipl, F., Hartl, W., Day, A. G., & Küchenhoff, H. (2019). Penalized estimation of complex, non-linear exposure-lag-response associations. *Biostatistics*, **20**(2), 315 – 331.

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

Sylvestre, M. P., & Abrahamowicz, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Statistics in medicine*, **28**(27), 3437 – 3453.

Part III

Modelling single-nucleotide polymorphism to assess genetic contribution to disease progression

Mazin Aouf[1], Kenan M. Matawie[1]

[1] School of Computer, Data and Mathematical Sciences, Western Sydney University, Australia

E-mail for correspondence: 16006690@student.westernsydney.edu.au

Abstract: This paper investigates the association between the Apolipoprotein E4 single nucleotide polymorphism (SNP) and brain Tensor Based Morphometry (TBM) as a potential candidate gene for Alzheimer's disease. The Linear Mixed Model (LMM) is used to analyse and model the data. The imputed data was used to compare the results of the LMM to those obtained from the complete dataset, allowing for the selection of the optimal model and the assessment of missing values' impact on performance. Ultimately, the LMM with a random intercept and random slope was found to be the most representative model. These models and such analysis could have important implications for understanding the underlying mechanisms of Alzheimer's disease. This study provides valuable modeling insights into the relationship between genetic factors, brain structure, and the development of Alzheimer's disease.

Keywords: Longitudinal data; Linear Mixed Model; Brain TBM; SNPs; Missing data.

1 Introduction

Genome-wide association studies (GWAS) and Tensor Based Morphometry (TBM) are techniques used in genetics and brain research to identify genetic variants, such as single-nucleotide polymorphisms (SNPs), and measure structural changes in the brain over time using magnetic resonance imaging (MRI), respectively. The linear mixed model (LMM) analyses longitudinal data by modelling fixed and random effects to understand how factors influence outcomes over time. Integrating genetic data with TBM and LMM data can improve dementia treatments by aiding in the understanding of

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

the brain structure, function, and treatment effects. SNPs set analysis in GWAS has grown (Park et al., 2019; Sikorska et al., 2013). Researchers have also used different statistical approaches such as Bayesian methods to analyse longitudinal data (Ariyo et al., 2020). The objective of this paper is using the LMM to further analyse and examine the relationship between TBM and SNPs over time to identify the factors that contribute to the development of AD.

Alzheimer's disease Neuroimaging Initiative (ADNI) is a multisite that provides researchers with patient study data in order to track the progression of the AD. Data of 806 patients was collected and prepared for this study. Each patient was tested repeatedly using multiple scans over the time. This data consisted of 3 main attributes: the time periods after 3, 6, 12 and 24 months of the baseline MRI scan, the TBM score (difference between the scores of the current scan and the baseline) and the SNPs values. After merging the original dataset with the TBM dataset, the longitudinal data was generated to have 2575 records (Table 1).

TABLE 1. TBM scores over the time.

Time	Patient	Percent Scan		Mean TBM	SD TBM
0	806	Yes	No	-0.0120	0.0329
03	742	92%	8%	-0.0136	0.0207
06	668	89%	11%	-0.0129	0.0138
12	653	97.7%	2.3%	-0.0103	0.0096
24	512	72.46%	27.54%	-0.0124	0.0222
	2575	89.33%	10.66%	-0.0123	0.0221

Longitudinal studies frequently encounter missing values due to various factors such as illness or personal circumstances, resulting in incomplete patient scan data. To address this issue, researchers may choose to either exclude the incomplete observations or employ statistical methods to impute missing values. Multiple studies, including Molenberghs and Verbeke (2001) and Donders et al. (2006), have concluded that Missing at Random (MAR) missingness is commonly encountered in clinical data. Numerous studies, such as White et al. (2011), have demonstrated the high effectiveness of the Multiple Imputation by Chained Equations (MICE) method in handling missing values in clinical data.

2 Longitudinal data modelling

Longitudinal data variables are measured repeatedly at different points over time. In addition, time is a crucial factor in this study to understand TBM scores variation, as a response component. The latter is associated to SNPs with respect to Time. Accordingly, LMM were proposed to identify SNPs

association with evolution of TBM score over time. The model is expressed as follows:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \quad i = (1, \ldots, n). \tag{1}$$

where n is the total number of patients (806); Y_i is a $m_i \times 1$ vector of TBM values for the ith patient; X_i is the $m_i \times p$ matrix of fixed effects corresponding to the p predictor variables (Time, SNP and interaction between them); β is a $p \times 1$ column vector of the fixed effect regression coefficients; Z_i is the $m_i \times q$ design matrix of the q random effects that describe the subject-specific of time for the ith patient; b is a $q \times 1$ vector of random effects. Here $q = 2$ yielding a random intercept b_0 and slope b_1). Finally, ϵ is $m_i \times 1$ vector of residual effects.

3 Results and discussion

To explore the association between TBM and SNPs over time, it is necessary to include an interaction term between the two variables in the LMM analysis. To address the missing values in the dataset, two approaches were employed: the *lme*4 R package with maximum likelihood estimation and the MICE technique. The best LMM model generated from both analysis is presented below:

$$\begin{aligned} TBM_{ij} = \beta_0 + \beta_1 SNP_{ij} + \beta_2 Time_{ij} + \\ \beta_3 MMSE_{ij} + \beta_4 SNP_{ij} \times Time_{ij} + b_{0i} + b_{1i} Time_{ij} + \epsilon_{ij} \end{aligned} \tag{2}$$

With i the patient and j the number of the MRI scan (1 to 4) taken at $Time_{ij}$ equal to months 3, 6, 12 and 24. The SNP_{ij} variable indicates the number of minor alleles and can take values from the interval 0 to 2. In order to explore the relationship between TBM and SNP, several multiple linear mixed models (LMMs) were employed. These models incorporated additional variables to enhance robustness and ensure accurate interpretation of estimates. Moreover, special attention was given to including highly correlated variables to avoid violating the assumptions of the LMMs.

The initial analysis using the *lme*4 package indicated that the model with both a random intercept and a random slope (equation 2) provided the best fit, as evidenced by higher R-squared values and lower AIC and BIC values (see Table 2). The results of the analysis showed that the SNPs and MMSE (Mini-Mental State Examination) variables were statistically significant (p-value < 0.05), indicating a significant relation between these predictors and the TBM variable. However, the time variable and the interaction between SNP and time were not found to be statistically significant, suggesting that these predictors did not have a significant influence on the TBM variable.

Next, the Multiple Imputation by Chained Equations (MICE) imputation technique was employed to handle missing data assumed to be Missing at Random (MAR) using the R package $MICE$. This approach allowed us to

TABLE 2. Summary statistics of the three LMMs

Model	R^2	AIC	BIC
Random intercept and fixed slope	58.3%	-10604	-10566
Fixed intercept and random slope	18.6%	-10273	-10235
Random intercept and random slope	71.8%	-10983	-10934

impute the missing values in the dataset, resulting in an augmented dataset 806 patients with 4 visits each, totaling 3224 records.
Subsequently, Linear Mixed Models (LMMs) were conducted on the imputed dataset. The same model (equation 2) with both a random intercept and a random slope, which was previously determined to be the best fit, was used for analysis. The LMM analysis on the imputed dataset yielded similar results to the previous analysis, with one notable difference, the Time variable and the response variable (TBM) were found to be statistically significant (p-value < 0.05) in this analysis. Additionally, the model demonstrated a higher R-squared value and smaller AIC and BIC values compared to the previous model. Detailed information regarding the model parameters can be found in Table 3. These findings indicate that the MICE imputation technique improved the model fit and provided more accurate results.

TABLE 3. Parameters estimation for imputed missing data.

Effect	Parameter	Estimate-Value	t-values	p-values
Intercept	β_0	-5.219e-02	-12.991	$<$ 2e-16
SNP	β_1	-4.621e-03	-3.386	0.000743
Time	β_2	1.168e-04	2.413	0.015961
MMSE	β_3	1.497e-03	11.081	$<$ 2e-16
SNP x Time	β_4	7.847e-05	1.402	0.161293

The LMM assumptions were evaluated for the best model selected (equation 2). Figure 1 presents a Q-Q plot (quantile-quantile plot) to assess whether the residuals of the best model (equation 2) are normally distributed. While there is only slight deviation from normality, it appears that the distribution possesses heavier tails compared to a normal distribution. As a result, the presence of light tails on either end suggests that alternative modelling approaches may be worth considering to satisfy the normality assumption of LMM more accurately.

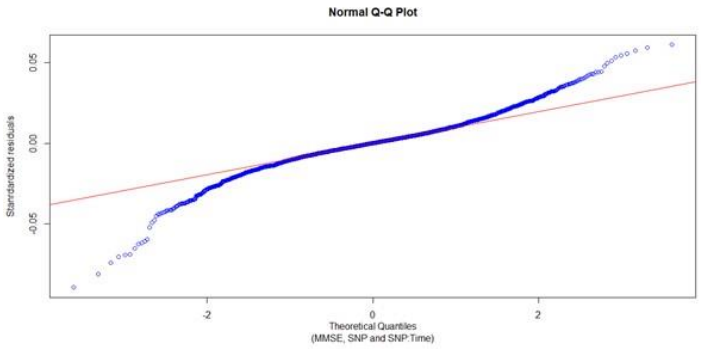


FIGURE 1. Normal Q-Q plot of residuals.

4 Conclusion

This study investigated the use of linear mixed models (LMMs) to analyse longitudinal data from the Alzheimer's disease Neuroimaging Initiative (ADNI) dataset. The analysis aimed to explore the relationship between a specific genetic marker, Apolipoprotein E4 (APOE4) single nucleotide polymorphism (SNP),and brain atrophy using Tensor-based Morphometry (TBM) method over time. The findings revealed that a LMM with both random intercept and random slope provided the best fit for the data. This approach handled missing data and enhanced our understanding of Alzheimer's disease progression. The study highlights the importance of considering genetic factors and demonstrates the effectiveness of LMMs in analysing longitudinal data in the context of Alzheimer's disease research. Future studies can utilise advanced modelling techniques to explore the relationship between genetic markers and brain atrophy in Alzheimer's disease, such as nonlinear mixed effects models, and other advanced approaches that can offer deeper insights into the complex dynamics and facilitate further analysis of the relationship between this biomarker and TBM.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19(6)**, 716 – 723.

Ariyo, O., Quintero, A., Muñoz, J., et. al. (2020). Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics*, **47(5)**, 890 – 913.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). A gentle introduction to imputation of missing values.*Journal of clinical epidemiology*, **59(10)** , 1087 – 1091.

Molenberghs, G., and Verbeke, G. (2001). A review on linear mixed models for longitudinal data, possibly subject to dropout. *Statistical Modelling*, **1(4)** ,235 – 269.

Park, J. Y., Wu, C., Basu, S., McGue, M., and Pan, W. (2018). Adaptive SNP-set association testing in generalized linear mixed models with application to family studies. *Behavior genetics*, **48**, 55 – 66.

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton,D., Pham, T. M et. al. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, 157 – 166.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461 – 464.

Sikorska, K., Rivadeneira, F., Groenen, P. J., Hofman, , et. al. (2013). Fast linear mixed model computations for genome-wide association studies with longitudinal data. *Statistics in medicine*, **32(1)**, 165 – 180.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, **30(4)**, 377 – 399.

Spatially adaptive Bayesian P-splines

Paul Bach[1], Nadja Klein[1]

[1] Research Center Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technische Universität Dortmund), Dortmund, Germany

E-mail for correspondence: paul.bach@tu-dortmund.de

Abstract: We propose a flexible Bayesian approach for spatially adaptive smoothing using tensor product P-splines. We define a spatially adaptive smoothness prior and explain how posterior sampling can be implemented efficiently in Stan. We conduct a simulation study demonstrating the advantages over anisotropic Bayesian smoothing and consider an illustration for German precipitation data.

Keywords: Bayesian smoothing; Precipitation data; Spatially adaptive splines.

1 Spatially adaptive Bayesian smoothing

Consider the two-dimensional nonparametric regression model

$$y_i = f(x_{i1}, x_{i2}) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n,$$

where $f : [0,1]^2 \longrightarrow \mathbb{R}$ is an unknown function to be estimated. We assume that the smoothness of f varies across the domain $[0,1]^2$ so that spatially adaptive smoothing is appropriate. We expand f in terms of tensor product B-splines, which allows us to write the model in the form

$$y = B\beta + \epsilon = (B_1 \otimes_r B_2)\,\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I_n),$$

where B is the $n \times d$ tensor product B-spline design matrix (the row-wise Kronecker product $\otimes_r$ of the $n \times d_j$ marginal B-spline design matrices $B_j,\ j = 1, 2$). Isotropic or anisotropic smoothness priors are often used for Bayesian smoothing (see, e.g., Bach and Klein, 2022), but these priors do not allow us to capture locally varying smoothness adequately. To address

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

this shortcoming, we follow Rodríguez-Álvarez et al. (2022) and consider the spatially adaptive penalty matrix

$$K(\lambda_1, \lambda_2) = D_1^{\mathrm{T}} \mathrm{diag}(\lambda_1) D_1 + D_2^{\mathrm{T}} \mathrm{diag}(\lambda_2) D_2,$$

where $D_1 = \Delta_1 \otimes I_{d_2}$ and $D_2 = I_{d_1} \otimes \Delta_2$. The $\Delta_j,\ j = 1, 2$ are second order difference matrices of sizes $(d_j - 2) \times d_j$ and λ_1, λ_2 are two vectors of positive smoothing parameters of lengths $(d_1 - 2)d_2$ and $d_1(d_2 - 2)$, respectively. To reduce the overall number of smoothing parameters, we follow Rodríguez-Álvarez et al. (2022) and set

$$\lambda_1 = X_1 \xi_1 \quad \text{and} \quad \lambda_2 = X_2 \xi_2,$$

where X_1 and X_2 are two tensor product B-spline design matrices with respect to pseudo-data on regular grids and ξ_1, ξ_2 are new vectors of positive smoothing parameters of shorter lengths p_1 and p_2. With this, we can write the spatially adaptive penalty matrix in the form

$$K(\xi) = D_1^{\mathrm{T}} \mathrm{diag}(X_1 \xi_1) D_1 + D_2^{\mathrm{T}} \mathrm{diag}(X_2 \xi_2) D_2.$$

In the present Bayesian setting we endow the tensor product B-spline coefficients $\beta \in \mathbb{R}^d$ with the corresponding partially improper Gaussian prior

$$p(\beta \mid \xi) \propto Det(K(\xi))^{1/2}\ \exp\left(-\beta^{\mathrm{T}} K(\xi) \beta / 2\right),\ \beta \in \mathbb{R}^d. \tag{1}$$

Thereby, Det is the generalized determinant, which is defined as the product of nonzero eigenvalues. To complete the prior specification, we use the Jeffreys prior for σ^2 and exponential priors $\xi_j \overset{iid}{\sim} Exp(\theta)$ for the smoothing parameters $\xi_j,\ j = 1, \ldots, p$, with $p = p_1 + p_2$.

2 How to implement the model in Stan

We use Stan (Carpenter et al., 2017) for efficient posterior sampling. To address the partial impropriety of (1), we make use of the Stan command `target +=`. To increase MCMC sampling efficiency, we exploit the equality

$$Det(K(\xi)) = det(K(\xi) + H_0),\ \xi \in (0, \infty)^p, \tag{2}$$

where det is the usual determinant and H_0 is the unique orthogonal projector onto the nullspace of $K(\xi)$. Equality (2) is easy to prove and allows us to compute the generalized determinant $Det(K(\xi))$ using a Cholesky decomposition instead of a spectral decomposition, which is more efficient.

3 Simulation study

In this section we compare the performance of spatially adaptive Bayesian P-splines and anisotropic Bayesian P-splines (see, e.g., Bach and Klein,

2022) in terms of root mean squared error (RMSE). Similar to Scenario II in Rodríguez-Álvarez et al. (2022) we consider the test function

$$f(x_1, x_2) = \exp\left(-50\ \{(x_1 - 1/2)^2 + (x_2 - 1/2)^2\}\right).$$

The design points (x_{i1}, x_{i2}), $i = 1, \ldots, n$, with $n = 500$ are iid uniform on the domain $[0, 1]^2$. The residual variance σ^2 is equal to $1/4$. Figure 1 shows the RMSE across 100 replications.

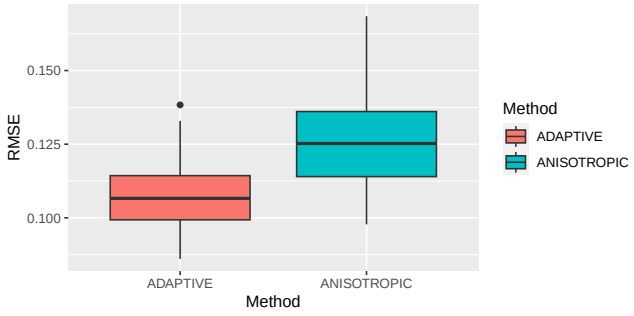


FIGURE 1. Root mean squared error across 100 replications for spatially adaptive Bayesian P-splines (left) and anisotropic Bayesian P-splines (right).

4 Application: Precipitation in Germany

In this section we use spatially adaptive Bayesian P-splines to analyze the spatial distribution of precipitation in Germany. We analyze publicly available data for the year 2020 retrieved from the German Meteorological Service, the DWD (cdc.dwd.de/portal). Spatially adaptive modelling is appropriate because in general the amount of precipitation can be expected to be similar for nearby locations. However, because of geographical features such as mountain ranges there may also be abrupt changes in the amount of precipitation. Figure 2 shows the estimated precipitation surface.

5 Discussion

We introduce spatially adaptive Bayesian P-splines as fully Bayesian counterpart of the approach of Rodríguez-Álvarez et al. (2022), which is based on restricted maximum likelihood (REML). We demonstrate the advantages over anisotropic Bayesian smoothing and apply the approach to analyze precipitation data from Germany. An interesting point for discussion is our use of Stan for posterior sampling: While this increases the flexibility of the approach, it limits its scalability. On the one hand, it is very straightforward to carry the presented approach over to non-Gaussian regression

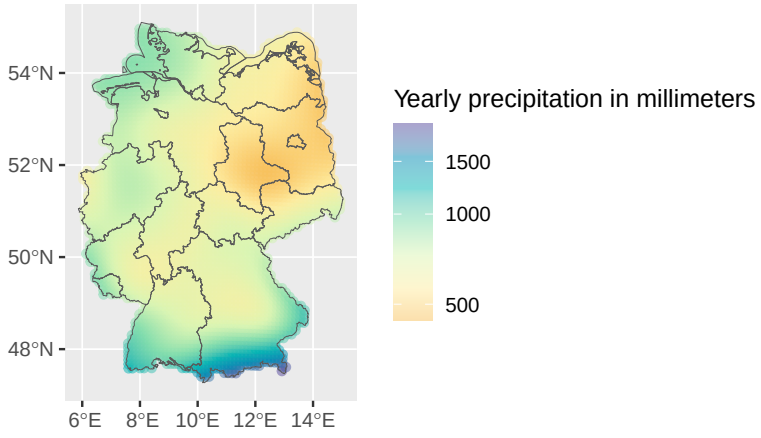


FIGURE 2. Yearly amount of precipitation in Germany. The amount of precipitation is lowest in the northeast and highest near the Alps. This is in line with the literature (see, e.g., Jung and Schindler, 2019).

models such as robust regression models, binary or count regression models. To this end, one only needs to change the likelihood in Stan's model block. On the other hand, however, the scalability is limited and the sampling time increases significantly when using large sample sizes n or a large number of tensor product B-splines d. Similarly, the sampling time increases significantly when moving from a two-dimensional setting to a three-dimensional setting (e.g. for spatio-temporal data). Interesting directions for our future research are:

- The derivation and implementation of a faster MCMC sampler that is tailored to spatially adaptive Bayesian P-splines.
- The investigation of approximate Bayesian methods such as (integrated nested) Laplace approximations or variational inference as an alternative to MCMC.

First steps in the latter direction using the variational inference engine implemented in Stan seem promising.

Acknowledgments: The authors gratefully acknowledge support by the German research foundation (DFG) through the Emmy Noether grant KL 3037/1-1. This work has been partly supported by the Research Center Trustworthy Data Science and Security, one of the Research Alliance centers within the University Alliance Ruhr (UA Ruhr).

References

Bach, P. and Klein, N. (2022). Anisotropic multidimensional smoothing using Bayesian tensor product P-splines. Preprint. *arXiv:2211.16218*.

Carpenter, B., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**, 1 – 32.

Jung, C. and Schindler, D. (2019). Precipitation atlas for Germany (GePrA). *Athmosphere* **10**, 737.

Rodríguez-Álvarez, M.X., et al. (2022). Multidimensional adaptive P-splines with application to neurons' activity studies. *Biometrics*.

A weighted curve clustering approach for analyzing pass rush routes in American football

Robert Bajons[1], Kurt Hornik[1]

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

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

Abstract: We present a weighted K-means approach for clustering weighted curves, i.e. curves which may be assigned weights at each observation of the curve. The methodology is applied to routes of defending players in American football, where the aim is to automatically detect effective pass rushing routes from specific players or teams. Preliminary results demonstrate that the methodology used is able to cluster pass rushing routes effectively and much better than a classical (unweighted) K-means approach.

Keywords: Sports Analytics; Curve Clustering; Weighted K-Means.

1 Introduction

Recently, as new data formats such as event stream data, play by play data and specifically tracking data have been developed, the problem of clustering curves has found attention in sports modelling. In (team) sports, such as American football or European Football (Soccer), players naturally move on the pitch in specific trajectories. Since usually the paths of players on the pitch follow certain criteria defined by the players position as well as the tactics of the team, interesting analyses can be derived from studying common pattern in these movements. Miller and Bornn (2017) for example studied player trajectories in Basketball by clustering possession into groups of similar offensive structure. Chu et. al. (2020) used similar techniques to cluster routes of wide receivers in football and derive a database of predefined routes.
The main motivation of this work is distinct from previous approaches and is based on the idea that in football and soccer, routes or possessions (or in

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

general trajectories of players/events) can be assigned weights. Instead of considering routes or possessions only as observed (x, y)-pairs of trajectories on the field, they can rather be viewed as a sequence of triplets (w, x, y). An example from football are pass rushing routes from defensive players. It is possible to assign to each observation of the trajectory a pressure probability, which would serve as a weight for each (x, y)-pair. Then, it makes sense to find structure in the weighted trajectories instead of the original curves.

In this paper, we present a weighted K-means approach to cluster the (weighted) trajectories of pass rushing defenders in American football, i.e. defenders, whose aim is to attack the quarterback and hinder him from throwing a pass. We consider a dataset provided by NextGenStats via the NFL Big Data Bowl 2023 competition on Kaggle, which contains tracking data of every player on every passing play from the first 8 weeks of the 2021 season of the NFL. For each player and play the data contains (x, y)-coordinates of the trajectories until some event (usually when the ball is thrown). We first build a model which assigns probabilities of quarterback pressure at (roughly) every timepoint in order to obtain weighted trajectories for each defensive player.

2 Methodology

Formally, we consider data $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, where each $\boldsymbol{y_i}$ is an $m_i \times 3$ dimensional matrix of weighted trajectories, comprising of a vector of x-coordinates, y-coordinates and a vector of weights w. Since m_i is not fixed but varies for each data point due to fact that some plays take longer than others, it is necessary to unitize the data in order to use a K-means approach. We thus approximate each trajectory by a Bézier curve evaluated at a fixed number M of points. Details about this adjustment are omitted to comply with the predefined page limit of the short paper.

We proceed by briefly describing the clustering methodology. The classical K-means approach tries to find an optimal partition of the n observations $(x_i, \ldots, x_n)$ into K cluster g_i, $i = 1, \ldots, K$, such that the within cluster sum of squares

$$S = \sum_{k=1}^{K} \sum_{i:g_i=k} (x_i - p_k)^2 \tag{1}$$

is minimized. The prototype p_k is given as the cluster mean , $p_k = \frac{1}{N_k} \sum_{i:g_i=k} x_i$. In the case of weighted observations it makes sense to take the weights into account by adjusting equation (1) such that instead the aim is to minimize

$$\sum_{k=1}^{K} \sum_{i:g_i=k} v_i (x_i - p_k)^2. \tag{2}$$

The resulting optimal prototypes for a cluster are found as the weighted averages $p_k = \frac{\sum_{i:g_i=k} v_i x_i}{\sum_{i:g_i=k} v_i}$.
To adapt the algorithm to the data at hand, each observation $\boldsymbol{y}_i \in \mathbb{R}^{M\times 3}$ is transformed such that an $\tilde{M} = 2M$-dimensional vector $\boldsymbol{z}_i = (x_{1,i}, \ldots, x_{M,i}, y_{1,i}, \ldots, y_{M,i})$ and a corresponding weights vector $\boldsymbol{w}_i = (w_{1,i}, \ldots, w_{M,i}, w_{1,i}, \ldots, w_{M,i})$ of the same dimension is obtained. The problem is then to find clusters and prototypes such that the following expression is minimized:

$$\min_{(p_{jk}),(g_i)} \sum_{k=1}^{K} \sum_{i:g_i=k} \sum_{j=1}^{\tilde{M}} w_{i,j}(z_{i,j} - p_{k,j})^2. \tag{3}$$

In analogy to classical K-means algorithms we implemented an iterative refinement procedure which is initialized by an appropriate starting assignment of clusters and then alternates between finding the optimal prototypes for given cluster assignments and finding the optimal cluster assignment given prototypes, until convergence is achieved, i.e. the change in the function to optimize is below some tolerance. The optimal prototypes for given cluster assignment are given by

$$p_{k,j} = \frac{\sum_{i:g_i=k} w_{i,j} z_{i,j}}{\sum_{i:g_i=k} w_{i,j}}, \tag{4}$$

whereas the optimal cluster assignment given prototypes is found by minimizing

$$\sum_{j=1}^{M} w_{i,j}(z_{i,j} - p_{k,j})^2, \tag{5}$$

over k.

3 Results

The left frame of Figure 1 shows the result of the weighted K-means algorithm described in the previous section for the defensive football players when using 12 clusters. Note that the X-axis is scaled, such that plays are from left to right (from the viewpoint of the offensive team) and the value 0 indicates the line of scrimmage for the play. The aim is to derive clusters of similar routes, where the weights, representing the probability of putting pressure on the quarterback (QB), are taken into account. In essence the idea is to distinguish between routes with high pressure outcome and low pressure outcome. If the algorithm is able to do so automatically it is possible to identify strengths of players as well as teams. It can be observed

nicely that there are outside as well as inside route clusters, and it is possible to identify effective routes and less effective routes (as given by the pressure weights) for both categories. The grey cluster in the left part of Figure 1 (Cluster 8) seems odd at first sight but upon examination it is clear that it comprises of coverage routes. This is particularly nice, as when analyzing pressure on the quarterback, such routes are not of importance and/or interest. In theory one could simply omit them for the clustering exercise, however from the data it is not clear how to distinguish them. Although there are role labels for defenders ("Pass Rush" and "Coverage"), often coverage players also attack the quarterback. To emphasize the importance of using a weighted K-means approach as opposed to a classical K-means algorithm, the results from the latter approach are shown in the right frame of Figure 1. From a pure route specific point of view the clusters seem reasonable, we observe pass rush clusters and coverage clusters. However, there are two main issues. First, when analyzing pressure on the QB, coverage routes are not of much interest as is also evident from the weights, which are (almost) 0 for these route clusters (clusters 2,6,7,10,12 in the right frame of Figure 1). Second, judging from the weights of the pass rush route clusters (clusters 1,3,9,11), we are not able distinguish between more threatening and less threatening routes, so the clustering is useless when trying to identify which teams or players are effective at which position.

References

Chu, D., Reyers, M., Thomson, J., and Wu, L. (2020). Route identification in the National Football League: An application of model-based curve clustering using the EM algorithm. *Journal of Quantitative Analysis in Sports,*. 16(2), 121-132.

Miller, A.C., and Bornn, L. (2017). Possession sketches : Mapping NBA strategies In: *Proceedings of the 2017 MIT Sloan Sports Analytics Conference.*

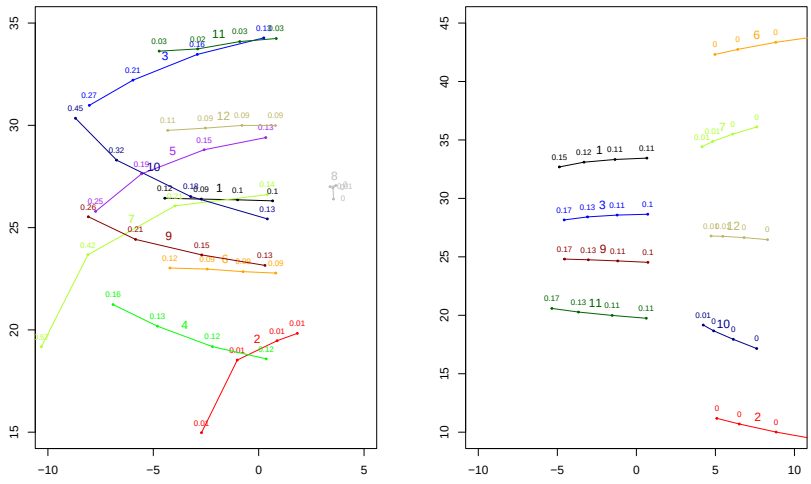


FIGURE 1. Left: 12 cluster as obtained from the weighted K-means algorithm with average weights at each observation point of the trajectory. Right: 12 clusters as obtained from the usual (non-weighted) K-means algorithm (only 9 clusters shown).

Playful introduction to data competencies for economic students

Julia Berginski[1], Alexander Silbersdorff[1]

[1] University of Göttingen, Germany

E-mail for correspondence: julia.berginski@uni-goettingen.de

Abstract: Following the aim to give students of business science a tangible introduction to data competencies, a game for the use in a statistics lecture was developed. During this game, students are tasked with collecting and using data in order to maximize the profit of their hypothetical company and win the game. The game uses a hands-on approach with paper and few work materials. It further uses R-code to compute additional data for the groups. The game will be used in an analysis of motivation and stress of students regarding statistics.

Keywords: Statistics; Teaching; Data Competencies; Introduction.

1 Introduction

To introduce economic students at the University of Göttingen to data competencies, a game dealing with data and data collection was developed. Data competency is an important skill to have. From an economic standpoint, data literacy improves the quality of decisions in firms (Ghasemaghaei et al., 2018) as well as the quality of manufacturing planning (Chae et al., 2014). Moreover, data collection and its usage can improve the quality of health care if done well (Bose, 2003). These are brief examples of the practical relevance of data competencies being taught in university.

2 Outlining the Game

The given goal within the developed game is profit maximization of hypothetical companies. In contrast to other business simulation games, much of the given information is uncertain while the mechanics of the game are kept very simple. The game consists of three rounds with six phases each.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

The phases are: (1) information, (2) contract, (3) buying, (4) production, (5) selling and (6) destruction phase. An explanation of the contents of these phases can be found in Figure 1. The firms - made up of groups of five to seven students - can produce and sell three different products: a rectangle, a triangle and a circle. Each shape is made out of paper by the means provided to each firm (see below). The shapes are shown in Figure 2. Only very little variation to the given measurements is allowed.

(1) Information Phase: The groups get information about expected paper cost, selling prices, price development and their account balance.
(2) Contract Phase: Each group has to make an offer over the number of products and the respective selling price via a contract form. After consideration of all offers, the market decides which offers will be accepted. An exemplary upper limit for the accepted price is the 80% quantile. Additional policies are: the market does not accept more than offered and if the firm delivers less than offered, the price is reduced.
(3) Buying Phase: The paper in A4 format is bought at the given market price. An order form has to be handed to the market and the resources need to be collected from the market within the give time.
(4) Production Phase: The groups produce what they have offered with the bought paper and the given material. Ideally, the groups have members who count and quality check their products.
(5) Selling Phase: The seller brings the products to the market, who decides on the fulfillment of the contract after a quality check.
(6) Destruction Phase: In this variant, residual and unused paper is collected. Therefore, the groups start the next round with zero stock.
Completion After the last destruction phase, the winner is announced. Subsequently, the students are asked to reflect the game and answer some questions regarding data collection, development and usage.

FIGURE 1. Description of the game phases.

Next to the craft required to produce these products (which many of the students focus on fervently), one catch of the game lies in limited working materials given to each group: one pencil, one ruler, one triangle ruler, one pair of scissors, one compass and one glue stick. Therefore, not every member of the group can produce at the same time. To aid this intended division of labour, the groups are given exemplary recommendations for roles within a group: production, buyer, seller, accounting, management and consulting. Theoretically ample spare capacity is left for (statistical) analysis to aid the optimisation of the firms performance. The groups decide in which way they may want to distribute tasks among their members. The lecturer (with additional help) embodies the role of the market.
The time for each phase is short (three to five minutes), therefore collaboration within the group is important. The students need to figure out what

to produce and how. Some reflected students may realise from the outset that to this end an explicit analysis of the data they have at their disposal and/or the generation of data may explicitly aid their performance. However, our experience is that many teams don't reflect upon the potential use that data can have, which ultimately yields a plastic experience of the use of statistics in the follow-up reflection.

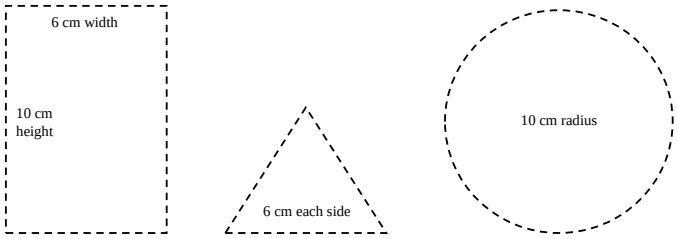


FIGURE 2. Types and dimension of the products.

Still, during and after the game the students should have reflected on the significance of the mean and prognosis because of the given information in the information phase. They also should have thought about different types of data. Besides metric data like buying and selling prices, they also have to use ordinal data, for example who can produce the most (most accurate, fastest, ...) products. In a later lecture, the game can also be used to calculate an example of confidence intervals. The calculations regarding the account balances and the decision over which selling prices are accepted are made with `R` (R Core Team, 2023). Here lies the possibility to show the students the specific code or even task them with writing their own code.

3 Preliminary evaluation

In a preliminary evaluation of the usefulness of the game in the statistics curriculum, we used a sample size of 110 students, who were offered to play the game during a lecture. For most of the participation the statistics module is mandatory. Accordingly, we consider the sample to be roughly representative of the population of economics students at our university, although the usual set of potential selection biases cannot be ruled out. The game was played at the beginning of the semester, and most of the students had not yet heard a lot of the teaching material of the course. A fixed number of six groups was picked at the outset to limit the administrative requirements. That led to groups consisting of 15 to 20 members. Each group was given the material, contract and order forms and an inventory list. During the game, it became clear that not all groups wrote down notes about the rules, so there were uncertainties at different stages.
Not all groups assigned roles or tasks to their members, which expressed itself during the phases. Some groups did not collect necessary data during

the information phase. As a consequence, they were unsure how to price their products and how much paper to buy. Overall, this was not bad, because it gave them a good opportunity to reflect which data they should have collected. The 80% quantile was used for the accepted selling prices. For the groups, this implied that they also should gather data about the other groups, e.g. their offered prices, in order to possibly adjust their own prices in the next round according to the other groups and the data from the information phase. Furthermore, some groups had trouble being on time during the contract, buying and selling phase because they were unsure who would fill out the forms and contact the market. In addition, most of the groups did not have a (sufficient) quality check of the products and the market had to reject a lot of products, which diminished their profit.
In the completion, some students reported that the first round was more like a trial round for them. Therefore, they were better able to use the given information about the prices and their collected data (e.g. regarding the possible number of producible products) during the following rounds.

4 Planned analysis

For the next conduction of the game, some adjustments will be made for further analysis. Besides adding shortened rules to the given material, the group size is going to be reduced ($5-7$ members) so that roles and tasks can be better distributed. The given number of groups is chosen accordingly.
Data related to the experienced emotions of students during the statistics course and their motivation regarding statistics is going to be collected. In particular, their attitude towards statistics before and after the game as a pre- and posttest shall be examined. For this, students will be asked to fill in a questionnaire, e.g. the Survey of Attitudes Toward Statistics (SATS-28 or SATS-36), which was developed by Schau et al. (1995). The components of these questionnaires consist e.g. of *Affect*, which uses six items to measure the feelings of students towards statistics. Other components are *Value* and *Cognitive Competence*, which measure the perceived usefulness and relevance of statistics in students lifes as well as their own assessment of their ability to learn statistics. Within the SATS-36 the component *Interest* as a measure for students' interest in statistics is added (Schau, 2003). Furthermore, we are interested in the mental health impact of statistics on students. On the one hand, we are interested in the students' perception of stress. To obtain data on this, we plan on using questions like those used in an empirical report about stress perception among students in Germany from Herbst et al. (2016). In those, the students are asked to indicate how they perceive stress in different situations, e.g the level of demands of their university courses as well as group work and the way material is being taught (Herbst et al., 2016). On the other hand, data on the relationship with fellow students will be collected. Therefore, questions

regarding social support are necessary. We plan to use a questionnaire adapted to our conditions (concentrating on fellow students) like the short form of the Social Support Questionnaire (F-SozU) (Fydrich et al., 2009). Our proposed model for analysis is a cumulative ordinal regression model (Fahrmeir et al., 2013). Our response variable $Y_i \in \{1, ..., c+1\}$ consists of the attitude of students towards statistics as well as their perceived stress and social support. It is categorical and measured on an ordinal 5 or 7-point Likert scale. Regarding the attitude towards statistics, 1 equals "Strongly Disagree" and 7 equals "Strongly Agree" with 4 equaling "Neither Disagree or Agree" (Schau, 2003). Our covariates x_i are going to be the age, gender, math grade in university course and in school and the number of prior statistics courses of the students. Additional covariates on a Likert scale are going to be the students own rating of their mathematical abilities (Lavidas et al., 2020), their enjoyment of the game and their feeling of loneliness/social inclusion in the student body of the respective semester. Furthermore, the model contains covariates regarding the degree of participation in the game and the comprehension of the relevance of data.

For the analysis we plan to use a cumulative ordinal regression model:

$$Y_i = r \iff \theta_{r-1} < u_i \leq \theta_r, \quad r = 1, ..., c+1, \tag{1}$$

where Y_i is the ordinal response variable, that is linked to a latent variable u_i via the ordered thresholds $-\infty = \theta_0 < \theta_1 < ... < \theta_{c+1} = \infty$.
The covariates are modelled as linear predictors as follows:

$$u_i = -\boldsymbol{x_i}\boldsymbol{\beta} + \varepsilon_i. \tag{2}$$

Using this set up, we plan to estimate the probability of a student being in or below a certain category in terms of attitude towards statistics, perceived stress and social support (Tutz, 2011). For this, our Y_i will be students' affect, cognitive competence, interest and value regarding statistics as well as their risk of loneliness and stress associated with group work and data. Using a 5 and 7-point scale, we have four and six thresholds respectively.

5 Conclusion and outlook

For further didactic purposes the game can be adapted in different ways. For example, an additional penalty for insufficient contract fulfillment can be added to further risk awareness regarding ones prediction. All in all, the game offers an easy introduction to data competencies for economic students. Within this simple game they can experience the need for different types of data. They also experience first hand that not all available data is always necessary and that some additional data (to the freely given information) should be gathered. They also come into contact with different calculations within `R`. The goal of our planned analysis is to see if a

hands-on approach such as the presented game of an introduction to data competencies is beneficial to the motivation of students regarding statistics.

References

Bose, R. (2003). Knowledge management-enabled health care management systems: capabilities, infrastructure, and decision-support. *Expert systems with Applications*, **24(1)**, 59 – 71.

Chae, B.K., Yang, C., Olson, D., and Sheu, C. (2014). The impact of advanced analytics and data accuracy on operational performance: A contingent resource based theory (RBT) perspective. *Decision support systems*, **59**, 119-126.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). Categorical Regression Models *Regression: Models, Methods and Applications*. Berlin Heidelberg: Springer-Verlag, 325 – 347.

Fydrich, T., Somer, G., Tydecks, S., and Brähler, E. (2009). Fragebogen zur sozialen Unterstützung (F-SozU): Normierung der Kurzform (K-14). *Zeitschrift für Medizinische Psychologie*, **18(1)**, 43 – 48.

Ghasemaghaei, M., Ebrahimi, S., and Hassanein, K. (2018). Data analytics competency for improving firm decision making performance. *The Journal of Strategic Information Systems*, **27(1)**, 101 – 113.

Herbst, U., Voeth, M., Eidhoff, A.T., Müller, M., and Stief, S. (2016). *Stu-dierendenstress in Deutschland – eine empirische Untersuchung*. Berlin: AOK-Bundesverband.

Lavidas, K., Barkatsas, T., Manesis, D., and Gialamas, V. (2020). A structural equation model investigating the impact of tertiary students' attitudes toward statistics, perceived competence at mathematics and engagement on statistics performance. *Statistics Education Research Journal*, **19(2)**, 27 – 41.

R Core Team (2023). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*

Schau, C., Stevens, J., Dauphinee, T.L., and Del Vecchio, A. (1995). The development and validation of the Survey of Attitudes Toward Statistics. *Educational and Psychological Measurement*, **55**, 868 – 875.

Schau, C. (2003). Students' attitudes: The "other" important outcome in statistics education. In: *Proceedings of the joint statistical meetings*, San Francisco, California, 3673 – 3681.

Tutz, G. (2011). *Regression for Categorical Data*. Cambridge: Cambridge University Press, 241 – 268.

Accounting for clustering in automated variable selection using hospital data: A comparison of different LASSO approaches

Stella Bollmann[1,2], Andreas Groll[3], Michael M. Havranek[1]

[1] Competence Center for Health Data Science, Faculty of Health Sciences and Medicine, University Lucerne, Switzerland
[2] Department of Educational Sciences, University Zurich, Switzerland
[3] Department of Statistics, TU Dortmund University, Germany

E-mail for correspondence: stella.bollmann@uzh.ch

Abstract: Automated feature selection methods such as the Least Absolute Shrinkage and Selection Operator (LASSO) have recently gained importance in the prediction of quality-related outcomes. However, the methods that have been used so far, usually, do not account for the fact that patient data are typically nested within hospitals. Therefore, we aim to demonstrate how to account for the multilevel structure of hospital data with LASSO and compare the results of this procedure with a LASSO variant that ignores the multilevel structure of the data. We find that inserting hospitals leads, at least partly. to better predictions and in some instances, the variable importances differ greatly between the methods. In summary, we show that it is possible to take the multilevel structure of data into account in automated predictor selection and that this leads, to better predictive performance. From the perspective of variable importance, including the multilevel structure is crucial to selecting predictors in an unbiased way under consideration of the structural differences between hospitals.

Keywords: LASSO; feature selection; mulitlevel model; mixed model; hospital quality indicators.

1 Background

Performance metrics such as *Duration Stay* or indicators of the quality of care such as *Mortality Rates* are widely used to assess healthcare providers' performance. A broad branch of research has identified predictors that explain differences in such measures both at the patient as well as the provider level.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

Recent improvements in data availability have increased the number of healthcare-related variables as possible predictors of these performance metrics (e.g., with regard to quality of care, see Schwartz et al., 2017). Consequently, the use of automated feature selection tools like the *least absolute shrinkage and selection operator* (LASSO) have become more and more popular to facilitate variable selection, and studies have demonstrated its superiority in prediction accuracy (e.g. Harris et al., 2019). In general, there has been great progress in collecting large amounts of data to identify predictors for quality of care using automated variable selection methods. However, all these studies face a methodological challenge that remains unresolved: The nested structure of the hospital data.
All in all, there is relatively great unanimity regarding the general importance of accounting for the multilevel structure of hospital data (see e.g. Dimick et al., 2012) and it has been shown that estimated relationships change dramatically depending on whether hospitals are included as random effects or not (Austin and Alte, 2003; Hofstede et al., 2018). Unfortunately, there is a lack of research regarding the integration of the multilevel structure of data sets in the automated feature selection process.
To address this research gap, our goals were, on the one hand, to examine whether taking hospital clustering into account provides better predictive performance, and on the other hand, to see whether it leads to a different set of selected variables. Therefore, we investigated a LASSO variant that incorporates the estimation of random effects, i.e., a LASSO for general linear mixed models (GLMMs; Groll and Tutz, 2014).

2 LASSO for clustered data

The idea of LASSO is to impose a penalty term on the size of the coefficients that addresses their absolute values:

$$J_{lasso}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j| \,.$$

This penalty term comes with the desirable feature of enabling variable selection, as the coefficient estimates can be shrunk down to exactly zero due to the absolute value function. However, the corresponding optimization becomes more cumbersome and the LASSO estimator does not exist in closed form. Instead, numerical methods must be used. To implement the classic LASSO, we use the R package `glmnet` (Friedman et al., 2010).
In the case of clustered data, however, it is also necessary to account for potential cluster-specific and unobserved heterogeneity during the variable selection process. One option is to incorporate random intercepts for hospitals into the regularized regression model within the GLMM framework. We use the implementation in the `glmmLasso` R package (Groll, 2022).

3 Application to example data

We used a national health administration data set provided by the Swiss Federal Statistical Office. It contains all inpatient cases treated in Swiss hospitals in 2019 with a main diagnosis of chronic obstructive pulmonary disease (COPD, n = 12,404) with 32 potential predictors including clinically relevant variables and demographic information. We predicted the continuous outcome *Duration Stay* that indicates how many days a patient was hospitalized. Additionally, we predicted the binary outcome *Mortality* that indicates whether a patient died during the hospitalization. In doing so, we aimed to examine how well the different LASSO variants can deal with both continuous and binary data.

We used a 20-fold sub-sampling procedure to evaluate the predictive performance of the different LASSO methods. Since some of our models also contained random effect estimates for each hospital, which can be used for the calculation of the deviance, we decided to include the requirement that each hospital from the test data was also represented in the training data with at least one observation. Moreover, note that the outcome *Duration Stay* was highly positively skewed and therefore needed to be log-transformed before the analyses.

To assess predictive performances, for the outcome *Duration Stay*, we used the mean squared error (MSE) on the test data. For the binary outcome *Mortality*, we used the area under the Receiver Operator Characteristic (ROC) curve (AUC, Hanley & McNeil, 1982). Both measures were averaged over the 20 sub-samples.

Additionally, for the outcome *Duration Stay*, we examined how variable importance and thereby variable selection changes when hospitals are included. To this end, we registered the five variables from all 20 data sets with the largest absolute coefficients in the model selected via cross-validation (see Table 1).

As previously mentioned, we generally distinguished between standard, fixed-effects-only LASSO models via `glmnet` (*No hosps*) and the LASSO variant including random effects via `glmmLasso` (*Hosps random*). In addition, we included `glmnet` with fixed-effects for the hospitals (*Hosps fixed*) as an intermediate solution for comparison.

Additionally, we used different procedures for the tuning of λ: (1) K-fold cross-validation (CV) and (2) information criteria-based approaches using AIC or BIC.

4 Results

The MSEs for the outcome *Duration Stay* are summarized in Figure 1. It can be seen that including the hospitals, whether as fixed or as random effects, leads to considerably better results (i.e. lower MSEs). However, there

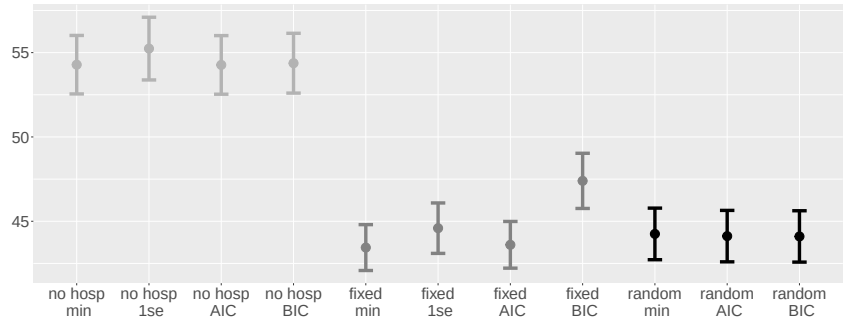


FIGURE 1. MSE (means and standard error bars) for the outcome *Duration Stay* for different LASSO variants.

is little difference between the different variants of including the hospitals and the various optimization criteria within those variants, except that the MSE is larger for the *fixed BIC*.

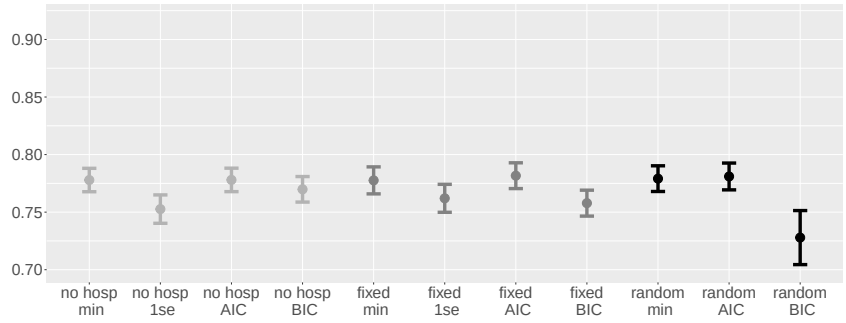


FIGURE 2. AUC (means and standard error bars) for the outcome *Mortality*.

Figure 2 shows mean AUC values and standard error bars for the results of the different LASSO variants on the binary outcome *Mortality*. According to generally accepted thresholds (see e.g. Hosmer et al., 2013), all variants achieve acceptable performance. Additionally, the AUC values show only small differences between the different variants and it does not seem to make a substantial difference whether hospitals are included or not. The only exception is *random BIC* that yields slightly lower performances.
A summary of the variable importance in Table 1 shows a substantial difference between the different variants depending on the inclusion of hospital effects. Particularly interesting is the variable *Planned admission*, that specifies whether patient admissions were planned and thus distinguishes elective admissions from emergency admissions. It has a large positive ef-

TABLE 1. Variable importances for the CV selected model concerning COPD data for the outcome *Duration Stay* for different LASSO variants.

Variable	**no hosps**	**hosps fixed**	**hosps random**
Adm from hospital	100% (.16)	100% (.11)	100% (.10)
Pre-MDC	100% (.22)	100% (.17)	100% (.19)
Part medical	100% (-.21)	100% (-.19)	100% (-.23)
MDRG E65	100% (.25)	100% (.13)	100% (.18)
elix 23	0% (–)	100% (.11)	100% (.11)
Planned admission	100% (.13)	0% (–)	0% (–)
elix 1	0% (–)	100% (.07)	0% (–)
elix 30	0% (–)	100% (.07)	0% (–)
elix 24	0% (–)	70% (.06)	0% (–)
Emergency	0% (–)	30% (-.06)	0% (–)

Note. Rates of being among the five most important variables (according to absolute values of regression coefficients) across the 20 sub-samples in percent; in brackets: mean estimated regression coefficient over all sub-samples.

fect on *Duration Stay* whenever hospitals are not specifically included in the variable selection.

5 Discussion

We found an improvement in predictive performance in models that include hospitals compared to those that do not include them, at least for the continous outcome. More importantly, however, we found that variable importance changes when hospitals are included. Some variables go from being among the top five for all 20 sub-samples to no longer being in the top five for any of the sub-samples when hospital effects are included. Such differences can, of course, lead to completely different interpretations of dependencies between predictor variables and the outcome. Another consequence of our results is that the comparison of selection results from different approaches (i.e., with or without the consideration of hospital effects) could be used to learn about dependencies within the data in a data-driven way. More specifically, it may support researchers to assess which variables are mainly associated with hospital characteristics and which are primarily related to patient characteristics.

Based on our findings, we recommend that the natural clustering of hospital data already should be considered when selecting predictor variables for prediction purposes or the risk-adjustment of quality indicators, in contrast to only considering it in the final modeling stage (once the predictors have been selected). This approach leads to predictions that are better or at least

as good as when not considering hospital effects. In addition, when hospital effects are not included, interpretations of the importance of predictor effects and, therefore, dependencies among variables can be distorted.

References

Austin, P.C., Alte, D.A. (2003). Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: should we be analyzing cardiovascular outcomes data differently? *American heart journal*, 145(1):27–35.

Dimick, J.B., Ghaferi, A.A., Osborne, N.H., Ko, C.Y., Hall, B.L. (2012). Reliability adjustment for reporting hospital outcomes with surgery. *Annals of surgery*, 255(4):703–707.

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

Groll, A. (2022). `glmmLasso:` Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation. R package version 1.6.2.

Groll, A., Tutz, G. (2014). Variable selection for generalized linear mixed models by L1-penalized estimation. *Statistics and Computing* 24(2):137 – 154.

Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1):29 – 36.

Harris, A.H., Kuo, A.C., Weng, Y., Trickey, A.W., Bowe, T., Giori, N.J. (2019). Can machine learning methods produce accurate and easy-to-use prediction models of 30-day complications and mortality after knee or hip arthroplasty? *Clinical orthopaedics and related research*, 477(2):452.

Hofstede, S., van Bodegom-Vos, L., Kringos, D.S., Steyerberg, E., Marang-van de Mheen, P.J. (2018). Mortality, readmission and length of stay have different relationships using hospital-level versus patient-level data: an example of the ecological fallacy affecting hospital performance indicators. *BMJ Quality & Safety*, 27(6):474 – 483.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Schwartz, J., Wang, Y., Qin, L., Schwamm, L.H., Fonarow, G.C., Cormier, N., Dorsey, K., McNamara, R.L., Suter, L.G., Krumholz, H.M., Bernheim, S.M. (2017). Incorporating stroke severity into hospital measures of 30-day mortality after ischemic stroke hospitalization. *Stroke*, 48(11):3101 – 3107.

An active deep learning method for high out-of-sample predictive performance in image classification

Ludwig Bothmann[1,2], Lisa Wimmer[1,2], Omid Charrakh[3], Tobias Weber[1,2], Hendrik Edelhoff[4], Wibke Peters[4,5], Hien Nguyen[4], Caryl Benjamin[6], Annette Menzel[6,7]

[1] Department of Statistics, LMU Munich, Germany
[2] Munich Center for Machine Learning (MCML), Germany
[3] Munich Center for Mathematical Philosophy, LMU Munich, Germany
[4] Wildlife Biology and Management Research Unit, Bavarian State Institute of Forestry (LWF), Freising, Germany
[5] Wildlife Biology and Management Unit, Technical University of Munich, Freising, Germany
[6] Ecoclimatology, TUM School of Life Sciences, Technical University of Munich, Freising, Germany
[7] TUM Institute for Advanced Study, Garching, Germany

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

Abstract: Human annotation of large image data sets is time-consuming, but annotated images are often necessary to tackle subsequent research questions. Pre-trained neural networks are often bound to a fixed set of classes, which is a major obstacle to using them for new class sets. For an applied use-case of wildlife camera trap images, we propose an image classification pipeline of object detection and image classification. We design an active learning component that allows us to re-train existing models using new data very efficiently – also for new class sets. We propose a tuning strategy for optimizing the hyperparameters of the pipeline using nested resampling. Empirical results show how tuning and active learning improve out-of-sample predictive performance.

Keywords: Deep Learning; Active Learning; Out-of-Sample Prediction; Nested Resampling; Image Data

---

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

1 Introduction

Wildlife ecologists all over the world use camera traps for estimating the population sizes of different animal species and for other research questions such as assessing animal interactions, animal-plant interactions, or animal-human interactions. These camera traps produce high amounts of image data which have to be classified with respect to animal species before tackling downstream tasks. Annotation of images by human experts is possible but very time-consuming, which is why this task shall be automated. While there exist many methods from the field of deep learning (DL) that tackle image classification and claim good predictive performance, oftentimes it is challenging to reproduce similarly convincing results in applied use cases. There are, among others, two main reasons that hinder applied researchers from making use of previously proposed DL methods: (A) there is a variety of hyperparameters that have a strong influence on the performance – selecting the best hyperparameters is a complex process which calls for careful resampling strategies and (B) there is no all-encompassing answer to the question, how many images of each class are needed to learn a good model. We propose a framework to solve this (A) by designing a resampling strategy that allows optimizing the hyperparameters with respect to out-of-sample predictive performance, (B) by developing an active learning pipeline to enable human-in-the-loop training to allow training with increasing amounts of images until a sufficient performance is achieved, and (C) by publishing an open-source software package in Python implementing this framework.

2 Data and Methods

To demonstrate our framework, we use image data from a study site in Bavaria, Germany, comprising 48,116 images of 8 classes (European hare, red deer, red fox, red squirrel, roe deer, wild boar, others, empty) at 37 camera stations.

Image Classification Pipeline: The core image classification pipeline consists of two main parts: (A) an object detection model which detects bounding boxes of animals in the images and (B) an image classification model which classifies the previously identified bounding boxes with respect to the above classes. As an *object detection model*, we use the *Megadetector* by Beery et al. (2019). This outputs the coordinates of the bounding boxes of animals, together with a confidence $c^{(i)} \in [0, 1]$ that the detected object in bounding box i is in fact an animal. Bounding boxes where the confidence exceeds a threshold α (a tunable hyperparameter) are passed on to the image classifier – images without such bounding boxes are considered empty. For the *image classifier*, we use transfer learning based on different pre-trained image classification networks. The choice of the network and the number of finetuned layers are tunable hyperparameters.

Hyperparameter Tuning: Tuning the hyperparameters of this two-stage classification pipeline demands a carefully designed resampling strategy to avoid any information leakage between training, validation, and test data. Figure 1 visualizes our version of nested resampling for hyperparameter tuning.

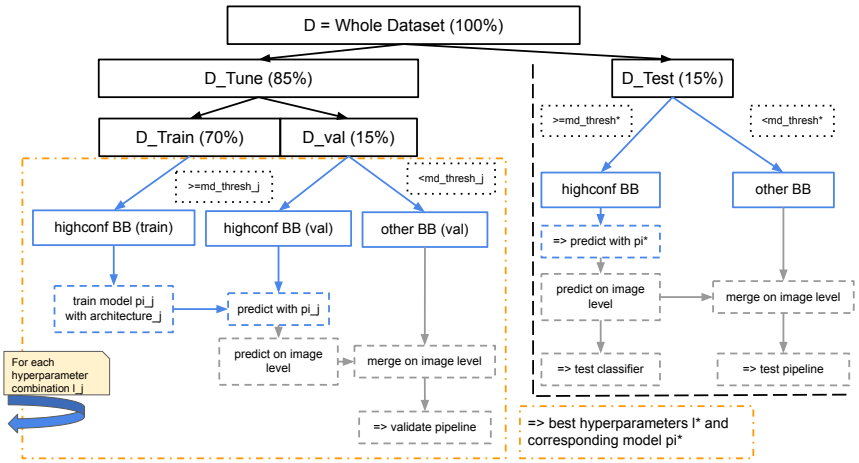


FIGURE 1. Resampling strategy. `md_thresh` is the Megadetector threshold α, `highconf BB` are bounding boxes with confidence above `md_thresh`.

Active Learning Pipeline: We propose an active learning (AL) pipeline in order to use the available data most efficiently as human-in-the-loop. The pipeline consists of the steps *(1) Initialization* (provide unlabeled data) *(2) AL loop* (while predictive performance does not exceed pre-specified threshold, do:) *(2a) Image selection* (select unlabeled images via an acquisition function such as softmax entropy) *(2b) Manual labeling* (expert labels images) *(2c) Model training* (tune, train, evaluate model) *(3) Prediction* (predict remaining unlabeled images with the latest model).

3 Results

As a glimpse of the results, we show how hyperparameter tuning improves the performance at the example of the confidence threshold α in Table 1. The most common choice (as, e.g., in Norouzzadeh et al., 2021) is $\alpha = 0.9$ – tuning ($\alpha = 0.253$) improves the false positive rate (FPR) by a factor of > 2.

Figure 2 shows how the predictive performance increases with using more images during active learning. Using a relative sample size of about 40%, we already reach 97% of the 8-class accuracy of the upper baseline.

TABLE 1. Performance for empty (0) vs. non-empty (1) images.

Confidence α	Accuracy	Precision	Recall	FPR	F1
0.253	**0.953**	0.964	**0.976**	**0.024**	**0.970**
0.5	0.949	0.973	0.963	0.037	0.968
0.9	0.942	**0.978**	0.948	0.052	0.963

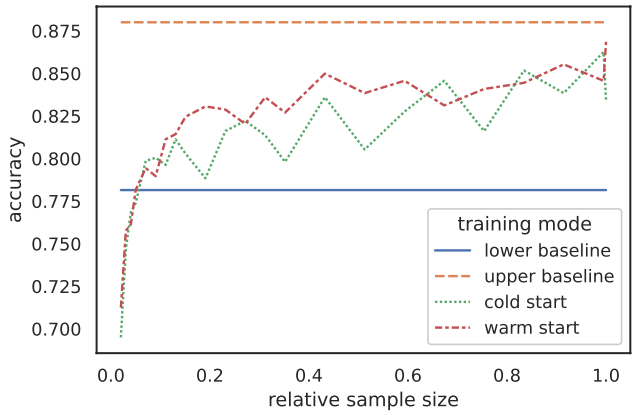


FIGURE 2. Comparison of 8-class accuracy for different amounts of training images used in active learning. Lower baseline: Training only on other camera stations. Upper baseline: Training with all available data. Warm start: Initialize weights as trained on other camera stations.

4 Conclusion

We proposed to improve image classification tasks by using active learning and sensible hyperparameter tuning and showed the benefit of both components with real-world data. An implementation of the framework and example code to reproduce the results will be made available upon acceptance.

References

Beery, S., Morris, D. and Yang, S. (2019). *Efficient Pipeline for Camera Trap Image Review*. Technical Report arXiv:1907.06772

Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N., and Clune, J. (2021) *A deep active learning system for species identification and counting in camera trap images*. Methods in Ecology and Evolution, **12(1)**, 150 – 161,

A smooth Laplace regression model

Kevin Burke[1]

[1] University of Limerick, Ireland

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

Abstract: We consider a novel smooth Laplace distribution, i.e., a version of the distribution but where the density function is differentiable. This is achieved using a differentiable approximation to the absolute value function, and facilitates Newton-type estimation in the robust regression context.

Keywords: differentiable approximation; Laplace distribution; robust regression

1 Introduction

Consider the linear regression model

$$y_i = x_i^{\mathrm{T}}\beta + \sigma\varepsilon_i,$$

which is fundamental to statistical modelling. Here, y_i is the response variable for the ith individual, x_i is the vector of covariates with coefficients β, σ is a dispersion parameter, and ε_i a random error. Assuming a symmetric error distribution, this is a model for the mean of y_i, i.e., $\mu = E(Y_i) = x_i^T\beta$. Classically, estimation may proceed using least *squares*,

$$\min_{\beta} \sum_i (y_i - x_i^{\mathrm{T}}\beta)^2,$$

equivalent to assuming a Gaussian error density, $f(\varepsilon) = \exp(-\varepsilon^2/2)/\sqrt{2\pi}$, or least *absolute deviations*

$$\min_{\beta} \sum_i |y_i - x_i^{\mathrm{T}}\beta|,$$

equivalent to assuming a Laplace error density $f(\varepsilon) = \exp(-|\varepsilon|)/2$. Least absolute deviations yields β estimates that are less sensitive to outliers, and is therefore known as robust regression. However, unlike least squares, it is non-differentiable so that simplex optimisation procedures are typically used, e.g., see the `L1pack` package in `R` (Osorio, F. and Wolodzko, 2022).

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Differentiable optimisation

Rather than using simplex procedures, we instead make use of the fact that the absolute value function $|z|$ can be approximated by

$$a_\tau(z) = \sqrt{z^2 + \tau^2} - \tau,$$

for $\tau > 0$, in the sense that $a_\tau(z) \to |z|$ as $\tau \to 0$. The advantage of the function $a_\tau(z)$ is that it is differentiable, i.e.,

$$a'_\tau(z) = \frac{z}{\sqrt{z^2 + \tau^2}},$$

meaning that, from a statistician's perspective, more standard Newton-type estimation can proceed (in contrast to the use of simplex methods). This differentiable approximation has been used previously in the context of smooth penalised regression for the purpose of variable selection (Jaouimaa et al, 2019; Burke and Patilea, 2021).

Therefore, in the context of Laplace regression, we consider β estimation using the differentiable objective

$$\min_\beta \sum_i a_\tau(y_i - x_i^{\mathrm{T}}\beta),$$

which is equivalent to the use of a smooth (differentiable) Laplace error density,

$$f(\varepsilon) = c_\tau e^{-a_\tau(\varepsilon)}/2,$$

where c_τ is a normalising constant that is required upon replacing $|\varepsilon|$ with $a_\tau(\varepsilon)$ in the Laplace density function, but note that $c_\tau \approx 1$ when τ is small.

We can also form a log-likelihood function for β and σ

$$\ell(\beta, \sigma) = n \log c_\tau - n \log \sigma - \sum_i a_\tau \left(\frac{y_i - x_i^{\mathrm{T}}\beta}{\sigma} \right),$$

and this is differentiable in β and σ; for simplicity of implementation, we make use of the `nlm` optimiser in `R`. Note that the constant c_τ is not needed for the estimation of these parameters, but, of course, would be required for comparing this model to other models, for example, using the Bayesian Information Criterion. In any case, although c_τ must be computed numerically, it can be pre-computed for the given τ value that is being used.

3 Data example

We consider the so-called "stack loss" dataset, which is commonly used as an exemplar for robust regression (Brownlee, 1960). It relates to an industrial process for oxidising ammonia to nitric acid, where the response

variable `stack_loss` is a measure of inefficiency, and there are three input process variables, namely, `air_flow`, `water_temp`, and `acid_conc`.

Table 1 displays the estimated β coefficients from the smooth Laplace regression for three values of τ along with the output of the `L1pack` (which makes use of a true absolute value). We can see that the results for the smooth optimiser are very similar to that of `L1pack`, especially at the smallest value of $\tau = 0.01$ displayed here. In particular, we see that `air_flow` and `water_temp` increase inefficiency while `acid_conc` decreases inefficiency.

TABLE 1. Fitted models

Covariate	$\tau = 0.5$	$\tau = 0.1$	$\tau = 0.01$	`L1pack`
`air_flow`	0.83	0.83	0.83	0.83
`water_temp`	0.69	0.59	0.57	0.57
`acid_conc`	-0.10	-0.07	-0.06	-0.06

4 Summary

The use of a smooth approximation to the absolute value function generates differentiable Laplace regression (i.e., robust regression). This enables the use of Newton-type optimisation, which will be more familiar to statisticians. We believe that the the differentiable absolute is useful more generally outside of robust regerssion, and, indeed, it has been used previously in penalised regression for the purpose of variable selection.

Acknowledgments: This work was supported by the Confirm Smart Manufacturing Centre funded by Science Foundation Ireland (Grant Number: 16/RC/3918).

References

Brownlee, K. A. (1960). Statistical Theory and Methodology in Science and Engineering.

Burke, K. and Patilea, V. (2021). A likelihood-based approach for cure regression models. TEST.

Jaouimaa, F.Z., Ha, I.D. and Burke, K. (2019). Penalized Variable Selection in Multi-Parameter Regression Survival Modelling. arXiv.

Osorio, F. and Wolodzko, T. (2022). Routines for L1 estimation. R package version 0.41-2. https://cran.r-project.org/package=L1pack

`TwoTimeScales`: an R-package for smoothing hazards with two time scales

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

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

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

Abstract: We introduce the R-package `TwoTimeScales` for the analysis of time to event data with two time scales. The package provides tools to estimate a smooth hazard that varies over two time scales and also, if covariates are available, to estimate a proportional hazards model with such a two-dimensional baseline hazard. We describe the features of the package and illustrate options for presentation of results. As an example we analyse mortality of patients with a recurrence of colon cancer.

Keywords: R-package; Time-to-event data; Time scales; P-splines.

1 Introduction

Time to event data can involve more than one time scale. For example, in medical and epidemiological studies, time since disease onset and age of the patient (which is time since birth) may jointly determine the occurrence of an event, such as death or relapse. The hazard over two time scales can be modelled by two-dimensional P-splines, and Carollo et al. (2020) described an approach to estimate this model. In case there are covariates available, the approach can be extended to a proportional hazards (PH) model with a baseline hazard varying over two time scales. To efficiently estimate this model array algorithms are employed (see Currie et al. (2006)). This is particularly relevant in the PH setting.
To make statistical models accessible for data analyses and to promote their usage statistical software is essential. Therefore we developed the R-package `TwoTimeScales` which implements the model presented in Carollo

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

et al. (2020). Here we describe the features of the package and illustrate some of its capacities to present results.
As an example, we will use data on colon cancer patients and will study mortality of patients who experienced a relapse. The two time scales are time since randomization (after surgery) and time since recurrence. Additional covariates are included in a PH regression model. The data are available in the R-package `survival` (Therneau (2023)). In the following section we briefly summarize the essential components of the P-spline model for two-dimensional hazard models. In Section 3 we then describe the functionality of `TwoTimeScales` along the colon cancer example.

2 Smoothing hazards with two time scales

The two time scales are denoted by t (here: time since randomization) and s (here: time since recurrence), where the origin of s is later than the origin of t, so that $t > s$. The vector of covariates is denoted by z. In the example the event of interest is death. We would like to fit the PH model $\lambda(t, s; z) = \lambda_0(t, s) \exp(\beta^{\mathrm{T}} z)$, where $\lambda_0(t, s)$ is the smooth baseline hazard over t and s. β is the vector of regression coefficients. Equivalently we can express the model via $u = t - s$, where u is the time when $s = 0$ (here: time between randomization and recurrence). The value of u differs between individuals. Hence

$$\lambda(t, s; z) = \lambda_0(t, s) \exp(\beta^{\mathrm{T}} z) \equiv \breve{\lambda}_0(u, s) \exp(\beta^{\mathrm{T}} z), \tag{1}$$

and we can estimate $\breve{\lambda}_0(u, s)$ over $u > 0, s > 0$.
To estimate the model the (u, s)-plane is divided into small squares of equal size, and for each individual the number of events and the time at risk in each square is determined. The smooth log-baseline hazard $\eta_0(u, s) = \ln \breve{\lambda}_0(u, s)$ is modelled as a linear combination of tensor products of B-splines, the coefficients are constrained by a roughness penalty. The model is fitted via penalized Poisson regression (for details see Carollo et al. (2020)). Model estimation hence requires two steps: Transforming the input data into the matrices of events and exposures, and then maximizing the penalized Poisson log-likelihood (using array algorithms) for an optimal choice of smoothing parameters.

3 The `TwoTimeScales` package

The R-package `TwoTimeScales` provides a suite of functions to estimate model (1) and to present results. The user typically will have to interact with three functions only, see also Figure 1.

`prepare_data()` performs the two-dimensional binning in individual event- and exposure-matrices and it also sets up the covariate matrices (in the case of a PH model) for the array algorithm. The user

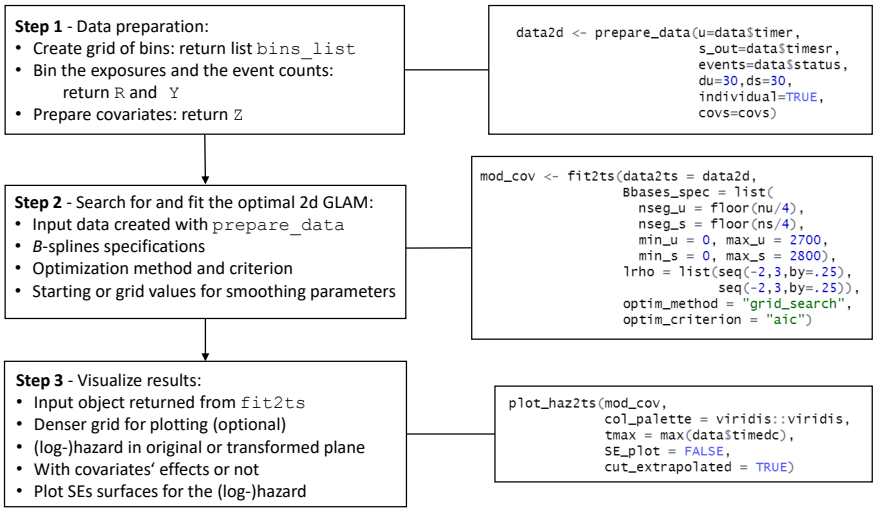


FIGURE 1. Flowchart of main package functions and example of their usage.

provides the bin widths for u and s and the names of variables that contain entry and exit times, respectively, the event indicator and, if applicable, the covariates.

`fit2ts()` performs the actual estimation. Its input most likely was created by `prepare_data()`. The user provides information on the B-splines specification for the baseline hazard, the criterion that should be minimized (AIC or BIC) and whether a grid search over the smoothing parameters should be performed or numerical optimization should be used. Correspondingly, either grid values or starting values for the smoothing parameters can be given. The results of this function contain, among others, the parameter estimates (including standard errors), optimal smoothing parameters, AIC/BIC values, the effective dimension of the estimated model.

`plot_haz2ts()` offers several figures to visualize the results. This includes the estimated (log-)hazard in (t, s)- or (u, s)-coordinates and an image plot of its standard errors. If a PH model was estimated, the regression coefficients and their 95% confidence intervals (or the corresponding hazard ratios) can be plotted. If a grid search was performed the AIC- (BIC-) profile can be shown. Graphical parameters, such as the color palette, can be changed. Figures 2 and 3 show some output for the colon cancer example.

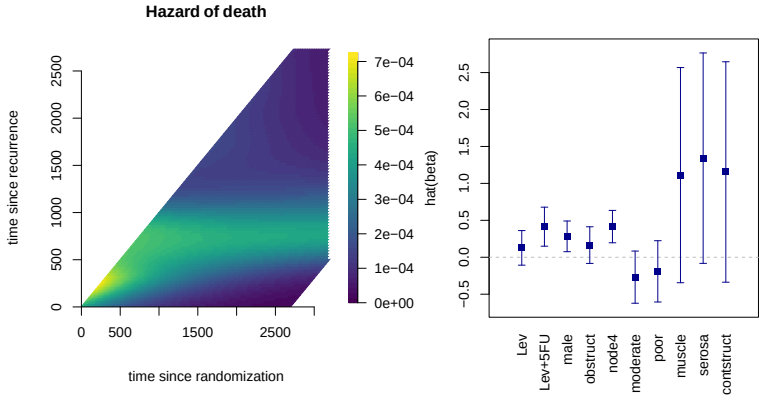


FIGURE 2. Baseline hazard by time since randomization and time since recurrence (left). Estimated covariate effects and 95% confidence intervals (right).

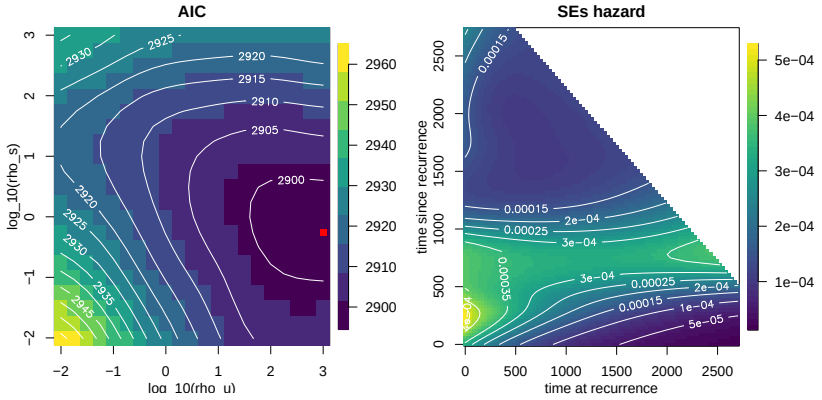


FIGURE 3. AIC profile (left), with ρ_u and ρ_s smoothing parameters. Standard errors (right) for hazard in Fig. 2.

References

Carollo, A., H. Putter, P. Eilers, and J. Gampe (2020). Hazard smoothing along two time scales. In: *Proceeding of the 35th International Workshop on Statistical Modelling*, Bilbao, Spain 31 – 34.

Currie, I. D., M. Durban, and P. H. C. Eilers (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B*, 68 (2), 259–280.

Therneau (2023). A Package for Survival Analysis in R. *R package v. 3.5-5*.

A new statistical methodology to detect earnings management

M. Chavent[1], V. Darmendrail[2], D. Feral[1], H. Lorenzo[1], F. Pourtier[2], J. Saracco[1]

[1] Univ. Bordeaux, CNRS, INRIA, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France
[2] Univ. Bordeaux, IRGO, UR 4190, F-33000 Bordeaux, France

E-mail for correspondence: jerome.saracco@inria.fr

Abstract: This communication presents a new statistical methodology to detect earnings management associated with the zero earnings threshold. An EM algorithm is proposed to estimate the underlying parameters. The good numerical behavior of the methodology is illustrated on simulated data close to real data.

Keywords: EM algorithm; density estimation; Mixture model; Truncated Gaussian distribution; Earnings management.

1 Motivation and underlying mixture model

The practice of earnings management by firms have been long recognized by researchers. Accounting literature have shown that they are inclined to manage their accounting profits and losses to exceed specific thresholds such as zero earnings, earnings of prior year or analysts' forecasts (Burgstahler & Chuk, 2017). Since the pervasiveness of earnings management significantly compromises the integrity of financial reporting, regulators or investors are interested in detecting earnings management and identifying its frequency and its magnitude.
In this communication, a new statistical methodology to detect earnings management associated with the zero earnings threshold is presented and is illustrated on simulated data very similar to real data.
In the following, a statistical model will first be detailed in order to model this earnings management phenomenon. Then a simplified model will then be considered. The procedure for estimating the parameters of the simplified model will then be described.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Statistical model for earnings management

To recover the frequency and the magnitude of earnings management across all the firms, the true earnings must be modelled. Let us denote by X_i the true earnings of the firm i, $i = 1, \ldots, n$. Note that these true earnings are unobservable. The true earnings X_i are assumed to be independent and identically distributed: X_i follows a mixture of two Gaussian distribution of density

$$\pi\varphi_A(x) + (1-\pi)\varphi_B(x),$$

where $\pi \in]0,1[$ and φ_A (resp. φ_B) is the density of the Gaussian distribution with mean μ_A (resp. μ_B) and variance σ_A^2 (resp. σ_B^2). The choice of a Gaussian mixture allows modelling a distribution with tails of different weights, as it is often observed on the underlying data.
When a firm has an actual true earning value below the zero threshold value, this firm may engage in earnings management and the corresponding reported earning will be greater than the zero threshold value. Only the reported earnings Y_i, $i = 1, \ldots, n$, are observable.
When earnings management is effective for a firm i, the reported earning Y_i is assumed to follow the Exponential distribution with parameter λ, otherwise Y_i follows the same distribution as X_i. In order to model the transition to earnings manipulation for a firm i, the random variable T_i is introduced: when X_i is less than the zero threshold value, T_i given $X_i = x$ follows a Bernoulli distribution of parameter $\tau(x)$. Note that the true frequency of earning management p can be obtained from $\tau(.)$.

3 A simplified statistical model and its parameter estimation procedure

The underlying idea is to focus only on the subpopulation of firms associated with the Y_i's greater than the zero threshold value. From a mathematical point of view, the work is implicitly developed conditionally to $Y \geq 0$. The corresponding reported earnings of these firms can be viewed as a mixture between "true" earnings and "managed" earnings, which can be modelled as follows:

$$\forall y \geq 0, \quad f(y) = qf_1(y) + (1-q)f_2(y), \tag{1}$$

where $q \in]0,1[$, f_1 is the density of the Exponential distribution with rate parameter λ, and f_2 is the density of a mixture of two truncated Gaussian distributions:

$$\forall y \geq 0, \quad f_2(y) = \tilde{\pi}\varphi_A^{(+)}(y) + (1-\tilde{\pi})\varphi_B^{(+)}(y),$$

where $\varphi_A^{(+)}$ (resp. $\varphi_B^{(+)}$) is the density of the truncated Gaussian distribution with mean μ_A (resp. μ_B) and variance σ_A^2 (resp. σ_B^2)

defined as follows: $\forall t \geq 0, \quad \varphi_A^{(+)}(t) = \dfrac{\varphi_A(t)}{C_A}$, with $C_A = \displaystyle\int_0^{+\infty} \frac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left(-\frac{(s-\mu_A)^2}{2\sigma_A^2}\right) ds = \overline{\Phi}\left(-\frac{\mu_A}{\sigma_A}\right)$ with $\overline{\Phi}(t) = 1 - \Phi(t)$ and Φ the cumulative distribution function of the standardized Gaussian distribution. The underlying parameter of the simplified model is thus:

$$\theta = \left(q, \lambda, \tilde{\pi}, \mu_A, \sigma_A^2, \mu_B, \sigma_B^2\right).$$

Given the observed sample of the reported earnings $\mathcal{D}_n = \{y_1, \ldots, y_n\}$, the objective is to estimate θ as the maximum of the likelihood

$$\hat{\theta} = \left(\hat{q}, \hat{\lambda}, \hat{\tilde{\pi}}, \hat{\mu}_A, \hat{\sigma}_A^2, \hat{\mu}_B, \hat{\sigma}_B^2\right) := \arg\max_{\theta} \ \ell(\theta; \mathcal{D}_n)$$

where $\ell(\theta, \mathcal{D}_n) = \prod_{i=1}^n f(y_i)$. For that purpose, an expectation–maximization (EM) algorithm (see Dempster et al., 1977) has been developed. This algorithm relies on the complete-data likelihood based on two dichotomous latent variables: $\tilde{T}_i$ (resp. $\tilde{Z}_i$) which follows a Bernoulli distribution with parameter q (resp $\tilde{\pi}$). The latent variable $\tilde{T}_i$ indicates whether the firm i has managed earnings or not, while the variable $\tilde{Z}_i$ is used to manage the mixture of the two truncated Gaussian distributions. The "completed" density is:

$$f(y, t, z; \theta) = \left(q f_1(y)\right)^t (1-q)^{1-t} \left(\tilde{\pi}\varphi_A^{(+)}(y)\right)^{(1-t)z} \left((1-\tilde{\pi})\varphi_B^{(+)}(y)\right)^{(1-t)(1-z)}.$$

Assuming that the $(Y_i, \tilde{T}_i, \tilde{Z}_i)$'s are independent, the "completed" log-likelihood is written this way:

$$\begin{array}{rcl} \ell\ell(\theta; \mathcal{D}_n, \mathcal{L}_n) & = & \sum_{i=1}^n \left[t_i \log(q) + (1-t_i)\log(1-q) + t_i \log(f_1(y_i))\right. \\ & + & (1-t_i) z_i \log(\tilde{\pi}) + (1-t_i) z_i \log\left(\varphi_A(y_i)\right) \\ & + & (1-t_i)(1-z_i)\log(1-\tilde{\pi}) + (1-t_i)(1-z_i)\log\left(\varphi_B(y_i)\right) \\ & - & \left.(1-t_i) z_i \log(C_A) - (1-t_i)(1-z_i)\log(C_B)\right], \end{array}$$

where $\mathcal{L}_n = \{t_1, \ldots, t_n, z_1, \ldots, z_n\} \in \{0,1\}^{2n}$.
The EM algorithm is an iterative method which starts from $\hat{\theta}_{(0)}$, an initial value of the parameter θ, and provides at each iteration j

$$\hat{\theta}_{(j+1)} := \arg\max_{\theta} \mathbb{E}\left[\ell\ell(\theta; \mathcal{D}_n, T_1, ..., T_n, Z_1, ..., Z_n) \mid \mathcal{D}_n; \hat{\theta}_{(j)}\right] \qquad (2)$$

where $\hat{\theta}_{(j)} = \left(\hat{q}_{(j)}, \hat{\lambda}_{(j)}, \hat{\tilde{\pi}}_{(j)}, \hat{\mu}_{A(j)}, \hat{\sigma}^2_{A(j)}, \hat{\mu}_{B(j)}, \hat{\sigma}^2_{B(j)}\right)$ is a current estimation of θ. Note that, since the EM algorithms are known to be sensitive to initialization, an "expert" data-driven initialization (not detailed here) has been proposed and provides relevant numerical results in the considered simulation study. The two steps (E and M) can be described as follows.

• **The E step** provides estimates of the conditional expectations of the latency variables such as, $\forall i = 1, \ldots, n$,

$$\begin{aligned} < t_i >_{j+1} &= \hat{\mathbb{E}}\left(\tilde{T}_i | \hat{\theta}_{(j)}, \mathcal{D}_n\right) = \hat{\mathbb{P}}\left(\tilde{T}_i = 1 | \hat{\theta}_{(j)}, y_i\right) \\ &= \frac{\hat{q}_{(j)} f_{1,\hat{\lambda}_{(j)}}(y_i)}{\hat{q}_{(j)} f_{1,\hat{\lambda}_{(j)}}(y_i) + (1-\hat{q}_{(j)})\left(\hat{\bar{\pi}}_{(j)} \hat{\varphi}_{A(j)}(y_i) + (1-\hat{\bar{\pi}}_{(j)})\hat{\varphi}_{B(j)}(y_i)\right)}, \end{aligned}$$

and

$$\begin{aligned} < z_i >_{j+1} &= \hat{\mathbb{E}}\left(\tilde{Z}_i | \hat{\theta}_{(j)}, \mathcal{D}_n\right) = \hat{\mathbb{P}}\left(\tilde{Z}_i = 1 | \hat{\theta}_{(j)}, y_i\right) \\ &= \frac{\hat{\bar{\pi}}_{(j)} \hat{\varphi}_{A(j)}(y_i)}{\hat{\bar{\pi}}_{(j)} \hat{\varphi}_{A(j)}(y_i) + (1-\hat{\bar{\pi}}_{(j)})\hat{\varphi}_{B(j)}(y_i)}, \end{aligned}$$

where $f_{1,\hat{\lambda}_{(j)}}$ stands for the Exponential distribution with rate parameter $\hat{\lambda}_{(j)}$ and $\hat{\varphi}_{A(j)}$ (resp. $\hat{\varphi}_{B(j)}$) for the Gaussian distribution with mean $\hat{\mu}_{A(j)}$ (resp. $\hat{\mu}_{B(j)}$)and variance $\hat{\sigma}^2_{A(j)}$ (resp. $\hat{\sigma}^2_{B(j)}$).

• **The M step** solves the optimization problem (2) which provides:

$$\begin{aligned} \hat{q}_{(j+1)} &= \frac{1}{n}\sum_{i=1}^n < t_i >_{j+1}, \quad \hat{\lambda}_{(j+1)} = \frac{\sum_{i=1}^n < t_i >_{j+1}}{\sum_{i=1}^n < t_i >_{j+1} x_i}, \\ \hat{\bar{\pi}}_{(j+1)} &= \frac{\sum_{i=1}^n (1- < t_i >_{j+1}) < z_i >_{j+1}}{\sum_{i=1}^n (1- < t_i >_{j+1})}, \\ \hat{\mu}_{A(j+1)} &= \frac{\sum_{i=1}^n (1- < t_i >_{j+1}) < z_i >_{j+1} x_i}{\sum_{i=1}^n (1- < t_i >_{j+1}) < z_i >_{j+1}} \\ &- \hat{\sigma}_{A(j+1)} \frac{\varphi\left(-\hat{\mu}_{A(j+1)}/\hat{\sigma}_{A(j+1)}\right)}{\overline{\Phi}\left(-\hat{\mu}_{A(j+1)}/\hat{\sigma}_{A(j+1)}\right)}, \\ \hat{\sigma}^2_{A(j+1)} &= \frac{\sum_{i=1}^n (1- < t_i >_{j+1}) < z_i >_{j+1} (x_i - \hat{\mu}_{A(j+1)})^2}{\sum_{i=1}^n (1- < t_i >_{j+1}) < z_i >_{j+1}} \\ &+ \hat{\sigma}_A \hat{\mu}_{A(j+1)} \frac{\varphi\left(-\hat{\mu}_{A(j+1)}/\hat{\sigma}_{A(j+1)}\right)}{\overline{\Phi}\left(-\hat{\mu}_{A(j+1)}/\hat{\sigma}_{A(j+1)}\right)}, \\ \hat{\mu}_{B(j+1)} &= \frac{\sum_{i=1}^n (1- < t_i >_{j+1})(1- < z_i >_{j+1}) x_i}{\sum_{i=1}^n (1- < t_i >_{j+1})(1- < z_i >_{j+1})} \\ &- \hat{\sigma}_{B(j+1)} \frac{\varphi\left(-\hat{\mu}_{B(j+1)}/\hat{\sigma}_{B(j+1)}\right)}{\overline{\Phi}\left(-\hat{\mu}_{B(j+1)}/\hat{\sigma}_{B(j+1)}\right)}, \\ \hat{\sigma}^2_{B(j+1)} &= \frac{\sum_{i=1}^n (1- < t_i >_{j+1})(1- < z_i >_{j+1})(x_i - \hat{\mu}_{B(j+1)})^2}{\sum_{i=1}^n (1- < t_i >_{j+1})(1- < z_i >_{j+1})} \\ &+ \hat{\sigma}_{B(j+1)} \hat{\mu}_{B(j+1)} \frac{\varphi\left(-\hat{\mu}_{B(j+1)}/\hat{\sigma}_{B(j+1)}\right)}{\overline{\Phi}\left(-\hat{\mu}_{B(j+1)}/\hat{\sigma}_{B(j+1)}\right)} \end{aligned}$$

where φ is the density of the standardized Gaussian distribution.
One might notice that $\hat{p}_{(j+1)}$, $\hat{\lambda}_{(j+1)}$ and $\hat{\pi}_{(j+1)}$ have explicit forms, while $(\hat{\mu}_{A(j+1)}, \hat{\sigma}^2_{A(j+1)})$ (resp. $(\hat{\mu}_{B(j+1)}, \hat{\sigma}^2_{B(j+1)})$) are solutions of a nonlinear system with two equations (which can be numerically solved using a specific iterative algorithm, not detailed here).

4 Illustration on simulated datasets

In this numerical study, two scenarios have been considered. The corresponding simulated datasets of size $n = 2000$ have been generated from Model (1) with the value of θ described in Table 1. In the first case, it is relatively easy to identify and estimate all the components of θ. The second one is more complex, but corresponds to more realistic situations in the context of earnings management. It is strongly inspired by the first dataset used by Chen et al. (2010).

TABLE 1. Parameters used in the numerical study.

	q	λ	$\tilde{\pi}$	μ_A	σ_A	μ_B	σ_B
Simu. 1	0.1	12	0.3	1	1	4	0.5
Simu. 2	0.05	185.3	0.632	0.007	0.123	0.06	0.034

The obtained numerical results are respectively provided in Figures 1 and 2. The true density and parameter values are plotted in blue, and the corresponding estimated ones are plotted in green. In order to have visibility on the variability of the estimators, $B = 100$ bootstrap samples were generated, and the associated parameters were then estimated. The variability of the estimated density is provided via the Bootstrap 90% confidence interval (in red), as well as those of the estimate of the components of θ via the corresponding boxplots. It can be clearly observed that the estimation procedure makes it possible to properly recover the θ parameters and thus the true density of the Y_i's.

5 Concluding remarks

From the estimated parameter $\hat{\theta}$ of the simplified statistical model, it is possible to obtain an estimation of the initial statistical model and to retrieve the frequency and the magnitude of earnings management on the whole population of firms. Based on the Bootstrap variability, it will also be possible to compare models based on two populations, as different countries, different sub-periods, or different categories of firms. Furthermore, as listed by Byzalov and Basu (2019), there are other fields of application for such methods, where a decision-maker has both an opportunity and an incentive to move from just below to just above a benchmark or vice versa, through manipulation and/or extra effort.

References

Burgstahler, D. and Chuk, E. (2017). What have we learned about earnings management? Integrating discontinuity evidence. *Contemp. Account. Res.*, 34(2), 726749.

Byzalov, D. and Basu, S. (2019). Modeling the determinants of meet-or-just-beat behavior in distribution discontinuity tests. *Journal of Accounting and Economics*, 68(23), 101266.

Chen, S.K., Lin, B-X., Wang, Y. and Wu, L. (2010). The frequency and magnitude of earnings managements: Time-series and multi-threshold comparisons. *International Review of Economics and Finance*, **19**, 671 – 685.

Dempster, A.P., Laird, N.M. and Rubin, D. B.(1977).Maximum Likelihood from Incomplete Data via the EM Algorithm. *J.R. Stat. Soc., Ser. B*, **39**,1 – 38.

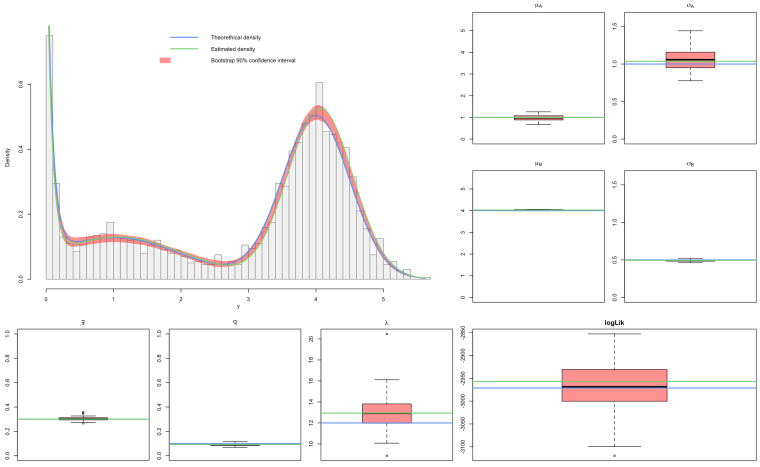


FIGURE 1. Simulation 1: true density (in blue) and estimated density (in green), as well as true and estimated values of θ, with bootstrap variability (in red).

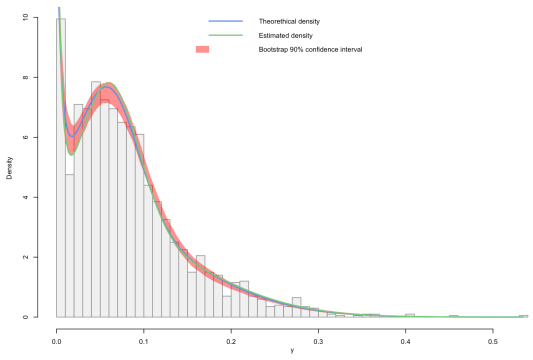


FIGURE 2. Simulation 2: true density (in blue) and estimated density (in green) with bootstrap variability (in red).

Automatic effect selection for generalized additive models

Claudia Collarin[1], Matteo Fasiolo[2], Claudio Agostinelli[3]

[1] Department of Statistical Sciences, University of Padua, Padua, Italy
[2] School of Mathematics, University of Bristol, Bristol, UK
[3] Department of Mathematics, University of Trento, Trento, Italy

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

Abstract: Forecasts of electricity net–load, consumption less embedded generation, are key inputs for operations such as trading and power production planning. Here we focus on Great Britain's power network, which is divided into fourteen regions, the grid supply points groups. Each exhibits specific demand and embedded power generation characteristics, implying that different models should be used to predict net–load in each region. In addition, regional net–load is determined by several social and meteorological factors, which interact with each other. Given that including all possible interactions is infeasible from both a statistical and a computational perspective, here we propose a model selection method to automatically choose the main effects and the first–order interactions to include in, region–specific, generalized additive models. The proposed method combines gradient boosting for effect exploration and the Lasso for effect selection. The results from simulations and net–load electricity data demonstrate the selective and predictive power of the proposed algorithm.

Keywords: Generalized Additive Models; Gradient Boosting; Model Selection; Electricity Net–load Forecast.

1 Introduction

Electricity operators need accurate and interpretable short–term power demand forecasts to make production planning, grid management, and trading decisions. In this context, Generalized Additive Models (GAM, Hastie and Tibshirani, 1987) offer an attractive balance between flexibility and interpretability. However, selecting which effects to include in a GAM is a non–trivial task, particularly when interactions are considered. While GAM model selection can be performed jointly with fitting via the L_2Boost

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

algorithm (Bühlmann and Yu, 2003), the selected model often include non–significant effects. This motivated the development of several methods (see, e.g., Bühlmann and Hothorn (2010) and Strömer et al. (2022)) aimed at enhancing variable selection for gradient boosting. Here we propose a new method, which combines gradient boosting with the lasso (Tibshirani (1996)), and that is specifically aimed at improving model selection when considering interactions between covariates.

2 Methodology

Let X be a $n \times k$ matrix of covariates. Assume the response y_i, with $i = 1, \ldots, n$, follows the GAM model $y_i = \sum_{s \in \mathcal{I}} f_s(x_{i,s}) + \sum_{t \in \mathcal{C}} g_t(x_{i,t_1}, x_{i,t_2}) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. where $\mathcal{I}$ and $\mathcal{C}$ denote the set of main effects and interactions, respectively. Then, proceed as follows:

Step 1: Use gradient boosting to fit a GAM model based on the base–learners f_s for $s \in \mathcal{I}_1$, where $\mathcal{I}_1$ is user–defined.

Step 2: Let $F_1 = (f_s(x_s))_{s \in \mathcal{I}_1}$ be the $n \times |\mathcal{I}_1|$ matrix whose columns are the effects fitted in step 1. Hence, the fitted values from step 1 are $\hat{\mathbf{y}} = F_1 \, \mathbf{j}_1, \ \mathbf{j}_1 = (1, \ldots, 1)^{\mathrm{T}}$. Regress $\mathbf{y}$ on F_1 via the lasso and let $\mathcal{I}_2$ be the set of non–zero coefficients.

Step 3: Let $\mathcal{C}_1 = \{(i, j) \mid i, j \in \mathcal{I}_2, \ i < j\}$ be the set of unique pairs of indices in $\mathcal{I}_2$. Run L_2Boost with main and interactive effects in $\mathcal{I}_2$ and $\mathcal{C}_1$ as base learners and store in $\mathcal{I}_3$ and $\mathcal{C}_2$ the selected effects indices.

Step 4: Denote with $F_2 = (f_s(x_s), g_t(x_{t_1}, x_{t_2}))_{s \in \mathcal{I}_3, t \in \mathcal{C}_2}$ the $n \times (|\mathcal{I}_3| + |\mathcal{C}_2|)$ matrix whose columns are the effects fitted in step 3. The fitted values from step 3 are then $\hat{\mathbf{y}} = F_2 \, \mathbf{j}_2, \ \mathbf{j}_2 = (1, \ldots, 1)^{\mathrm{T}}$. Regress $\mathbf{y}$ on F_2 via the lasso and let $\mathcal{I}_4$ and $\mathcal{C}_3$ be the final sets of selected covariate indexes and interactions.

3 Simulation Study

Table 1 ana 2 shows the results of a simulation study, consisting of 70 runs of 2000 independent training and 2000 testing observations, generated by $y_i = \sum_{j=1}^{8} f_j(x_{i,j}) + f(x_{i,2}, x_{i,3}) + \varepsilon_i, \ i = 1, \ldots, 2000$, where $x_i \sim U([0, 1])$, $\varepsilon_i \sim N(0, \sigma^2)$ and with f_j's summarized below.

$f_0(x_0) = x_0^4$	$f_1(x_1) = x_1^2$	$f_4(x_4) = \frac{1}{3}(1 + x_4)^5$
$f_5(x_5) = -(3x_5 - 1.5)^7 + -(4x_5 - 0.5)^3$	$f_6(x_6) = x_6(1 + x_6)^3$	$f_7(x_7) = e^{2(x_7+1)}/2$
$f_8(x_8) = -10x_8$	$f_{23}(x_2, x_3) = -\frac{1}{2}\left(3\cos(x_2 + x_3) - 2\sin(x_2)\right)^4$	

Additional 20 noise covariates following $U([-0.5, 0.5])$ were added to each observation. Thin–plate and cubic regression splines were used as base–learners for main effects and interactions, respectively. The step size ν in gradient boosting was fixed to 0.1.

TABLE 1. Results on simulated data. σ^2 is the noise variance, MSE_i and MSE_{gam} are mean squared errors obtained in steps 1, 4 and by fitting the response with the true GAM model. R^2 is the variance explained in step 4.

σ^2	$\textbf{MSE}_1$		$\textbf{MSE}_4$		$\textbf{MSE}_{gam}$		R^2	
	Mean	**sd**	**Mean**	**sd**	**Mean**	**sd**	**Mean**	**sd**
2.8	5.995	0.667	2.883	0.268	2.813	0.278	0.952	0.002
4.0	6.522	1.128	4.067	0.527	4.022	0.519	0.903	0.003
8.8	10.352	3.499	8.945	2.503	8.833	2.755	0.659	0.011
10.5	11.794	3.882	10.680	3.558	10.538	3.866	0.578	0.011

TABLE 2. Frequency of selection for different values of σ. The last column, denoted as f_k, indicates frequencies of erroneously selected effects.

σ	$f_0(x_0)$	$f_1(x_1)$	$f_2(x_2)$	$f_3(x_3)$	$f_4(x_4)$	$f_5(x_5)$	$f_6(x_6)$	$f_7(x_7)$	$f_8(x_8)$	$f_{23}(x_2,x_3)$	f_k
2.80	1	0.99	1	1	1	1	1	1	1	1	0
4	0	0	1	1	1	1	1	1	1	1	0
8.80	0	0	1	1	1	1	1	1	1	1	0.07
10.50	0	0	1	1	1	1	1	1	1	1	0.51

4 Forecasting of regional net–load in UK

Half–hourly net–load and weather forecast data from Browell and Fasiolo (2021) cover the period 2/1/2014–31/12/2018, totaling 91726 observations for each GSP group. We focus on four groups: East England (A), London (C), South West England (L) and North Scotland (P). We use 2018 data for testing, the rest is split into three folds to select the number of boosting steps via cross–validation.

To avoid bias selection all base–learners should have equal degrees of freedom (see Hofner et al., 2011). We fixed them to six. As base–learners, we used thin–plate and cubic regression splines for main effects and interactions, respectively. The step size ν in gradient boosting was fixed to 0.1. Results are summarized in Tables 3 and 4.

5 Conclusion

Interpretability is indispensable for the operational adoption of new models in high–stakes applications such as net–load forecasting. As demonstrated by the examples discussed here, the proposed GAM model selection approach aids interpretability by reducing the number of effects in the model, relative to gradient boosting, while not compromising accuracy. It also finds

TABLE 3. Selected effects for GSP group models. From 1 to 8: day of the week, day of the year, time of day, 2–week rolling mean net–load, wind speed at 10 metres, 48–hour rolling temperature, temperature, solar irradiance.

GSP	1	2	3	4	5	6	7	8	2–3	2–4	3–4	3–8	6–7
A	•	•	•	•	•	•	•	•	•	•		•	
L	•	•	•	•	•	•		•	•	•	•	•	
P	•	•	•	•	•	•						•	•
C	•	•	•			•	•	•					

TABLE 4. Results on net–load data, for each GSP group. MSE_i and n_i are the mean squared error and the number of smooths selected in steps 1, 3 and 4. R^2 is the variance explained at the end of the procedure.

GSP	$\mathbf{n}_1$	$\mathbf{n}_2$	$\mathbf{n}_3$	$\mathbf{MSE}_1$	$\mathbf{MSE}_3$	$\mathbf{MSE}_4$	R^2
A	11	32	11	0.108	0.097	0.097	0.881
L	11	32	11	0.195	0.173	0.192	0.820
P	11	17	6	0.684	0.698	0.258	0.804
C	11	21	8	0.062	0.057	0.057	0.939

and selects first–order interactions in an automatic fashion. This is practically important when modelling net–load at a regional (or lower) level of aggregation, because manual model selection is made difficult by the fact that each region has different characteristics, and that the number of candidate interactive effects is large.

Acknowledgments: Claudia Collarin PhD scholarship is funded by PON "Research and Innovation" 2014–2020 Action IV.5 "PhDs on Green issues." – Ministerial Decree 1061/2021.

References

Bühlmann, P., and Hothorn, T. (2010). Twin boosting: improved feature selection and prediction. *Statistics and Computing*, **20**, 119 – 138.

Bühlmann, P., and Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, **98(462)**, 324 – 339.

Hastie, T., and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, **82(398)**, 371 – 386.

Hofner, B., Hothorn, T., Kneib, T., and Schmid, M. (2011). A framework

for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, **20(4)**, 956 – 971.

Browell, J., and Fasiolo, M. (2021) Probabilistic forecasting of regional net–load with conditional extremes and gridded NWP. *IEEE Transactions on Smart Grid*, **12(6)** 5011 – 5019.

Strömer, A., Staerk, C., Klein, N. , Weinhold, L., Titze, S., and Mayr, A. (2022). Deselection of base–learners for statistical boosting—with an application to distributional regression. *Statistical Methods in Medical Research*, **31(2)**, 207 – 224.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58(1)**, 267 – 288.

A multifidelity framework for wind speed data

Pietro Colombo[1], Claire Miller[1], Ruth O'Donnell[1], Xiaochen Yang[1]

[1] University of Glasgow, UK

E-mail for correspondence: pietro.colombo@glasgow.ac.uk

Abstract: Monitoring wind speed is essential to develop offshore wind farms. However, recorded wind data often lack the necessary accuracy for understanding the profitability of the wind farm, and even when they exist, they are scarce in time or space. Intuitively, using multiple data sources could balance the trade-off between scarcity and accuracy. A multi-fidelity framework in the form of the autoregressive Gaussian process is introduced to analyze wind speed reanalysis data fusing datasets of different reliability and resolution to provide a more accurate wind speed data product.

Keywords: Multifidelity; Gaussian process; Wind speed.

1 Introduction

Offshore wind speed data are obtained through different means such as in-situ field sampling using anemometric or Lidar technologies, or processed satellite retrievals. The former provides high-fidelity (high quality) and high-resolution measurements but is limited in temporal and spatial coverage, while the latter offers larger coverage but with low-fidelity (low quality) and low-resolution. Data fusion of two products can, in principle, provide a more informative data stream.
The development of a series of offshore wind farms on the Italian Adriatic coast is our motivational study case. The project known as Agnes (Adriatic green network of energy sources) aims to build a hub for renewable energy. For the wind speed data, the companies involved in the project rely on two main data sources:

1. The ERA5 reanalysis data [ERA5 Data], which contains hourly wind speed measurements from 1979 until the present.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2. The wind climatology obtained through two Lidar installations. These measurements tend to be more reliable than those of ERA; however, they come as point samples, covering a minimal spatial surface.

We developed a simulation study using ERA5 reanalysis data (from the Agnes location), from which we derived two datasets, one of high-fidelity (HF), closer to the true wind speed, but with low temporal sampling rate, the other of low-fidelity (LF) but with high temporal sampling rate. This paper evaluates the performance of an autoregressive Gaussian process (ARGP) [Le Gratiet L. & Garnier J.(2014)] for data fusion of multi-fidelity data. Through the multi-fidelity framework, we aim to return predictions that are more accurate and abundant than those based only on a single data source.

2 Experimental Design

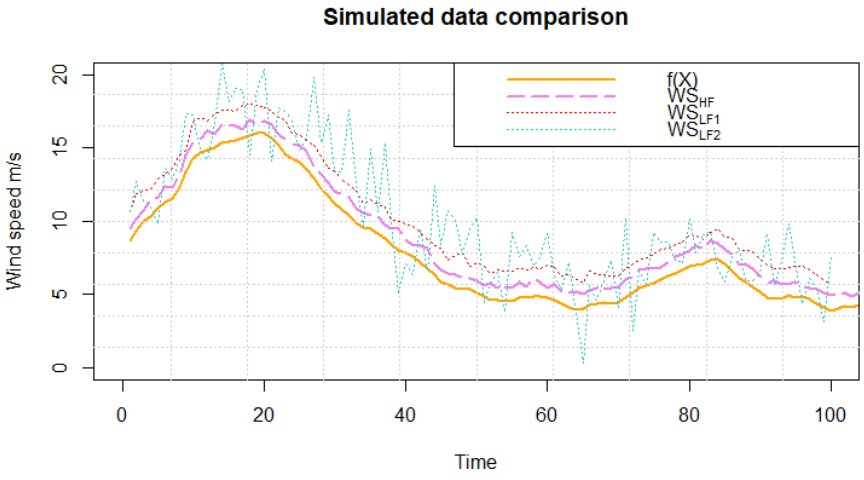
We have simulated two data sources (time-series) that resemble wind speed measurements such that:

$$ws_{HF} = f(x) + e_{HF}(x) \tag{1}$$

$$ws_{LF} = f(x) + e_{LF}(x). \tag{2}$$

The ws_{HF} represents the high-fidelity measurements, therefore closer to $f(x)$, the true wind speed at the index location x, while ws_{LF} is a low-fidelity measurement. The time-series are distinguished by the normally distributed corruptions e_{HF} and e_{LF}, with the low-fidelity corruptions ($e_{LF1} \sim N(2, 0.2)$ and $e_{LF2} \sim N(2, 1)$) being roughly double that of the high-fidelity corruption ($e_{HF} \sim N(1, 0.2)$). HF data are typically scarce, therefore we performed an additional sampling of them of size $N1 < N$. Starting from the ERA5 reanalysis wind speed data, we proceed with a series of decompositions to extract the deterministic part of these data: $f(x)$, which we assume to be composed of a long term trend, a seasonal pattern and potentially other cyclical components. Given a roughly normal remainder, we can generate different ws_{HF} and ws_{LF} adding normally distributed errors. Figure 1 depicts an example of data constructed with such an approach. Given such a design, we compared the model performances of ARGP with two mono-fidelity models: the quantile gradient-boosted regression tree (QGBRT) [Kriegler B. & Berk R. (2010)], a model often used in wind forecasting that provides the prediction for all quantiles of a distribution (hence a deeper understanding of the uncertainty), and a standard GP. We controlled for different $N1$ sample sizes and used as a performance metric the mean absolute deviation (MAE) of the residuals $r = f(x) - P_i$, where P_i is the i^{th} model prediction. The experimental comparison has $N = 850$, 100 replications of randomly drawn errors e_{LF} and e_{HF} and the sub-sample $N1$ index position.

FIGURE 1. Comparison of the four generated signals: in orange $f(x)$ the assumed true wind speed, in violet ws_{HF} the high-fidelity time-series, in dark orange a low noise low-fidelity time-series ws_{LF1} and in turquoise a noisy version of the low-fidelity data ws_{LF2}.



3 Models

3.1 Multifidelity:ARGP

In the ARGP model, the high-fidelity data are modelled as a scaled sum of the lower-fidelity data:

$$GP_{HF}(x) = \rho GP_{LF}(x) + \epsilon(x), \tag{3}$$

where $GP_{HF}(x)$ is a Gaussian process modelling the HF data, $GP_{LF}(x)$ a Gaussian process modelling the LF data, ρ is the degree of correlation between the HF and LF data and $\epsilon(x) \sim GP(\mu_\epsilon, \Sigma_\epsilon)$ is an independent Gaussian process denoting the error structure between HF and LF data. Our simulation design presents a nested structure $D_{HF} \subset D_{LF}$; therefore, we can use the recursive formulation of the model proposed by [Le Gratiet L. & Garnier J.(2014)], which guarantee an efficient maximum likelihood inference.

3.2 Monofidelity: GP and QGBRT

In opposition to ARGP, we tested a standard Gaussian process (GP) fitted only with LF and HF separately, denoted by the notation GP_{HF}, GP_{LF} and

a QGBRT in which a quantile loss function is combined with a gradient-boosted regression tree, also fitted using only one dataset and denoted by $QGBRT_{LF}$ and $QGBRT_{HF}$.

4 Results and Discussion

By comparing the models, with x being a time index, for different $N1$ high fidelity sample sizes, and low-fidelity noise setting $e_{LF1} \sim N(2, 0.2)$, we obtained the results in Table 1. ARGP outperformed the other models for small $N1$ sample sizes, while for $N1 > 160$ its performance was equivalent to those of GP_{HF} and $QGBRT_{HF}$. It also appears there is no notable advantage with highly noisy data. For our simulation design, with $N1$=32, the estimates from the ARGP which combined the low and high-fidelity data were, on average, 1 m/s closer to the $f(x)$, which consists of a 16% improvement compared to the unprocessed information. The multi-fidelity framework has been successfully applied to multiple environmental applications; however, its application to a wind case study is new. This work has illustrated that potentially modelling wind speed data with the multi-fidelity framework is appropriate. To further improve the methodology, three directions of future development have been identified: expansion to include the spatial dimension, exploration of non-linear methodologies for noisy data, and integration of techniques to address data skewness.

TABLE 1. MAE summary from 100 replications for a simulationw with different $N1$ high- fidelity samples size, with low fidelity data error structure equal to $e_{LF} \sim N(2, 0.2)$; The table contains the performance of 5 models: two mono-fidelity GP, two mono-fidelity QGBRT and a multi-fidelity ARGP. In the parenthesis, the MAE standard deviation.

MODELS	$N1$=32(sd)	$N1$=96(sd)	$N1$=160(sd)
GP_{LF}(Time)	0.54(0.018)	0.54(0.018)	0.54(0.018)
GP_{HF}(Time)	0.43(0.048)	0.32(0.011)	0.30(0.008)
ARGP($Time$)	0.29(0.047)	0.29(0.060)	0.29(0.015)
QGBRT$_{LF}$(Time)	0.55(0.020)	0.55(0.020)	0.55(0.020)
QGBRT$_{HF}$(Time)	0.52(0.005)	0.39(0.021)	0.34(0.016)

References

ERA5 Data https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels?tab=overview.

Kriegler, B., & Berk, R. (2010) Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting. *Ann. Appl. Stat* **4(3)**: 1234-1255.

Le Gratiet, L., & Garnier, J. (2014) Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, **4(5)**.

Group penalized models with an adaptive non-convex penalty function

Daniele Cuntrera[1], Vito M.R. Muggeo[1], Luigi Augugliaro[1]

[1] Università degli studi di Palermo, Dip.to Sc Econom, Az e Statistiche, Palermo

E-mail for correspondence: daniele.cuntrera@unipa.it

Abstract: Group penalized models have gained much interest recently due to their ability to handle high-dimensional data with complex structures. This paper proposes an extension of the adaptive non-convex penalty function introduced in Cuntrera et al. (2022a) to a group penalized context. The proposed method respects the grouping structure of the variables and simultaneously shrinks entire groups towards zero to produce a sparse model. The performance of the proposed method is evaluated through a numerical study and compared with the principal competitors. Results show that the proposed method performs well and is a viable option for variable selection.

Keywords: concave group selection; grouped data; sparsity

1 Introduction

In recent years, group penalized models have attracted much interest due to their ability to deal with high-dimensional data with complex structures. A grouping structure that collects the coefficients into groups with similar properties or functions is present in many real-world scenarios where the data are high-dimensional and exhibit the grouping feature. Typical examples of grouped variables are the dummies relevant to the same categorical variable or the powers variables for polynomial terms. This is particularly helpful in fields such as genetics and finance, where variables often occur in groups and are naturally interdependent. Variable selection methods have been proposed to address the challenges posed by these structures, such as the group lasso (Yuan and Lin, 2006) and several concave group selection methods (e.g. group SCAD (Wang et al., 2007) and group MCP. By simultaneously shrinking entire groups of variables towards zero, group penalised models attempt to take advantage of this structure and produce a

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

sparse model. The interested reader can refer to Huang et al. (2012) for an exhaustive review.
In this paper, we aim to extend the adaptive non-convex penalty function introduced in Cuntrera et al. (2022a) to group penalized context, evaluating the performance against the principal competitors by a numerical study.

2 Method

Although the setting of the problem may look different, the approach is the same as the well-known statistical modelling problem with a penalty function. The problem is to optimize a penalized objective function, such that

$$\hat{\beta} = \arg\min_{\beta} \ \mathcal{L}(\beta) + p_\lambda(|\beta|), \tag{1}$$

where $\mathcal{L}(\beta)$ is the loss function, and $p_\lambda(|\beta|)$ is the penalty function, indexed by the tuning parameter λ that controls the sparsity of the vector β (the higher the value, the sparser the vector of coefficients). Regarding the first term we can consider the general regression loss function $\mathcal{L} = \left\| y - \sum_{j=1}^{J} X_j \beta_j \right\|_2^2$, where y is an $n \times 1$ vector of the centred response variable, X_j is an $n \times p_j$ matrix of factor (i.e. groups of variables) and β_j is the p_j-size coefficient vector, where $j = 1, ..., J$. Equation (1) allows for either categorical or continuous factors. Notice that the ANOVA model is a typical case where all factors are categorical, whereas the additive model is one where all factors are continuous. However, it is possible to simultaneously include both categorical and continuous factors in the equation. The penalized model is also easily extended to the context of generalised linear models (as will be seen below).
The difference in the penalisation of ungrouped variables lies in the specification of the penalty function. Instead of penalizing the individual regression coefficient, the penalty is based on the L_2 norm of the J different group of coefficients. Thus, for a given value of the tuning parameter λ, if a j-th group of coefficients is selected, all the coefficients within it will differ from 0. Generalising the penalty function introduced in Cuntrera et al. (2022a), the group penalty in (1) is

$$p_\lambda(|\beta|) = \lambda \sum_{j=1}^{J} \sqrt{2p_j \pi} \nu \Phi \left(\frac{\|\beta_j\|_2}{\nu} \right). \tag{2}$$

We call it as group Adaptive Non-Convex penalty function (group ANP), where ν is an additional shape parameter. The selection of the additional parameter ν is crucial since it determines the degree of bias of the non-null estimates and has implications for both computational and inferential aspects. Furthermore, the convergence rate of the estimated coefficients to

the maximum likelihood estimates of the non-zero coefficients is also influenced by the choice of ν: the lower ν, the faster the convergence rate. As ν goes to infinity, the penalty function takes the form of the lasso penalty. On the other hand, selecting a small value of ν results in a highly non-convex penalty function, which can behave local optima in the objective function. To ensure the uniqueness of the solution, it is possible to find a lower bound for ν, denoted by $\nu_{\lambda,min}$, for any given λ value. In this way, the solution enjoys the oracle property defined by Fan and Li (2001), avoiding computational issues. The interested reader can find details of the proposed penalty function in Cuntrera et al. (2022b). Finding the solution to the objective function is not trivial. We propose using the Alternating Direction Multiplier Method (ADMM), an algorithm that decomposes complex optimisation problems into smaller, more manageable problems (Boyd et al., 2011).

Let us consider the standard local quadratic approximation of the log-likelihood function (McCullagh and Nelder, 1983). More precisely, assume that $\hat{\beta}^t$ is an appropriate initial point and define $\eta_i^t = x_i^\top \hat{\beta}^t$. Then, we can approximate the minimizer of the objective function (1) using (2) as:

$$\hat{\beta}^{t+1} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} V_i^t (y_i^t - x_i^\top \beta)^2 - \lambda \sqrt{2p_j \pi} \nu \sum_{j=1}^{J} \Phi\left(\frac{\|\beta_j\|_2}{\nu}\right),$$

where $V_i^t = V(\eta_i^t)$ and $y_i^t = \eta_i^t + \{y_i - \mu(\eta_i^t)\}/V_i^t$ denotes the ith working response of the IWLS algorithm. The above approximation shows that estimating $\hat{\beta}$ involves solving a series of penalised weighted least squares regression problems, which can be efficiently solved using the ADMM algorithm. The problem results in the following linear equality-constrained problem:

$$\begin{aligned} \min_{\beta,\tilde{\beta}} \quad & f(\beta) + g(\tilde{\beta}) \\ \text{s.t.} \quad & \beta - \tilde{\beta} = 0, \end{aligned}$$

where:

$$f(\beta) = \frac{1}{n} \sum_{i=1}^{n} V_i^t (y_i^t - x_i^\top \beta)^2, \quad \text{and} \quad g(\tilde{\beta}) = -\lambda \sqrt{2p_j \pi} \nu \sum_{j=1}^{J} \Phi\left(\frac{\left\|\tilde{\beta}_j\right\|_2}{\nu}\right).$$

3 Numerical study

We present the results of a numerical study aimed to evaluate the performance of the extended adaptive non-convex penalty function in a group penalised context compared to the main competitor, i.e group lasso, group SCAD and group MCP. We perform the simulation following one of the settings used by Yuan and Lin (2006): we simulated 15 latent variables, $Z_1, ..., Z_{15}$, following a centred multivariate normal distribution with

Toeplitz correlation matrix $\Sigma_{jk} = 0.5^{|j-k|}$. We then trichotomized each Z_j into three categories (0, 1, or 2), depending on whether it was less than $\Phi^{-1}(1/3)$, greater than $\Phi^{-1}(2/3)$, or in between. Finally, we simulated the response variable y from the following equation

$$y_i = 1.8I(z_{1,i} = 1) - 1.2I(z_{1,i} = 0) + I(z_{3,i} = 1) + 0.5I(z_{3,i} = 0) + \\ + I(z_{5,i} = 1) + I(z_{5,i} = 0) + \epsilon_i, \qquad i = 1, ..., n,$$

where $I(\cdot)$ is the indicator function, $\epsilon_i \sim N(0, 1)$ and $n = 50$. For each of the 100 simulations, we calculated the average model error $ME(\hat{\beta}) = (\hat{\beta} - \beta)'E(X'X)(\hat{\beta} - \beta)$ for all the λ-values used, along with the AUC.

TABLE 1. Simulation results: averages and standard deviations in parenthesis of model error and AUC.

	Group ANP	Group lasso	Group SCAD	Group MCP
AUC	0.930	0.922	0.928	0.918
	(0.07)	(0.08)	(0.07)	(0.09)
Model error	0.441	0.614	0.654	0.650
	(0.08)	(0.10)	(0.11)	(0.12)

Table 1 shows the results of the simulation. Looking at the AUC, all the estimators have roughly the same performance. In terms of model error, however, the differences are greater: the group ANP has a lower error than its competitors, around 30% lower. Thus, the estimator is able to identify the correct subset of groups of variables not equal to 0 slightly better than its competitors, committing the smallest error (along the entire λ path).

4 Conclusion

In this paper, we have extended the adaptive non-convex penalty function to the group penalized setting, which allows for the simultaneous selection and estimation of group coefficients in high-dimensional data with complex structures. We evaluated the performance of our method against the main competitors, in a numerical study using simulated data. Our results show that our proposed method outperforms the competitors in terms of model error, indicating its superiority in selecting important groups of variables with the lowest bias w.r.t. the true coefficients.

Acknowledgments: Luigi Augugliaro and Vito M.R. Muggeo gratefully acknowledge financial support from the University of Palermo (FFR2021-22)

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3(1)**, 1 – 122

Cuntrera, D., Muggeo, V. M. R. and Augugliaro, L. (2022a). Variable selection with unbiased estimation: the CDF penalty. *51th Scientific Meeting of the Italian Statistical Society: Book of Short Papers*, 1835–1840

Cuntrera, D., Augugliaro, L., and Muggeo, V. M. R. (2022b). The CDF penalty: sparse and quasi unbiased estimation in regression models. *arXiv preprint arXiv:2212.08582.*

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* **96(456)**, 1348 – 1360

Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Stat Sci*, **27(4)**, 481 – 499

McCullagh, P. and Nelder, J. A. (1983). Generalized linear models. *Routledge*

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B*, **68(1)**, 49 – 67.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinform*, **23(12)**, 1486 – 1494.

Gradient boosting for GAMLSS using adaptive step lengths

Alexandra Daub[1], Andreas Mayr[2], Boyao Zhang[1], and Elisabeth Bergherr[1]

[1] University of Goettingen, Germany
[2] University of Bonn, Germany

E-mail for correspondence: alexandra.daub@uni-goettingen.de

Abstract: Estimating generalized additive models for location, scale and shape by means of a non-cyclical gradient-based boosting algorithm with fixed step lengths can result in imbalanced submodel updates and long run times. Shrunk optimal step lengths have been shown to solve these issues. We propose a new way for obtaining adaptive step lengths based on algorithm intrinsic information and implement a boosting algorithm for GAMLSS with the different step length options for normal, negative binomial and Weibull distributed response variables. We show in a simulation study that the new adaptive step length approach yields similar results as shrunk optimal step lengths while reducing the run time. Additionally, the algorithm is applied to model mean and overdispersion of the number of doctor's visits using data from the Australian Health Survey.

Keywords: Step Lengths; Gradient Boosting; GAMLSS.

1 Boosting GAMLSS

In order to benefit from the known advantages of machine learning methods, component-wise boosting algorithms are used for estimating statistical models, inter alia generalized additive models for location, scale and shape (GAMLSS). GAMLSS are specified by

$$g_k(\theta_k) = \eta_{\theta_k} = \beta_{0\theta_k} + \sum_{j=1}^{J_k} f_{j\theta_k}(x_{kj}) \,, \;\; k \in \{1, ..., K\} \,,$$

where $x_{k1}, ..., x_{kJ_k}$ are the J_k covariates for modelling the predictor η_{θ_k}, which is linked to the corresponding distributional parameter θ_k via the

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

known monotonic function $g_k(\cdot)$. $\beta_{0\theta_k}$ refers to the intercept of the kth submodel and $f_{j\theta_k}$ represents the assumed type of effect between covariate j and predictor η_{θ_k}. In the following, only linear effects $f_{j\theta_k}(x_{kj}) = x_{kj}\beta_{kj}$ will be considered.

While GAMLSS are typically estimated using penalized maximum likelihood, component-wise gradient boosting (Bühlmann and Yu, 2003) has the advantages of an intrinsic variable selection and a regularization of model coefficients. Due to the model structure, boosting GAMLSS is however not completely straight forward. Modifying the originally "cyclical" update structure for boosting GAMLSS (Mayr et al., 2012), Thomas et al. (2018) propose to update only the predictor η_{θ_k} whose update yields the largest improvement with respect to the loss function. Zhang et al. (2022) extend the algorithm by adaptive step lengths that ensure a natural balance in the updates of the different submodels and often need fewer iterations until stopping. They define the adaptive step length $\nu_{\theta_k}^{[m]}$ as shrunk optimal step length, i.e.

$$\nu_{\theta_k}^{[m]} = 0.1 \cdot \nu_{\theta_k}^{*[m]}, \text{ where } \nu_{\theta_k}^{*[m]} = \arg\min_{\nu_{\theta_k}} \rho\left(\boldsymbol{\eta}_{\theta_k}^{[m-1]} + \nu_{\theta_k} \boldsymbol{h}_{j*,\theta_k}^{[m]}\right). \quad (1)$$

$\boldsymbol{\eta}_{\theta_k}^{[m-1]}$ represents the predictor corresponding to θ_k after $m-1$ iterations and $\boldsymbol{h}_{j*,\theta_k}^{[m]}$ is the best-performing base-learner for updating $\boldsymbol{\eta}_{\theta_k}^{[m-1]}$. For solving this optimization problem, Zhang et al. (2022) implement a numerical optimization via line search and derived analytical results for the special case of a Gaussian response variable with simple linear base-learners.

In order to generalize this concept to other distributions, an optimal step length would either have to be derived for every submodel of every response variable separately, where in many cases an analytical closed form solution does not exist, or a line search could be used, which is however computationally more demanding. We therefore propose a third option to construct adaptive step lengths based on algorithm intrinsic information.

2 Alternative Construction of Adaptive Step Lengths Based on the Ratio of Base-Learner Norms

The new construction of an adaptive step length aims at GAMLSS that have a response variable with more than one distributional parameter. In particular, we propose to define the adaptive step length for updating $\eta_{\theta_l}^{[m-1]}, l \in \{1, ..., K\} \setminus \{k\}$ in iteration m as the step length of a reference parameter θ_k rescaled by the ratio of norms of the respective fitted base-learners, i.e.

$$\nu_{\theta_l}^{[m]} := \nu_{\theta_k}^{[m]} \frac{\|\boldsymbol{h}_{j_k^*,\theta_k}^{[m]}\|_2^2}{\|\boldsymbol{h}_{j_l^*,\theta_l}^{[m]}\|_2^2}, \quad (2)$$

where $\|x\|_2$ denotes the Euclidean norm of $x \in \mathbb{R}^n$.
This construction and especially the choice of the squared Euclidean norm as measure for the length of a vector is motivated by the observation that the ratios of optimal step lengths behave approximately inversely to the ratios of the Euclidean base-learner norms, i.e.

$$\frac{\nu_{\theta_l}^{[m]}}{\nu_{\theta_k}^{[m]}} \approx \frac{\|\boldsymbol{h}_{j_k^*,\theta_k}^{[m]}\|_2^2}{\|\boldsymbol{h}_{j_l^*,\theta_l}^{[m]}\|_2^2}\,.$$

Fig. 1 below displays both ratios for an exemplary simulation run of a negative binomial response variable with shrunk optimal step lengths.

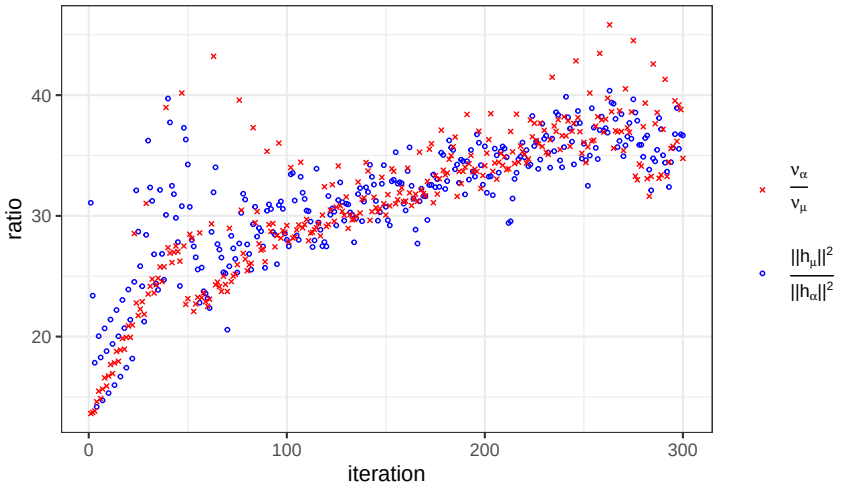


FIGURE 1. Inverse ratio of shrunk optimal step lengths and ratio of Euclidean base-learner norms over the iterations of an exemplary simulation run with a negative binomial response variable.

Defining the step size of a potential update by $\zeta_{\theta_k}^{[m]} := \nu_{\theta_k}^{[m]} \|\boldsymbol{h}_{j_k^*,\theta_k}^{[m]}\|_2^2$, the approximately inverse relationship of base-learner norms and step lengths can be rewritten as

$$\nu_{\theta_k}^{[m]} \|\boldsymbol{h}_{j_k^*,\theta_k}^{[m]}\|_2^2 = \zeta_{\theta_k}^{[m]} \approx \zeta_{\theta_l}^{[m]} = \nu_{\theta_l}^{[m]} \|\boldsymbol{h}_{j_l^*,\theta_l}^{[m]}\|_2^2\,,$$

where $\nu_{\theta_k}^{[m]}$ and $\nu_{\theta_l}^{[m]}$ are specified as in eq. 1. The step length definition in eq. 2 thus effectively fixes the step sizes of all updates to the step size of the reference parameter.
Base-learner ratio-based adaptive step lengths cannot stand on their own but a differently obtained reference step length is always necessary. In order to implement adaptive step lengths for negative binomial and Weibull

distributed response variables, we derive approximations of the optimal step lengths of one parameter per distribution analytically, in the negative binomial case, e.g., μ serves as reference parameter.

3 Simulation Study

In order to show that using the base-learner ratio-based approximation yields comparable results as shrunk optimal step lengths, we consider the following model with a negative binomial response variable $y_i \sim NB(\mu_i, \alpha_i)$ exemplarily:

$$\eta_{\mu,i} = \log(\mu_i) = 0.2 - 0.25\, x_{1i} + 0.2\, x_{2i} - 0.15\, x_{4i} + 0.2 x_{5i}$$
$$\eta_{\alpha,i} = \log(\alpha_i) = -1.5 - 0.25\, x_{2i} + 0.2\, x_{3i} - 0.1\, x_{5i} + 0.15\, x_{6i}$$

Note that in this model formulation x_{2i} and x_{5i} are shared between both predictors and that y_i has a variance of $\mu_i + \alpha_i \mu_i^2$. We simulate $n = 500$ observations of each variable, where x_{1i}, x_{2i}, x_{3i} are drawn independently from a uniform distribution on $[-1, 1]$ and for x_{4i}, x_{5i}, x_{6i} independent realizations of a Bernoulli distributed random variable with $p = 0.6$ are drawn. Except for the inclusion of binary covariates and different coefficients in order to have a setup with differing optimal step lengths, this simulation setup follows Thomas et al. (2018).
We apply a non-cyclical boosting algorithm with two adaptive step length approaches with analytically derived ν_μ, where the step length for an update of α is either obtained by line search ('LS') or based on the base-learner

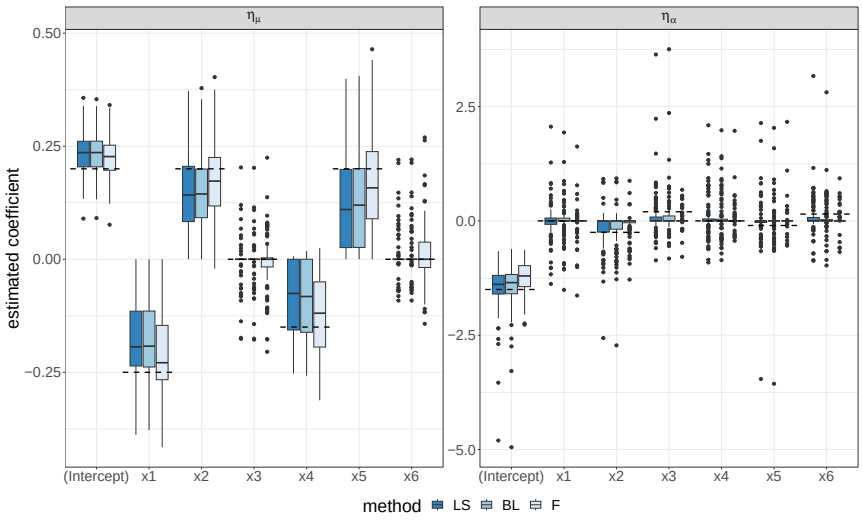


FIGURE 2. Coefficient estimates of the non-cyclical boosting algorithm with different step length approaches at the stopping iteration.

ratio ('BL'). Additionally, a fixed step length of 1 for both submodel updates is considered ('F'). The maximum number of iterations is set to 1,000 for the adaptive step length approaches and to 7,000 when using fixed step lengths. Fig. 2 displays the coefficient estimates at the stopping iteration obtained via 10-fold cross-validation for $B = 100$ simulation runs.

The simulation results show that the adaptive step length approaches yield similar estimated coefficients, while using fixed step lengths results in higher coefficient estimates in η_μ and lower coefficient estimates in η_α in absolute values. The differences in the resulting coefficients go along with a larger number of false positives in η_μ and false negatives in η_α when using fixed step lengths, where in 46 of the runs η_α is not updated at all. While overall being very similar to the numerically obtained shrunk optimal step lengths, the base-learner ratio-based results seem to exhibit a slight tendency towards the fixed step length approach.

The outlined differences between the adaptive and fixed step length approaches originate from the problem that the negative gradient vector sizes of the different parameters differ substantially in this model setup, which carries over to the size of the fitted base-learners. While in adaptive step length approaches this is compensated by the step lengths, with fixed step lengths the relation of the base-learner sizes is passed onto the predictor updates favoring updates of the predictor with the larger base-learners. The mean squared prediction errors on test data as well as the amount of false positives and false negatives indicate that this disbalance between submodel updates leads to an overfitting of μ and underfitting of α in the present case.

The simulation results moreover show that the intended reduction in run time could be achieved. The median run time of the base-learner ratio-based approach is about 45% lower than computing optimal step lengths numerically in this setup, while for the fixed step length approach the median run time until stopping is about 80% higher due to later stopping. Simulations with 10 and 150 additional non-informative variables show similar overall results.

4 Modelling the Number of Doctor's Visits

In the following, we apply the non-cyclical boosting algorithm with two adaptive ('LS' and 'BL') and a fixed ('F') step length scheme to a negative binomial location and scale model for the number of doctor's visits in Australia. The data was collected within the Australian Health Survey in 1977-1987. In addition to the number of doctor's visits, it comprises information on different characteristics of the individuals like the type of health insurance (e.g., 'levyplus' or 'freepoor') and variables referring to the health condition (e.g.,'chcond1') or recent treatments (e.g., 'medicine', 'prescrib') of 5,190 individuals. For further information on the variables, we refer to Cameron and Trivedi (1986) who compiled the data.

TABLE 1. Comparison of the coefficient estimates using different step length schemes.

		LS	BL	F		LS	BL	F
η_μ	(Intercept)	-2.012	-2.054	-2.094	freepoor	-0.184	-0.264	-0.363
η_α		0.476	0.339	0.264		0	0	0
η_μ	sex	0.111	0.128	0.147	chcond1	0	0	0.009
η_α		0	0	0		0	0	0
η_μ	income	0	-0.003	-0.010	medicine	0	0	-0.011
η_α		0	0	0		-0.056	-0.087	-0.113
η_μ	levyplus	0	0	0	prescrib	0.153	0.165	0.177
η_α		-0.080	-0.053	-0.030		-0.241	-0.139	-0.078

Table 1 displays selected coefficient estimates at the stopping iteration determined by 10-fold cross validation. The results show that with respect to η_μ the two adaptive step length approaches select the same variables except for the variable 'income', which has an estimated coefficient of -0.003 in the base-learner ratio-based approach. Also, size and sign of the coefficient estimates are similar. The fixed step length approach on the other hand selects 'income', 'chcond1' and 'medicine' additionally and results in higher coefficient estimates for all selected covariates. For η_α, the same variables are selected in all three cases with mostly similar coefficient estimates.
The mean squared error on a held-out part of the data set is 0.907, 0.941 and 1.029 for numerically obtained, base-learner ratio based and fixed step lengths respectively. The results in this data example are thus in line with the findings in the simulation study.

References

Bühlmann, P. and Yu, B. (2003). Boosting With the L2 Loss. *Journal of the American Statistical Association*, **98**, 324 – 339.

Cameron, A.C. and Trivedi, P.K. (1986). Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests. *Journal of Applied Econometrics*, **1**, 29 – 53.

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.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., and Hofner, B. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, **28**, 673 – 687.

Zhang, B., Hepp, T., Greven, S., and Bergherr, E. (2022). Adaptive step-length selection in gradient boosting for Gaussian location and scale models. *Computational Statistics*, **37**, 2295 – 2332.

Mixture confidence sequences for regression coefficients in generalized linear models

Claudia Di Caterina[1], Alessandra Salvan[2], Nicola Sartori[2]

[1] University of Verona, Italy
[2] University of Padova, Italy

E-mail for correspondence: claudia.dicaterina@univr.it

Abstract: Standard fixed-n confidence intervals produced at different sample sizes on accumulating data can be contradictory with high probability. We show a simple way to compute, instead, mixture confidence sequences for regression coefficients in generalized linear models. Simulations attest the need of using these always-valid inferences if more observations become available over time.

Keywords: Anytime-valid inference; Logit link; Probit link; Streaming data.

1 Introduction

Streaming datasets, i.e. datasets becoming available sequentially over time, are the rule rather than the exception in the *big data* era. Examples include observations from long-term clinical trials and online A/B tests. In this context, delivering safe inferences which are valid at any and all times of analysis appears especially crucial. While the coverage of usual confidence intervals (CIs) is guaranteed for a fixed sample size n, mixture confidence sequences (CSs) (Robbins, 1970) meet the coverage requirements under arbitrary enlargement of the sample (see, e.g., Howard et al., 2021; Ramdas et al., 2022, § 2.7). We extend here the use of the approximate CSs proposed in Pace et al. (2022) for scalar parameters of interest to the framework of generalized linear models (GLMs). This extension is based on the sequential update and the asymptotic normality of the renewable estimator by Luo and Song (2020) for GLMs regression coefficients.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Renewable estimation in generalized linear models

Luo and Song (2020) assume that data arrive in batches, and each batch $D_b = \{y_b, X_b\}$ $(b = 1, \ldots, B)$ with n_b independent observations (y_i, x_i) can be modeled by the same exponential dispersion family model $f(y_i; x_i, \beta, \phi)$. Covariates are in the single-batch $n_b \times p$ matrix X_b and the single-unit p-vector x_i, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the coefficients vector of interest and $\phi > 0$ is the dispersion parameter. The goal is to fit sequentially a GLM on the first b aggregated batches, i.e. estimate $\mu_i = E(y_i|x_i) = g(x_i^{\mathrm{T}}\beta)$ with $i = 1, \ldots, N_b = \sum_{j=1}^{b} n_j$ for some known link function $g(\cdot)$.
Since there is no closed form for the maximum likelihood (ML) estimator $\hat{\beta}_b^*$, it is not possible to update it sequentially based solely on the current data D_b and the previous ML estimate $\hat{\beta}_{b-1}^*$. Denote by $U_b(D_b; \beta) = \sum_{i \in D_b} \nabla_\beta \log f(y_i; x_i, \beta, \phi)$ and by $J_b(D_b; \beta) = -\nabla_\beta U_b(D_b; \beta)$ the bth batch score function and observed information, respectively. Let also the aggregated observed information be $\tilde{J}_b = \sum_{j=1}^{b} J_j(D_j; \tilde{\beta}_j)$. An incremental updating algorithm for $\tilde{\beta}_b$ (Luo and Song, 2020, eq. (11)) is

$$\tilde{\beta}_b^{(r+1)} = \tilde{\beta}_b^{(r)} + \{\tilde{J}_{b-1} + J_b(D_b; \tilde{\beta}_{b-1})\}^{-1} \tilde{U}_b^{(r)}, \tag{1}$$

where the adjusted score $\tilde{U}_b^{(r)} = \tilde{J}_{b-1}(\tilde{\beta}_{b-1} - \tilde{\beta}_b^{(r)}) + U_b(D_b; \tilde{\beta}_b^{(r)})$ is updated at each iteration. Implementing (1) requires just the availability of current data and summary statistics $\{\tilde{\beta}_{b-1}, \tilde{J}_{b-1}\}$ of previous data. The renewable estimator $\tilde{\beta}_b$ is consistent and asymptotically normal as $N_b \to \infty$, with estimated covariance matrix $\tilde{V} = \tilde{\phi}_b \tilde{J}_b^{-1}$. The estimator $\tilde{\phi}_b$ can be easily updated iteratively based on the previous $\tilde{\phi}_{b-1}$ (Luo and Song, 2020, p. 77).

3 Mixture confidence sequences

If inference is conducted as data batches are collected, it is important to limit the chance that contradictory conclusions are reached along the way. This translates into requiring that plausible regions for the parameters constructed at different sample sizes are compatible, i.e. overlap with high enough probability (Pace and Salvan, 2020).
Let the potentially observable data $y^{(n)}$ be realization of the random vector $Y^{(n)} = (Y_1, \ldots, Y_n)$ $(n = 1, 2, \ldots)$. We denote by $p_n(y^{(n)}; \theta)$ the density of $Y^{(n)}$ with support independent of $\theta \in \Theta \subseteq \mathbb{R}^q$. A sequence of estimation regions $\hat{\Theta}_n = \hat{\Theta}(y^{(n)}) \subseteq \Theta$ is a CS. The latter has persistence level $1 - \varepsilon$ if

$$\forall \theta \in \Theta \quad P_\theta \left(\theta \in \cap_{n \geq 1} \hat{\Theta}_n \right) \geq 1 - \varepsilon\,, \quad 0 < \varepsilon < 1.$$

This implies that the probability of observing incompatible conclusions when the sample enlarges is smaller than ε (Pace et al., 2022). Ville's

inequality (Ramdas et al., 2022, § 2.4) is the starting point to obtain confidence sequences of persistence level $1-\varepsilon$. Among those, mixture CSs (Robbins, 1970) have the form

$$\hat{\Theta}_{1-\varepsilon}(y^{(n)}) = \left\{ \theta \in \Theta \,:\, p_n(y^{(n)};\theta) > \varepsilon \int_\Theta p_n(y^{(n)};\theta)\pi(\theta)\,d\theta \right\},$$

where the weight function $\pi(\theta)$ is a preset probability density over Θ. Pace et al. (2022) consider the partition $\theta = (\psi, \lambda)$, with $\psi \in \Psi \subseteq \mathbb{R}^{p_0}$ component of interest and λ nuisance. Mixture CSs for ψ, based on the corresponding profile likelihood with $\hat{\lambda}_\psi$ ML estimate of λ for ψ fixed, are

$$\hat{\Psi}_{1-\varepsilon}(y^{(n)}) = \left\{ \psi \in \mathbb{R} \,:\, p_n(y^{(n)};\psi,\hat{\lambda}_\psi) > \varepsilon \int_\Theta p_n(y^{(n)};\theta)\pi(\theta)\,d\theta \right\}.$$

4 Monte Carlo experiments

We focus on binary observations with probability of success μ_i generated by logit and probit link functions. The p covariates include the intercept and both discrete and continuous variables, simulated by the R package `RenewGLM` (Luo and Song, 2019) as exchangeable with correlation $\rho = 0.5$. Consider 10^4 Monte Carlo sequences of samples with size n ranging from $n_{\min} = 150$ to $n_{\max} = 10^5 + n_{\min}$. Between these two sample sizes, data are observed in B batches of equal size n_b. Each time the bth batch comes in, we compute the renewable estimator in (1) for the assumed GLM and obtain the approximate confidence sequence of persistence level $1-\varepsilon$ for β_p. Thanks to the asymptotic normality of $\tilde{\beta}_{b,p}$, according to Pace et al. (2022, eq. (17)), with a $N(\beta_{0,p}, \tau_0^2)$ weight function we get

$$\tilde{\beta}_{b,p} \pm \sqrt{\tilde{V}_{pp} \log \frac{\tau_0^2 + \tilde{V}_{pp}}{\tilde{V}_{pp}} + \frac{(\tilde{\beta}_{b,p} - \beta_{0,p})^2}{\tau_0^2 + \tilde{V}_{pp}} - 2\log\varepsilon},$$

where $\tilde{\beta}_{b,p}$ and $\tilde{V}_{pp}$ are the pth element of $\tilde{\beta}_b$ and the (p,p)th entry of $\tilde{V}$, respectively, obtained as in Section 2 by adapting the code in `RenewGLM` to deal also with non-canonical links. The true value of β_p is set equal to 1.2 for logit (1.2/1.6 for probit), while $\beta_{0,p} = 0$ and $\tau_0 = 1$. For various p and n_b, Table 1 compares for logit regression $(1-\alpha)$-Wald CIs ($\alpha = 0.05, 0.01$) and $(1-\varepsilon)$-mixture CSs ($\varepsilon = 0.20, 0.05$) in terms of incompatibilities and uncoverages. The former are given by sequences with incompatible intervals, the latter by those not always covering the true parameter value. Mixture CSs ensure non-contradictory and reliable conclusions on β_p: except for one case, all incompatibilities and uncoverages are well below the nominal threshold ε. Those for Wald CIs, instead, always exceed α as they are generally shorter: for instance, if $p = 10$ and $n_b = 20$ the average length is 15.52 for 0.80-mixture CSs and 11.69 for 0.99-Wald CIs. Unshown probit results are in line. The optimization of mixture CSs for a specific sample size n^*, in the style of Howard et al. (2021, § 3.5), is left for future work.

TABLE 1. Empirical % of incompatibilities and uncoverages for the true $\beta_p = 1.2$ in the logit GLM for data in batches of size n_b with p covariates. The weight function used for mixture CSs is $N(0, 1)$. Results are based on 10^4 Monte Carlo sequences from $n_{\min} = 150$ to $n_{\max} = 10^5 + n_{\min}$.

			Wald CIs		Mixture CSs	
p	property (%)	n_b	$\alpha = 5\%$	$\alpha = 1\%$	$\varepsilon = 20\%$	$\varepsilon = 5\%$
5	incompatibilities	20	29.10	6.04	1.96	0.36
		100	22.90	4.50	1.06	0.18
	uncoverages	20	56.72	19.14	4.78	1.28
		100	50.94	15.86	3.72	1.02
10	incompatibilities	20	30.66	5.84	1.90	0.38
		100	23.46	3.86	0.82	0.22
	uncoverages	20	56.56	19.96	4.94	1.42
		100	51.42	15.82	3.14	0.70
20	incompatibilities	20	35.46	8.00	2.58	0.80
		100	38.52	18.18	11.48	8.36
	uncoverages	20	60.72	22.16	7.12	2.46
		100	62.08	29.92	14.98	9.98

References

Howard, S.R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, **49**, 1055 – 1080.

Luo, L., and Song, P. X.-K. (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *Journal of the Royal Statistical Society B*, **82**, 69 – 97.

Luo, L., and Song, P. X.-K. (2019). RenewGLM: Renewable Estimation and Incremental Inference in Generalized Linear Models with Streaming Datasets. https://github.com/luolsph/RenewGLM_pkg.

Pace, L. and Salvan, A. (2020). Likelihood, replicability and Robbins' confidence sequences. *International Statistical Review*, **88**, 599 – 615.

Pace, L., Salvan, A., and Sartori, N. (2022). Confidence sequences with composite likelihoods. *Canadian Journal of Statistics.* 10.1002/cjs.11749.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2022). Game-theore-tic statistics and safe anytime-valid inference. arXiv:2210.01948.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, **41**, 1397 – 1409.

On the nature of one–inflation in microbial diversity studies

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

[1] Sapienza University of Rome, Italy

E-mail for correspondence: davide.dicecco@uniroma1.it

Abstract: The phenomenon of one–inflation is gaining more and more attention in the recent literature on species abundance and capture–recapture analysis. When analysing frequency count distribution, the excess of singletons is often ascribed to the erroneous inclusion of spurious cases. Various works propose to estimate the true number of singletons relying on the higher, supposedly error–free, counts ("discounting" approach). We argument that, in the case of microbial diversity studies, the generating process of the spurious singletons can be described in terms of false negative record linkage errors. Errors in sequencing the RNA genomes result in chimeric sequences that cannot be associated to the correct species, and constitute missing links that are added to the true singletons. In this scenario, none of the observed frequency counts is assumed to be error–free, and we propose an ABC algorithm to estimate the true frequency counts. The number of true singletons estimated in this way may differ considerably from the discounting approach. This implies different estimates of the diversity as measured, e.g., by Shannon's index. However, curiously, the total population count estimates under the two approaches coincide.

Keywords: Species problem; Biodiversity; Linkage Errors; Approximate Bayesian Computing.

1 Introduction

The problem of estimating the number of species in a population given a sample arises in many applications in the natural sciences, in linguistics and computer science. Our focus is on applications in microbial ecology. The spread of next generation high-throughput sequencing technology allowed to analyse an unprecedented amount of data on microbial communities. In order to study the biodiversity in a microbial community, an environmental sample is processed to detect, amplify and sequence RNA genomes. The

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

sequences are clustered into distinct species (or Operational Taxonomic Units) on the basis of a similarity score. The diversity analysis is then conducted on the abundance frequency counts, i.e., the counts $\{n_j\}_{j=1,2,...}$ representing the number of species with j captured occurrences. In most microbial studies, the distribution $\{n_j\}_{j=1,2,...}$ is characterized by an unexpected number of low–abundance species, in particular singletons, accompanied by a low number of very common species. The nature of these singletons has been debated at length, and the presence of spurious singletons resulting from sequencing errors has been confirmed in various ways (e.g., Quince et al. 2011, Haas et al. 2011). While bioinformatics focuses on avoiding the formation of the so–called chimera sequences, or removing them in a pre–processing step, various statistical contributions attempt to estimate ex–post their number.

The study of one–inflation in frequency count distribution is gaining more and more attention also in the recent capture–recapture literature on human and animal population, which shares many methodological aspects with the species abundance problem, (see, e.g., Godwin and Böhning 2017, Böhning et al. 2019, Tuoto et al. 2022). The possible sources of one–inflation can be categorized as:

- a behavioural effect, where certain units, once captured, avoid subsequent captures;
- the presence of out-of-scope units, which enter the sample for a peculiar error mechanism and should be excluded;
- the presence of missing links in the record linkage procedure employed to create the frequency counts.

Various authors adopted a "discounting" approach to the problem of one–inflation. That is, they propose to ignore the data affected by errors, i.e., the observed singletons, and re-estimate their number on the basis of the counts n_j, $j \geq 2$, (see, e.g, Willis and Bunge 2015, Willis 2016, Chiu and Chao 2016). We argument that this approach is consistent for the second mechanism listed above: a model where out-of-scope singletons are added to the baseline distribution of the true counts. We believe that the nature of the spurious cases can alternatively be described by linkage errors. That is, we assume that random errors occurring in sequencing result in the impossibility of a correct classification of the specimen, which cannot be associated to the right existing species. Therefore, we can describe these cases as false negative linkage errors (or missing links), which are added to the true singletons. This approach implies a re–estimation of the "real" frequency counts for all the abundances, not just the singletons. We found that treating the excess of singletons in this way leads to significant differences in the diversity estimates with respect to the discounting approach.

In this work we adopt a secondary approach to the linkage problem, i.e., we try to estimate the linkage errors solely on the basis of the vector

$\{n_j\}_{j=1,2,...}$ and our distributional assumptions, as we do not have access to the actual linkage process. Modeling linkage errors in this secondary setting, appears quite complex from a computational point of view. We fix some simplifying assumptions on the type of error in order to tackle the issue, but we still resorted to a Bayesian likelihood–free approach as the most convenient approach.

2 One–inflation models

Say we get n species in our sample with abundances $y_1, ..., y_n$, and abundance frequency counts $\{n_j\}_{j\geq 1}$. Under an out–of–scope singletons model, the distribution of the abundances (whether the species are observed or not, spurious or not) results in the following mixture of a baseline distribution $\widetilde{f}$ of the non–spurious counts, and a Dirac measure over one:

$$P(Y_i = j \,;\, \widetilde{f}, \psi) = \begin{cases} (1-\psi)\widetilde{f}_1 + \psi & \text{if } j = 1; \\ (1-\psi)\widetilde{f}_j & \text{otherwise,} \end{cases} \tag{1}$$

where ψ denotes the portion of spurious cases over the total population count. Let $\widetilde{n}_j$ denote the number of species with j non spurious captures. Then, since we assumed $\widetilde{n}_j = n_j$ for $j \geq 2$, we just have to estimate the number of unsampled species $\widetilde{n}_0$, and the number of non–spurious singletons $\widetilde{n}_1$ as a portion of n_1. The estimate of the total number of distinct species $\widetilde{N}$ will result as:

$$\sum_{j\geq 0} \widetilde{n}_j = \widetilde{n}_0 + n - n_1 + \widetilde{n}_1.$$

A Bayesian estimation of this model presents no difficulties under various parametric families choices for $\widetilde{f}$. A simple Gibbs sampler scheme is the following: under a Beta prior for ψ, its posterior is easily updated. Then, a value for $\widetilde{n}_1$ is generated from a Binomial with parameters $1-\psi$ and n_1. Steps to update the values of $\widetilde{n}_0$ and of the parameters of $\widetilde{f}$ are easily found in literature (see, e.g., Tuoto et al. 2022).

Under our missing links proposal, we assume that each sequence has the same probability μ of being missclassified as a singleton independently from the other. Denote the true number of sampled distinct species as n^*, ($n^* < n$). For each species i with X_i^* true captures, we have M_i missing links, such that the registered abundance is reduced from X_i^* to $X_i = X_i^* - M_i$. M_i has the following distribution:

$$P(M_i = m_i \mid X_i^* = x_i^*) = \binom{x_i^*}{m_i} \mu^{m_i} (1-\mu)^{x_i^* - m_i}, \quad i = 1, ..., n^*. \tag{2}$$

Let f^* be the baseline distribution of the X_i^*. The distribution of the X_i results as a thinning process where a portion μ of captures disappear. Let

n_j^* denote the true number of species with j captures, and as $N^* = \sum_{j\geq 0} n_j^*$ the total number of distinct species according to the missing links model. Unlike the spurious singletons model, in this case all values $\{n_j^*\}_{j\geq 0}$ have to be estimated, as they will be, in general, different from the observed values. Denote as θ the parameters defining f^*. We adopted an ABC rejection algorithm with the following scheme:

1. generate values for (θ, N^*) from the priors $\pi(\theta)$ and $\pi(N^*)$;
2. generate values $(n_0^*, n_1^*, n_2^*, ...)$ conditional on N^* and θ;
3. generate a value for μ from the Beta prior $\pi(\mu)$ (independent from all the rest);
4. generate missing links at random according to the distribution described in (2), given $(n_0^*, n_1^*, n_2^*, ...)$ and μ. Each missing link modifies the observed count, and increments accordingly the number of singletons, thus obtaining the fictitious data D^*;
5. retain the current generated values if a measure of distance ρ between the generated data D^* and the observed data D is below a certain threshold ϵ:
$$\rho(D^*, D) < \epsilon.$$
In our application we utilized the euclidean distance.

As the simple ABC rejection scheme can have a low acceptance rate, we further adopted a sequential ABC to accelerate the procedure, as described in Marin et al. 2012.

A simulation study confirmed the correctness of the ABC algorithm under a Poisson, Geometric, and finite mixture of Poisson distributions for f^*. Our first finding in a further simulation study comparing the spurious cases and the missing links proposal, has been the substantial identity of the estimates of the total number of species under the two competing models. That is, if we choose f^* and $\widetilde{f}$ in the same family, despite the fact that the estimates of the true abundance frequencies differ under the two models (i.e., $\widetilde{n}_j \neq n_j^*$ for all j), we have $N^* = \widetilde{N}$.

To demonstrate this identity, consider the baseline distribution f^* of the values X_i^* introduced above. It is easily demonstrated that, under various parametric family for f^*, (notably, if f^* is any mixed Poisson), the distribution of the X_i belongs to the same parametric family. Then, under identifiability of that family, if we use model (1) and take $\widetilde{f}$ in the same family as f^*, $\widetilde{f}$ will be identified as the distribution of the x_i, for all $x_i > 0$, and ψ would represent the portion of missing links over the total population count. Let r_0 be the number of captured species whose occurrences where all missclassified, i.e., such that $M_i = X_i^*$. Let M be the total number of missing links: $M = \sum_{i=1}^{n^*} M_i$. Then we have

$$n^* = n - M + r_0 \quad \text{and} \quad \widetilde{n}_1 = n_1 - M.$$

The missing links mechanism does not affect the number of undetected species n_0^*, but under $\widetilde{f}$ the r_0 values are included in $\widetilde{n}_0$, i.e., we have $\widetilde{n}_0 = n_0^* + r_0$. Finally, we can write

$$\widetilde{N} = \widetilde{n}_0 + \widetilde{n}_1 + n - n_1 = \widetilde{n}_0 + n - M \;=\; n_0^* + r_0 + n - M = n_0^* + n^* = N^*.$$

As we have said, even if the estimates of the total number of species coincides under the two models, the abundance distribution will differ, and consequently, the estimated diversity will differ. To illustrate the effect of (ignoring) a missing links mechanism on the estimation of diversity, we utilized a simulation study. As a measure of diversity we considered Shannon's diversity H (see, e.g., Chiu and Chao 2016) calculated as:

$$H = \exp\left(-\sum_{j\geq 1} n_j \frac{j}{s} \ln \frac{j}{s}\right). \tag{3}$$

We generated various datasets under Poisson and Geometric baseline distributions, then simulated the effect of missing links to simulate from our model. Then, we estimated Shannon's diversity on the observed data (that is, ignoring any one–inflation mechanism), on the "adjusted" counts as derived from the spurious cases model (that is, trimming the observed number of singletons) and as derived from the ABC procedure for the missing links model. Note that in our Bayesian approach we can easily estimate the posterior distribution of (3). First, we concluded that ignoring an existing one–inflating mechanism, implies a severe overestimation of the diversity. Second, utilizing model (1) when missing links are the true source of error, reduces sensibly the overestimation, but still leads to different results than what can be achieved with an ABC simulating the actual generating process.

References

Böhning, D., Kaskasamkul, P., van der Heijden, P. G. (2019). A modification of Chao's lower bound estimator in the case of one-inflation. *Metrika*, **82(3)**, 361 – 384.

Chiu, C. H., Chao, A. (2016). Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ*, **4**, e1634.

Godwin, R. T., Böhning, D. (2017). Estimation of the population size by using the one–inflated positive Poisson model. *Journal of the Royal Statistical Society. Series C*, 425 – 448.

Haas, B.J., Gevers, D., Earl, A.M. et al. (2011) Chimeric 16s rRNA sequence formation and detection in Sanger and 454-pyrosequenced pcr amplicons. *Genome research*, **21(3)**, 494 – 504.

Marin, J. M., Pudlo, P., Robert, C. P., Ryder, R. J (2012). Approximate Bayesian computational methods. *Statistics and computing*, **22(6)**, 1167-1180.

Quince, C., Lanzen, A., Davenport, R. J. (2011). Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, **12(1)**, 1 – 18.

Tuoto, T., Di Cecco, D., Tancredi, A. (2022). Bayesian analysis of one-inflated models for elusive population size estimation. *Biometrical Journal*, **64(5)**, 912 – 933.

Willis, A., Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics*, **71(4)**, 1042 – 1049.

Willis, A. (2016). Species richness estimation with high diversity but spurious singletons. *arXiv preprint arXiv:1604.02598*.

Prediction of record performances in sports in a record-values model

Christina Empacher[1], Udo Kamps[1]

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

E-mail for correspondence: empacher@isw.rwth-aachen.de

Abstract: In order to statistically predict a future record performance in sports based on previous record values, we exemplarily consider a data set from men's javelin throw. The assumption of a Pareto distribution as being the underlying distribution of record values is discussed. Within this model, point and interval prediction of future world records are derived and further refinements are outlined.

Keywords: Extreme-value theory; Record values; Pareto distribution; Athletics; Javelin

1 Introducing the data

Data from athletics events have been analyzed by using techniques from extreme value theory, mainly. For example, Einmahl and Magnus (2008) estimated the endpoint of the underlying distribution, which can be interpreted as an ultimate record. In their work 'How far can man go?' Fraga Alves et al. (2013) follow a similar approach when studying men's long jump using extreme-value theory. Stephenson and Tawn (2016) focus on data from running events in athletics to evaluate performances of athletes. In Empacher et al. (2023) we studied lower records in the sense of statistical record values with a focus on an underlying power function distribution applied to running events, and we predicted the very next world record in the sense of point and interval prediction.
In the present analysis, data from men's javelin throw are studied. From
https://worldathletics.org/records/all-time-toplists/throws/javelin-throw/outdoor/men/senior?regionType=world&page=1&bestResultsOnly=true&firstDay=1900-01-01&lastDay=2022-

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

`12-31`, (accessed on 3 February 2023) the all time top list of an event up to 2022 is obtained, which consists of only the best result per athlete. Since every athlete appears in the data at most once, we assume the throwing distances to be realizations of stochastically independent random variables. A histogram of the data with sample size 1625 is shown in Figure 1.

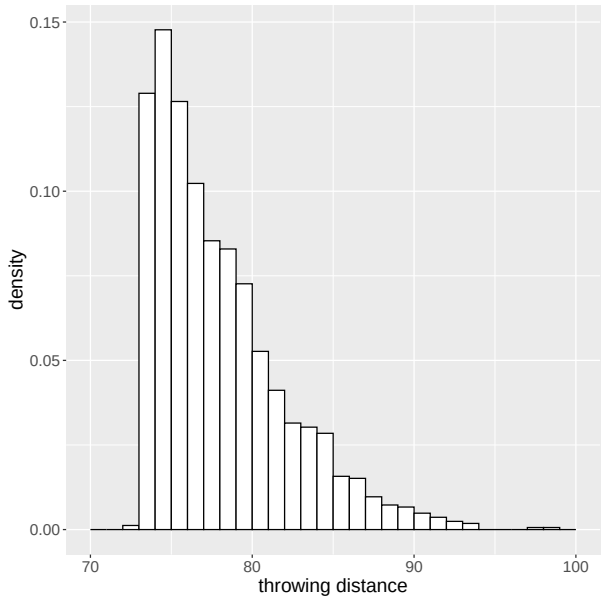


FIGURE 1. Results of men's javelin throws.

In order to further assume observations from identical distributions, we exclude data below a chosen threshold. The remaining throwing distances can be assumed to be achieved by athletes, who compete on the same professional level. In the men's javelin we choose the threshold $\lambda = 80$. Since the histogram of the respective data with sample size 420 takes the shape of a Pareto distributed sample with cumulative distribution function (cdf) and probability density function (pdf)

$$F_{\lambda,\beta}(x) = 1 - \left(\frac{\lambda}{x}\right)^{\beta}, \quad f_{\lambda,\beta}(x) = \frac{\beta\lambda^{\beta}}{x^{\beta+1}}, \quad x \in (\lambda, \infty),$$

respectively, the pdf of a Pareto distribution is fitted using the maximum likelihood estimate of the shape parameter β and the chosen λ, which will be considered a known parameter in the following study. The histogram of the throwing distances above threshold along with the graph of the fitted Pareto pdf can be found in Figure 2.

In order to evaluate the Pareto assumption, we further consider the quantile-quantile plot in Figure 3.

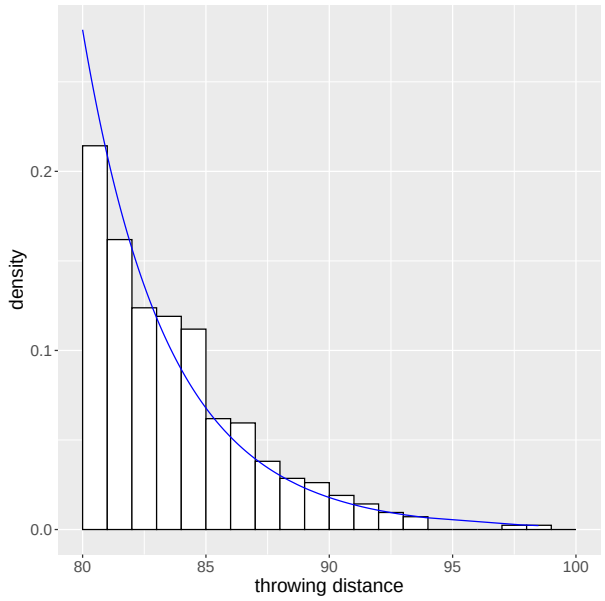


FIGURE 2. Results of men's javelin throws exceeding 80 m and pdf of the fitted Pareto distribution.

Although the density plot shows a good fit to the data, the Q-Q plot gives rise to question the underlying distributional assumption with respect to its tail behavior. In the present analysis, we keep the Pareto assumption, in which case explicit point and interval predictions of future record values can be derived.

2 Record values

Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of iid random variables with absolutely continuous cdf F and pdf f. The random variables

$$T(1) = 1, \quad T(n+1) = \min\{j > T(n) : X_j > X_{T(n)}\}, \quad n \in \mathbb{N}$$

are called upper record times and the quantities

$$R_n = X_{T(n)}, \quad n \in \mathbb{N}$$

are called upper record values (cf. Arnold et al. (1998) and Nevzorov (2001)).
The sequence of record values in men's javelin can be found at
`https://worldathletics.org/records/by-progression/17622?type=1`
(accessed on 3 February 2023). In order to predict the very next (future)

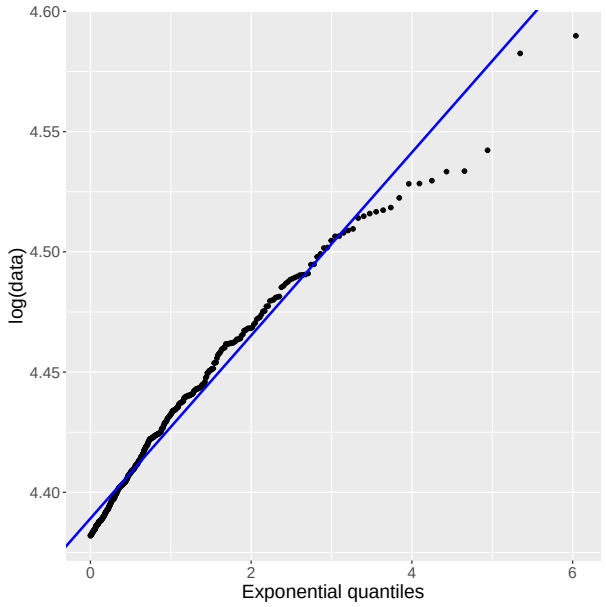


FIGURE 3. Pareto Q-Q plot for the results of men's javelin throws exceeding 80 m.

record based on r previous record values, we use the so called maximum product of spacings predictor (MPSP). According to Volovskiy and Kamps (2020) the MPSP of the $(r+1)$th record in the case of upper records values from a Pareto distribution is given by

$$\pi_{MPSP}^{(r+1)} = \lambda \left(\frac{R_r}{\lambda} \right)^{\frac{r+1}{r}}.$$

Furthermore, for $\alpha \in (0,1)$, exact and approximate $(1-\alpha)$-prediction intervals are shown in Empacher et al. (2023):

$$PI_1 = \left[R_1 \exp\left(\frac{\ln(R_r) - \ln(R_1)}{(1-\frac{\alpha}{2})^{\frac{1}{r-1}}} \right),\ R_1 \exp\left(\frac{\ln(R_r) - \ln(R_1)}{(\frac{\alpha}{2})^{\frac{1}{r-1}}} \right) \right],$$

$$PI_2 = \left[\lambda \exp\left(\frac{\ln(R_r) - \ln(\lambda)}{(1-\frac{\alpha}{2})^{\frac{1}{r}}} \right),\ \lambda \exp\left(\frac{\ln(R_r) - \ln(\lambda)}{(\frac{\alpha}{2})^{\frac{1}{r}}} \right) \right],$$

$$PI_3 = \left[R_r \left(1-\frac{\alpha}{2}\right)^{-\frac{\ln(R_r)-\ln(\lambda)}{r}},\ R_r \left(\frac{\alpha}{2}\right)^{-\frac{\ln(R_r)-\ln(\lambda)}{r}} \right].$$

While PI_1 and PI_2 are exact prediction intervals obtained by means of Pivot-statistics, the interval PI_3 is an approximate one, which is constructed by plugging in the maximum likelihood estimator of β based on

record values. In the last column of Table 1, the interval PI_2 is presented because it has shortest expected length of the two exact prediction intervals.
In the first column of Table 1 the number of the record value to be predicted is listed. In the second column the corresponding record is given. The first entry 85.74 in this column is the first javelin throw performance exceeding the chosen threshold 80 m and therefore is considered to be the first record value in the data set. The predictions in the third and forth column of each row are based on the record values in the second column and previous rows. Especially, in the first row a prediction without previously observed records is not possible. The record value in the last row is still waiting to be performed.

TABLE 1. Table of number of predicted record $r+1$, corresponding records, MPSP based on records $1, \ldots, r$ and prediction intervals with $\alpha = 0.1$.

$r+1$	record	MPSP	PI_2
1	85.74	-	-
2	87.66	91.89	[86.05; 319.86]
3	89.10	91.76	[87.87; 120.42]
4	89.58	92.36	[89.27; 107.17]
5	91.46	92.15	[89.71; 101.62]
6	95.54	93.94	[91.59; 102.08]
7	95.66	98.41	[95.69; 107.18]
8	98.48	98.13	[95.79; 105.24]
9	-	101.07	[98.61; 108.23]

For a small number of observed record values the upper bound of the prediction interval in Table 1 takes large values and thus respective prediction intervals are not meaningful. The length of the prediction interval decreases for larger r and then 90%-prediction intervals perform well. The seventh record value is the only one that does not lie in the corresponding prediction interval. The larger MPSP in that case also indicates that a greater improvement of the world record was expected according to the model. The point prediction of the eighth record is quite close to the observed record value. The MPSP in the last row indicates that the 100 m barrier may be broken by the next future world record performance.

3 Conclusion

In the prediction of future sports records based on previous record values several model assumptions are made. The underlying random variables are supposed to be independent and identically distributed and an underlying distribution has to be chosen, which, in our case, is the Pareto distribution.

Although the assumptions have to be discussed critically, the prediction results in the model of common upper record values show a promising performance with an increasing number of observed records. The analysis of models closer to reality, such as by incorporating a trend in the data over time, will be subject matter of our future work.

References

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (1998) *Records*. Hoboken, NJ: John Wiley & Sons, Inc.

Einmahl, J. H. J. and Magnus, J. R. (2008). Records in athletics through extreme-value theory. *Journal of the American Statistical Association*, **103**, 1382 – 1391.

Empacher, C., Kamps, U., and Volovskiy, G. (2023). Statistical prediction of future sports records based on record values. *Stats*, **6**, 131 – 147.

Fraga Alves, I., de Haan, L., and Neves, C. (2013). How far can man go? In: Torelli, N., Pesarin, F., Bar-Hen, A. (Eds) *Advances in Theoretical and Applied Statistics*, Springer, Heidelberg, 187 – 197.

Nevzorov, V. B. (2001) *Records: Mathematical Theory*. Providence, RI: American Mathematical Society.

Stephenson, A. G. and Tawn, J. A. (2016). Determining the best track performances of all time using a conceptual population model for athletics records. *Journal of Quantitative Analysis in Sports*, **9**, 67 – 76.

Volovskiy, G. and Kamps, U. (2020). Maximum product of spacings prediction of future record values. *Metrika*, **83**, 853 – 868.

Competing risk modelling for in-hospital length of stay

Juan Carlos Espinosa-Moreno[1], Fernando García-García[1], Dae-Jin Lee[1,2], María J. Legarreta-Olabarrieta[3], Susana García-Gutiérrez[3], Naia Mas[4]

[1] Basque Center for Applied Mathematics (BCAM); Bilbao, Spain.
[2] IE University, School of Science and Technology; Madrid, Spain.
[3] Galdakao-Usansolo University Hospital, Research Unit; Galdakao, Spain.
[4] Galdakao-Usansolo University Hospital, Critical Care Unit; Galdakao, Spain.

E-mail for correspondence: jcespinosa@bcamath.org

Abstract: In this study, we propose a framework for analysing in-hospital patient data from electronic health records. We transform longitudinal sparse vital signs measurements into cross-sectional data via descriptive statistics, imputing missing values, and evaluating variables strongly associated with time to mutually exclusive events (favourable medical discharge or deterioration). We employ competing risk and random survival forest techniques to predict patients' length of stay and evaluate models' performance via Brier score.

Keywords: Competing risks; Survival analysis; Variable selection; Electronic health records.

1 Introduction

The evaluation of the health status of hospitalised patients is often conducted based on electronic health records (EHR), tending to result in sparse datasets with problems such as high rates of missing values. In this work, we aim to study each patient's length of in-hospital stay (LoS) as a function of their cross-sectional vital signs statistics, alongside sex and age. Here, our target is to model the time-to-event until one of the two possible (mutually exclusive) final situations occur: either favourable discharge or clinical deterioration. For this, we employ competing risk models such as: cause-specific Cox proportional hazard regression, Fine and Gray's subdistribution hazard model, and cause-specific random survival forests.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Materials and methods

2.1 Data description and pre-processing

In the Galdakao-Usansolo University Hospital (Basque Country, Spain), a total of 19,602 hospitalisations (lengths of stay at least 24 hours) were collected during the year 2019, of which 852 (4.35%) resulted in deterioration. These data correspond to 55.8% males and 44.2% females. Those data are split into train and testing data (70% and 30%, respectively), via stratified random sampling, to keep the proportion of events. Training data has 13,722 hospitalisations with 597 (4.35%) that result in deterioration. Otherwise, the test has 5,880 hospitalisations with 255 (4.33%) in deterioration.

For each hospitalisation, we have the patient's sex and age, as well as longitudinal data along the hospitalisation for 7 vital signs: temperature, systolic and diastolic blood pressure, heart and respiratory rates, oxygen saturation and neurological state. We summarise these longitudinal data with the following statistics: maximum, minimum, first observation, last observation, mean, standard deviation, average percentage change (apc) and average change per time unit (acptu), transforming the original variables into a cross-sectional higher dimensional space. Then, we use the Multiple Imputation by Chained Equations (MICE) method for imputing missing cross-sectional values.

2.2 Variable selection

To detect which variables are strongly associated with time-to-event, where possible events are deterioration and favourable discharge, we employ the LASSO Regularized Cox Regression (Simon *et al.,* 2009) and Best Subset Selection (Wen *et al.,* 2017) in CoxPH models. In LASSO, we obtain the best regularisation parameter λ by k-fold cross-validation (CV). In each one, LASSO and BeSS (Best Subset Selection), we define two models: (a) One using deterioration as an event and favourable discharge as censored data, where we obtained a set s_1 of variables; (b) one with deterioration as censored and favourable discharge as the event, where we obtain a set s_2 of variables. Finally, we define the definite set of variables as $s = s_1 \cup s_2$, which is a subset of the full set of variables.

2.3 Time-to-event models

Given that hospitalisations can result in two mutually exclusive final states: favourable discharge or clinical deterioration, we opted for competing risk models. The first type of model that we use is the Cause-Specific Cox (Austin *et al.,* 2016) (CSC), where the hazard function denotes the instantaneous rate of occurrence of the k-th event in subjects who are currently

event-free. For each possible event $D \in \{1, \cdots, K\}$, it is calculated as described below.

$$\lambda_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t, D = k | T \geq t\right)}{\Delta t}, \tag{1}$$

where T is the random variable "baseline time until the occurrence of the event of interest" (such as death, failure, etc.), $t \in [0, \infty)$ and, in our case, $K = 2$. The second type of model, known as Fine and Gray (Austin *et al.,* 2016) (FG) or sub-distribution hazard function, defines the instantaneous risk of failure from the k-th event in subjects who have not yet experienced an event of type k (hazard function), as in Equation 2.

$$\lambda_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t, D = k | T > t \ \cup (T < t \cap D \neq k)\right)}{\Delta t}. \tag{2}$$

The third model is an adaptation of random survival forest (RSF) proposed by Ishwaran *et al.,* 2008. Analogous to the cause-specific Cox –which estimates cause-specific hazards using linear models, with all other events taken as censored–, here we employ a cause-specific RSF (CS-RSF), where each RSF is trained for a particular hazard function. The RSF hyperparameters are chosen by tuning in subsamples employing the out-of-sample error.

To evaluate the accuracy of a predicted survival function at a given time t, we use Brier score (BS). For a dataset of N individuals, survival times T_i, co-variables $\mathbf{X}_i$ and predicted survival function $\hat{S}(t)$, Brier score is defined as $\text{BS}(t, S) = E(1_{T_i \geq t} - \hat{S}(t|\mathbf{X}_i))^2$, calculated for each possible final state.

3 Results and conclusions

Employing LASSO and BeSS, Table 1 summarizes the variables that are discarded to model the time to patient deterioration or discharge. We can see that the statistic more discarded was the first observation and the mean.

TABLE 1. Variables discarded by LASSO and BeSS methods.

Vital sign	**Discarded by LASSO**	**Discarded by BeSS**
Temperature	apc	first
Systolic pressure	first, mean	first
Diastolic pressure	apc	—
Heart rate	mean	apc
Respiratory rate	max, first, mean, sd, acptu	min, last, mean, acptu
Oxygen saturation	mean, acptu	—
Neurological state	—	—

We compare CSC, FG and CS-RSF techniques concerning three models: the full model, which employs all the variables; a model with the variables

selected by LASSO Cox regression; finally, a model with the variables selected by BeSS. Then we calculated the Brier score for a LoS of 2, 3, 4 and 5 days (48, 72, 96 and 120 hours).

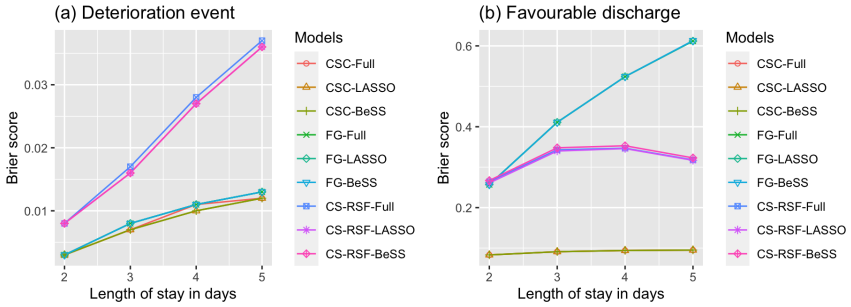


FIGURE 1. Brier score for the proposed models and variable selection methods.

As depicted in Figure 1, for both types of event, there is no meaningful gain due to variable selection. Across models, CSC achieves the highest prediction performance in BS for both time until both deterioration and favourable discharge. For deterioration, FG has similar BS to CSC, whereas CS-RSF had the worst performance. For favourable discharge, CS-RSF behaved better than FG, but both were far from CSC.

As a conclusion, in this work, we established a framework for modelling competing in-hospital LoS as a function of patients' vital signs (obtained from EHR). We obtained cross-sectional statistics of the time series data, dealt with high rates of missing values, and predicted length of stay by combining 'classical' time-to-event and machine learning models.

References

Austin P.C., Lee D.S., Fine J.P. (2016). *Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation*, **133(6)**, 601 – 609.

Ishwaran H., Kogalur UB., Blackstone EH., Lauer MS. (2008). *Random survival forests. The Annals of Applied Statistics*, 841 – 860.

Simon N., Friedman J., Hastie T., Tibshirani R. (2009). *Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw*, **39(5)**, 1 – 13.

Wen C., Zhang A., Quan S. and Wang X. (2020). *BeSS: An R Package for Best Subset Selection in Linear, Logistic and Cox Proportional Hazards Models. J Stat Softw*, **94(4)**.

Mixed nonlinear modelling in food engineering: determination of the salting time of boneless dry-cured Cerretan hams

Xavier Espuña[1], Lesly Acosta[1], Josep A. Sanchez-Espigares[1], Xavier Tort-Martorell[1]

[1] Department of Statistics and Operations Research, Polytechnic University of Catalonia, Barcelona, Spain

E-mail for correspondence: lesly.acosta@upc.edu

Abstract: A great challenge in producing a good cured ham is to reduce the variability of the salt content between pieces of ham and to obtain homogeneity in terms of flavour and quality in general. This reduction in variability would imply a reduction in salt content, a recommendation of the World Health Organisation (WHO, 2007). This work focuses on the salting process of boneless Cerretan hams and our aim is two-fold: 1) to build a mathematical model that enables —through predictions— the reduction of the variability of salt between pieces, and 2) to determine an 'appropriate' salting time for each ham.
We propose a novel strategy within the ham industry to determine appropriate hams extraction time from the salting pile and we postulate that it is statistically and practically advantageous to the habitual hams extraction strategy (removal based on fat and weight classification and all at the same time).
We build a non-linear mixed (NLM) model that, according to the final salt uptake target of 1.7%, would allow to decide each ham extraction time depending on the initial weight and fat, plus the weight decrease on day one. This model has to be applicable in industrial production, albeit in an approximate form. To account better for the salting-time estimated uncertainty, we run a nonparametric bootstrap. A further aim is to extrapolate the use of the NLM modelling methodology and proposed novel extraction strategy to other boneless hams industrial production systems in Europe.

Keywords: NLM model; Bootstrap; Cured-ham; Salt content; Extraction time.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

1 Introduction

Various European countries have specialities of cured hams that are salted, dried and matured off the bone, such as Speck Alto Adige, Schwarzwälder Schinken, Tyrolean Speck, Culatello di Zibello from Italy as well as the Cerretan ham in Spain.
A great challenge in producing a good cured ham is to reduce the variability of the salt content between pieces of ham and to obtain homogeneity in terms of flavour. This reduction in variability would imply a reduction in salt content, a recommendation of the World Health Organisation (WHO, 2007), although a too low salt content could also increase pastiness defects in the texture of the ham that usually occurs at a low threshold of 1.4%; see, e.g., Toldrá et al (1997), Coll-Brasas et al. (2019), and Martín-Gómez et al. (2022).
This work focuses on the salting process of boneless Cerretan hams and our final goal is to propose a NLM model that, according to the average salt uptake target of 1.7%, would allow us to decide on the appropriate date of removal of the hams from the salting pile depending on the initial fat and weight, plus the weight decrease on day one. This model has to be applicable in industrial production, albeit in an approximate form, to estimate the salt pile removal time. We also run a nonparametric bootstrap study, to account better for the estimated salting-time uncertainty.

2 Materials and methods

2.1 Ham samples and Salting process

Twenty-seven lean Cerretan hams were selected from a nearby slaughterhouse with an initial pH between 5.8–6 (García-Rey et al., 2004). The hams had a fat percentage, as measured by the Multiscan X-ray technology, of 15% to 24%. Each ham was weighed, and measured for fat. In the salting phase, first, a specific seasoning mixture is used (Gratacos et al., 2013). The hams are left to stand for a day to absorb the salt and then covered with recycled, moist (4%-5%) salt (T-3 sea salt) at a temperature of 2ºC to 4ºC, and at a humidity of 85% to 95%. Lastly, they are placed on a flat surface in a container that lets the exudate run off.
The time starts counting when the hams were covered with the recycled salt and after 24 hours, the salt content is measured and also the weight to determine its decrease on day one. The hams continue to be measured in terms of salt uptake over time during seven days, since we study the salt acquisition curve up to the moment of saturation. Note that hams are normally salted for 0.5 days per kg of initial weight, which would mean that they are salted for around 4 days. In this study, however, they were left in the salting pile for 7 days, since it is of interest to study the salt acquisition curve up to the moment of saturation.

2.2 Nonlinear mixed modelling

To study the daily evolution of the salt content of each of the 27 sampled Cerretan hams pieces, which would allow to model the expected value of salt content and characterize its variability, the NLM modelling statistical methodology is used (Pinheiro and Bates, 2000). Specifically, we built a NLM model of logistic-type to predict the amount of salt that each ham will have at each point in time. A general expression of the logistic NLM model that, apart from time (days), depends on three fixed parameters $(\beta_1, \beta_2, \beta_3)$' and three random effects b_{1i}, b_{2i} and b_{3i} on each parameter is given by:

$$\begin{aligned} Y_i &= f(t, \phi_i) + \epsilon_i, \\ &= \mu_i(t) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2 I) \quad i = 1, \ldots, k \end{aligned}$$

where

$$\mu_i(t) = \frac{\Phi_{1i}}{1 + \exp(-\frac{t - \Phi_{2i}}{\Phi_{3i}})}.$$

and

$$\phi_i = \begin{pmatrix} \Phi_{1i} \\ \Phi_{2i} \\ \Phi_{3i} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{pmatrix}.$$

The parameter Φ_{1i} represents the salt saturation and the parameter Φ_{2i} the time it takes to reach half of the saturation value. The parameter Φ_{3i} tells us how steep is the central part of the logistic S-shaped function, i.e. the instantaneous speed at which the salt is absorbed.
The vector of random effects $(b_{1i}, b_{2i}, b_{3i})'$ has a normal distribution with $b_i \sim N(0, \Sigma_b)$. The fixed effects vector of the parameters could be expressed as a linear function of other variables.
The general procedure, using the data of the 27 sampled ham pieces, for building a valid model to predict the salt average amount that each ham will have at each time point, and later to estimate when to remove the ham from the salting pile is:

1. Describe the time evolution of the 7-day salt measurements for each of the 27 sampled hams; in this case the observed S-shaped behaviour suggested the fit of a logistic NLM model.

2. Fit the logistic NLM model with salt-uptake as an outcome and only time as an independent variable. In this case, a strong correlation between the parameters Φ_{2i} and Φ_{3i} was observed and to avoid over-parameterization we get a reduced model by expressing Φ_{2i} in terms of the the parameter Φ_{3i}.

3. Fit an NLM model considering, apart from time as an independent variable, the covariates initial fat and weight, plus the weight decrease on day one. This model should be simplified to obtain a final model including only significant terms.

4. Validate the model and report the estimation results. In this case, the fitted model was properly validated (results not shown).

5. With the final model, predict the salt uptake per ham and also determine the extraction time from the salting pile required to reach the fixed target salt average content of 1.70%. A limitation, however arises, because it would not be industrially feasible to extract the hams at every predicted time of day; it would involve excessive labour costs.

6. Propose as a solution the strategy of considering five possible predefined extraction points in time (days: 1, 1.5, 2, 2.5 and 3) at which hams (as a percentage per day) could be removed from the salting pile. With this strategy, we improved the habitual hams extraction strategy (removal based on weight and fat similarities, and all at the same time) in terms of bias and uncertainty.

2.3 Nonparametric bootstrap

To better quantify the uncertainty of the salting time of extraction, we run a nonparametric bootstrap (Davison, A. C. and Hinkley, 1997). The idea is that the 1000 simulated data would have similar characteristics to our 27 experimental units and based on these trajectories, the time at which the target 1.7% salt content is reached can be estimated, and also the empirical distribution of those extraction salting times. This distribution, being per se continuous, is discretized using the five predefined extraction points in order to plan the removal of hams from the salting pile based on a feasible industrial production plan. This discretization will produce a small increase in the variability of the measured salt in removed hams, but despite that, the obtained uncertainty will be lower than the one obtained with the habitual hams extraction strategy.

References

Coll-Brasas, E., Arnau, J., Gou, P., Lorenzo, J.M., García-Pérez, J.V., and Fulladosa, E. (2019). Effect of high pressure processing temperature on dry-cured hams with different textural characteristics. *Meat Science*, **152**, 127 – 133.

Davison, A. C. and Hinkley, D. V. (1997). Bootstrap Methods and their Application. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

García-Rey, R.M., García-Garrido, J.A., Quiles-Zafra, R., Tapiador, J., and Luque-de-Castro, M.D. (2004). Relationship between pH before salting and dry-cured ham quality. *Meat Science*, **67(4)**, 625 – 632.

Gratacós-Cubarsí, M., Sárraga, C., Castellari, M., Valero, A., García Regueiro, J., and Arnau, J. (2013). Effect of ph24h, curing salts and muscle types on the oxidative stability, free amino acids profile and vitamin b2, b3 and b6 content of dry-cured ham. *Food Chemistry*, **141(3)**, 3207 – 3214.

Martín-Gómez, A., Segura-Borrego, M.P., Ríos-Reina, R., José Cardador, M., Callejón, R.M., Morales, M.L., Rodríguez-Estévez, V., and Arce, L. (2022). Discrimination of defective dry-cured Iberian ham determining volatile compounds by non-destructive sampling and gas chromatography. *LWT*, **154**, article no. 112785.

Pinheiro, J. and Bates, D.M. (2000). Nonlinear mixed-effects models: basic concepts and motivating examples. *Mixed-effects models in S and S-Plus*. Springer.

Toldrá, F., Flores, M., and Sanz, Y. (1997). Dry-cured ham flavour: enzymatic generation and process influence. *Food Chemistry*, **59(4)**, 523 – 530.

WHO. (2007). Reducing salt intake in populations: Report of a WHO forum and technical meeting. *World Health Organization Geneva, Switzerland*.

Learning Gaussian Bayesian networks from incomplete data - The Bayesian way

Marco Grzegorczyk[1]

[1] Bernoulli Institute, Groningen University, The Netherlands

E-mail for correspondence: m.a.grzegorczyk@rug.nl

Abstract: We propose a Bayesian Model Averaging (BMA) approach for learning Gaussian Bayesian networks (BNs) from data with missing values. We present a Markov Chain Monte Carlo sampling algorithm that allows for simultaneously sampling directed acyclic graphs (DAGs) as well as the values of the unobserved data points. We compare the network reconstruction accuracy of our new BMA approach with two non-Bayesian approaches for learning BNs from incomplete data. For the empirical evaluation we use protein data from the RAF pathway.

Keywords: Gaussian Bayesian networks; Incomplete data; Bayesian Model Averaging (BMA); Markov Chain Monte Carlo (MCMC)

1 Introduction

Bayesian networks (BNs) are an important model class for statistically modelling the conditional (in-)dependence relations among random variables $X_1, \ldots, X_n$. The variables become the nodes of a directed acyclic graph (DAG) whose edges encode the conditional (in-)dependence relations among them. Given a DAG $\mathcal{G}$ the n-dimensional joint distribution factorizes into n univariate conditional distributions:

$$P(X_1, \ldots, X_n|\mathcal{G}) = \prod_{i=1}^{n} P(X_i|\pi_{\mathcal{G}}^i) \tag{1}$$

where $\pi_{\mathcal{G}}^i$ is the parent node set of X_i implied by the DAG $\mathcal{G}$. We recall that X_j is a parent of X_i if there is an edge from X_j to X_i, symbolically $X_j \rightarrow X_i$. Learning BNs from data is challenging, as the number of possible DAGs grows super-exponentially in n and the acyclicity constraint does not allow this task to be decomposed and to be solved in parallel.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

If the available data set has missing values, not only the BN but also the missing data points have to be inferred from the observed data. A classical but still widely-applied approach for learning BNs from incomplete data is the structural Expectation Maximisation (EM) algorithm from Friedman (1997). It searches for the best DAG in terms of a penalized likelihood inside an EM algorithm. Conceptually easier is to employ penalized node-average log-likelihoods (NALs). The underlying idea is to compute the local scores $X_i|\pi^i_{\mathcal{G}}$ using the 'locally complete' observations and to scale them accordingly (Bodewes and Scutari, 2021). To the best of our knowledge, no Bayesian approach for learning BNs from incomplete data has been developed yet. To fill the gap, we build on the 'Bayesian metric for Gaussian networks having score equivalence' (BGe score) of Geiger and Heckermann (2002). We extend the structure Markov Chain Monte Carlo (MCMC) sampler (Madigan and York, 1995) to allow for simultaneously sampling directed acyclic graphs (DAGs) as well as the values of the unobserved data points from the posterior distribution. Like the competing methods, our approach assumes that values are 'missing completely at random'.

2 Outline of theory

The goal of BN structure learning is to infer DAGs $\mathcal{G}$ from data $\mathcal{D}$. Following the Bayesian paradigm, we have for the DAG posterior distribution:

$$P(\mathcal{G}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{G})P(\mathcal{G})}{P(\mathcal{D})} \tag{2}$$

where $P(\mathcal{G})$ is the DAG prior probability, $P(\mathcal{D}|\mathcal{G})$ is the marginal likelihood, and $P(\mathcal{D})$ is a normalization constant. In the absence of prior knowledge, we employ a uniform distribution for $P(\mathcal{G})$. The Gaussian BGe score from Geiger and Heckerman (2002) assumes the random vector $(X_1, \ldots, X_n)^\mathsf{T}$ to have an n-dimensional Gaussian distribution:

$$(X_1, \ldots, X_n)^\mathsf{T}|\mathcal{G} \sim \mathcal{N}_n\left(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}\right)$$

whose covariance matrix $\boldsymbol{\Sigma}^{\mathcal{G}}$ implies the factorization in Eq. (1). Complete DAGs (with the maximal number of edges) do not impose any conditional independenies. For complete DAGs the n-dimensional normal-Wishart is used as parameter prior distribution:

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}_n(\boldsymbol{\mu}_0, \alpha_\mu^{-1}\boldsymbol{\Sigma}) \quad \text{and} \quad \boldsymbol{\Sigma} \sim \mathcal{W}_n^{-1}(\alpha_w, \mathbf{R})$$

where $\alpha_\mu > 0$, $\alpha_w > n-1$, $\boldsymbol{\mu}_0 \in \mathbb{R}^n$ and $\mathbf{R} \in \mathbb{R}^{n,n}$ are hyperparameters. Geiger and Heckerman (2002) show that each sample $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the complete DAGs also specifies the parameters $(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$ of any DAG $\mathcal{G}$. Moreover, they show that the marginal likelihood of any DAG $\mathcal{G}$

$$P(\mathcal{D}|\mathcal{G}) = \int\int P(\mathcal{D}|\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}) p(\boldsymbol{\mu}^{\mathcal{G}}|\boldsymbol{\Sigma}^{\mathcal{G}}) p(\boldsymbol{\Sigma}^{\mathcal{G}})\, d\boldsymbol{\mu}^{\mathcal{G}} d\boldsymbol{\Sigma}^{\mathcal{G}}$$

can be computed analytically and also fulfills the conditional independence relations implied by $\mathcal{G}$; cf. Eq. (1). MCMC sampling can then be used to generate DAG samples from the posterior distribution shown in Eq. (2). We here make use of the structure MCMC sampler from Madigan and York (1995), and we implement it with the improvements from Giudici and Castelo (2003). From the posterior sampled DAGs the marginal posterior probabilities of the existence of all possible edges are estimated.
We assume that the data set $\mathcal{D}$ consists of N observations, where each observation features n individual values (i.e. one value for each node X_i). In total, there are then $n \cdot N$ values, and we assume that a fraction, p_m, of these values is 'missing completely at random'. The data $\mathcal{D}$ then consist of two parts: the observed data $\mathcal{D}_{obs}$ and the missing data $\mathcal{D}_{miss}$ and the posterior distribution becomes:

$$P(\mathcal{G}, \mathcal{D}_{miss}|\mathcal{D}_{obs}) \propto P(\mathcal{D}_{miss}, \mathcal{D}_{obs}|\mathcal{G}) \cdot P(\mathcal{G})$$

For generating posterior samples we propose the following MCMC sampling scheme, which consists of three consecutive sampling steps (**S1**-**S3**):

(**S1**) Given the missing values, we have that $\mathcal{D} := \{\mathcal{D}_{miss}, \mathcal{D}_{obs}\}$ is a complete data set. Hence, we can use the structure MCMC sampler for sampling DAGs $\mathcal{G}$ from the posterior distribution $P(\mathcal{G}|\mathcal{D})$ in Eq. (2). Loosely speaking, the DAG $\mathcal{G}$ is varied by proposing single edge additions, deletions and reversals and the new DAGs are accepted with the usual Metropolis Hastings acceptance probabilities.

(**S2**) Conditional on the complete data $\mathcal{D} := \{\mathcal{D}_{miss}, \mathcal{D}_{obs}\}$ and the DAG $\mathcal{G}$, we can sample the model parameters from the posterior distribution $P(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}}|\mathcal{D}, \mathcal{G})$. These parameters are the expectation vector and the covariance matrix of a multivariate Gaussian distribution $\mathcal{N}_n(\boldsymbol{\mu}^{\mathcal{G}}, \boldsymbol{\Sigma}^{\mathcal{G}})$, whose covariance matrix $\boldsymbol{\Sigma}^{\mathcal{G}}$ must imply the conditional (in-)dependence relations implied by the DAG $\mathcal{G}$, as indicated in Eq. (1). Sampling such parameters is not straightforward and can only be done indirectly. We propose the following algorithm for it:

(**S2a**) Sample the expectation vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ from the posterior distribution for complete DAGs:

$$\begin{aligned} \boldsymbol{\Sigma}|\mathcal{D} &\sim \mathcal{W}_n^{-1}(\alpha_w + N, \mathbf{T}) \\ \boldsymbol{\mu}|(\boldsymbol{\Sigma}, \mathcal{D}) &\sim \mathcal{N}_n\left(\boldsymbol{\mu}^{\diamond},\ (\alpha_\mu + N)^{-1}\boldsymbol{\Sigma}\right) \end{aligned}$$

where $\alpha_w,\ \alpha_\mu > 0$ are hyperparameters, N is the number of observations, and the matrix $\mathbf{T}$ and the vector $\boldsymbol{\mu}^{\diamond}$ can be computed from the data $\mathcal{D}$ and the hyperparameters $\mathbf{R}$ and $\boldsymbol{\mu}_0$, respectively. We can use $\boldsymbol{\mu}^{\mathcal{G}} := \boldsymbol{\mu}$, but the covariance matrix $\boldsymbol{\Sigma}$ refers to a complete DAG, and hence it does not imply the conditional independence relations implied by the DAG $\mathcal{G}$.

In steps (**S2b**-**S2c**) we extract from $\mathbf{\Sigma}$ a covariance matrix $\mathbf{\Sigma}^{\mathcal{G}}$ that is coherent with $\mathcal{G}$.

(**S2b**) Recall that $\pi^i_{\mathcal{G}}$ denotes the parent set of variable X_i implied by $\mathcal{G}$. Given $\mathbf{\Sigma}$ from step (**S2a**), compute the parameters of the n local univariate conditional Gaussian distributions:

$$X_i|\pi^i_{\mathcal{G}} \sim \mathcal{N}_1(\tilde{\mu}_i, \tilde{\sigma}^2_i) \qquad (i = 1, \ldots, n)$$

where $\tilde{\mu}_i$ and $\tilde{\sigma}^2_i$ are the expectation and the variance of the i-th univariate Gaussian. The expectations are of the form

$$\tilde{\mu}_i = \mu_i + \sum_{j \in \pi^i_{\mathcal{G}}} b_{i,j}(X_j - \mu_j) \qquad (i = 1, \ldots, n) \tag{3}$$

where μ_j is the unconditional expectation of X_j $(j = 1, \ldots, n)$ and $b_{i,j}$ can be thought of as a regression coefficient. As every DAG follows a topological order, the nodes $X_1, \ldots, X_n$ can be ordered and relabeled such that we can re-write Eq. (3) as:

$$\tilde{\mu}_i = \mu_i + \sum_{j=1}^{i-1} \tilde{b}_{i,j}(X_j - \mu_j) \qquad (i = 1, \ldots, n) \tag{4}$$

where $\tilde{b}_{i,j} = 0$ if $X_j \notin \pi^i_{\mathcal{G}}$.

(**S2c**) Henceforth, we have the n univariate conditional Gaussians

$$X_i|(X_1, \ldots, X_{i-1}) \sim \mathcal{N}_1\left(\mu_i + \sum_{j=1}^{i-1} \tilde{b}_{i,j}(X_j - \mu_j), \tilde{\sigma}^2_i\right) \tag{5}$$

and the recursive formula of Shachter and Kenley (1989) can be used to compute the covariance matrix $\mathbf{\Sigma}^{\mathcal{G}}$ of the joint Gaussian distribution of $(X_1, \ldots, X_n)^{\mathsf{T}}$. As we have $\tilde{b}_{i,j} = 0$ if $X_j \notin \pi_{\mathcal{G}^i}$, the resulting covariance matrix $\mathbf{\Sigma}^{\mathcal{G}}$ is coherent with $\mathcal{G}$; i.e. it implies the conditional (in-)dependence relations from Eq. (1).

(**S3**) The data $\mathcal{D}$ are a random sample from the $\mathcal{N}_n(\boldsymbol{\mu}^{\mathcal{G}}, \mathbf{\Sigma}^{\mathcal{G}})$ Gaussian distribution, symbolically, $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with $\mathbf{x}_k \sim \mathcal{N}_n(\boldsymbol{\mu}^{\mathcal{G}}, \mathbf{\Sigma}^{\mathcal{G}})$. Hence, we can loop through the observations and complete each observation by sampling the missing values conditional on the observed values from conditional Gaussian distributions. In case of missing data, $\mathbf{X}$ consists of two parts: the observed subvector $\mathbf{X}_{obs}$ and the unobserved subvector $\mathbf{X}_{miss}$. Given $\mathbf{X}_{obs} = \mathbf{x}_{obs}$, we have for $\mathbf{X}_{miss}$:

$$\mathbf{X}_{miss}|(\mathbf{X}_{obs} = \mathbf{x}_{obs}, \boldsymbol{\mu}^{\mathcal{G}}, \mathbf{\Sigma}^{\mathcal{G}}) \sim \mathcal{N}_k\left(\boldsymbol{\mu}^{\mathcal{G}}_{miss|obs}, \mathbf{\Sigma}^{\mathcal{G}}_{miss|obs}\right)$$

p_m	EM BIC	EM 0.4	EM 0.25	EM 0.1	NAL BIC	NAL 0.4	NAL 0.25	NAL 0.1	NEW
0	0.63	0.69	0.63	0.63	0.62	0.69	0.64	0.63	**0.75**
0.1	0.66	0.68	0.65	0.63	0.57	0.58	0.61	0.61	**0.75**
0.2	0.66	0.67	0.63	0.57	0.51	0.51	0.54	0.58	**0.75**
0.4	0.66	0.67	0.63	0.56	0.53	0.53	0.52	0.54	**0.71**

TABLE 1. **Average AUC scores for the RAF pathway.** The complete data set has $n = 11$ variables (proteins) and $N = 3530$ observations. For each p_m we generated 10 incomplete data sets by removing the fraction p_m of randomly selected data points. When comparing the mean AUCs of the 8 competing methods with the mean AUC of the Bayesian method (NEW), all two-sided paired t-test p-values were below the standard test level $\alpha = 0.05$.

where k is the dimension of X_{miss} and

$$\begin{aligned}\boldsymbol{\mu}^{\mathcal{G}}_{miss|obs} &:= \boldsymbol{\mu}^{\mathcal{G}}_{miss} + \boldsymbol{\Sigma}^{\mathcal{G}}_{miss,obs}\left\{\boldsymbol{\Sigma}^{\mathcal{G}}_{obs,obs}\right\}^{-1}\left(\mathbf{x}_{obs} - \boldsymbol{\mu}^{\mathcal{G}}_{obs}\right)\\ \boldsymbol{\Sigma}^{\mathcal{G}}_{miss|obs} &:= \boldsymbol{\Sigma}^{\mathcal{G}}_{miss,miss} - \boldsymbol{\Sigma}^{\mathcal{G}}_{miss,obs}\left\{\boldsymbol{\Sigma}^{\mathcal{G}}_{obs,obs}\right\}^{-1}\boldsymbol{\Sigma}^{\mathcal{G}}_{obs,miss}\end{aligned}$$

The subscripts 'obs' and 'miss' refer to the subvectors and submatrices that only contain the rows and columns that belong to observed or missing data points.

2.1 Competing methods

We compare our new Bayesian approach with two non-Bayesian approaches, namely the structural EM (Friedman, 1997) and the node-average Likelihood (NAL) approach (Bodewes and Scutari, 2021). For these methods we use the R implementation from Bodewes and Scutari (2021) and we apply them with the same four penalty parameters ('BIC', 0.4, 0.25 and 0.1).

3 Empirical results

For lack of space, we only present the results of a study on phosphorylation data from the RAF protein signalling pathway. Sachs et al. (2005) measured the phosphorylation sites of $n = 11$ proteins of the RAF pathway. We use the $N = 3530$ observational measurements and the gold-standard network of the RAF pathway from Sachs et al. (2005) as proxy for the true DAG. We distinguish four fractions of missing data $p_m \in \{0, 0.1, 0.2, 0.4\}$. For each p_m we generate 10 incomplete data sets by deleting different randomly selected data points. We then learn the network structure from each data

set and quantify the network reconstruction accuracy in terms of the Area Under the receiver operator characteristic Curve (AUC) scores. We recall that $0 \leq \text{AUC} \leq 1$ with higher AUCs indicating a better performance. Table 1 shows the average AUC scores. Each average AUC in Table 1 is across 10 data instantiations with different missing values. The new BMA approach leads consistently to the highest average AUC score.

4 Discussion and conclusions

We have proposed a new Bayesian Model Averaging (BMA) approach for learning Gaussian Bayesian networks from data with missing values. The new method builds on the Gaussian BGe score and it extends the structure MCMC sampler for DAG sampling (**S1**) by introducing new MCMC moves (**S2-S3**) that sample the missing data points. Our empirical results suggest that the new approach leads to a higher network reconstruction accuracy than two non-Bayesian state-of-the-art approaches.

References

Bodewes, T. and Scutari, M. (2021). Learning Bayesian networks from incomplete data with the node-average likelihood. *International Journal of Approximate Reasoning*, **138**, 145 – 160.

Friedman, N. (1997). Learning belief networks in the presence of missing values and hidden variables. In: *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, Nashville, Tennessee, 125 – 133.

Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, **30**, 1412 – 1440.

Giudici, P. and Castelo, R. (2003). Improving Markov Chain Monte Carlo Model search for Data Mining. *Machine Learning*, **50**, 127 – 158.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215 – 232.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. and Nolan, G.P. (2005). Protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523 – 529.

Shachter, R.D. and Kenley, R. (1989). Gaussian influence diagrams *Management Science*, **35**, 527 – 550.

Grouped regression modeling of proteins

Jonas Heiner[1], Jan Hengstler[2], Andreas Groll[1]

[1] TU Dortmund, Germany
[2] Leibnitz-Institut für Arbeitsforschung an der TU Dortmund, Germany

E-mail for correspondence: jonas.heiner@tu-dortmund.de

Abstract: The relation between the expression of a gene and the resulting levels of the corresponding protein is known to be positively correlated, but gene expression explains only a relatively small fraction of the variance of protein expression. This motivates the utilization of regression models in order to investigate the relationship between gene expression and protein levels. Co-expression analysis for gene grouping is used for the regression models to additionally consider the grouped genes as covariates for modeling a protein's expression. Quality measures are compared for the models, which show a clear improvement of the protein modeling when including grouping information of the genes.

Keywords: Multi-Omics; LASSO Regression; Co-Expression Analysis.

1 Introduction

Proteins are encoded by an organism's DNA and result via transcription and translation of the gene. In this way, every protein can be uniquely associated to a gene, resulting in standalone gene-protein-pairs. Thus, it is a general assumption that gene expression and protein counts are positively correlated. However, this is not always the case in practice. Thus, protein levels can not necessarily be modeled reliably solely by their associated gene as covariate. Hence, the idea of this work is to additionally consider congeneric genes from a gene co-expression analysis. As an organism's genome comprises a large amount of genes and the number of samples in toxicological studies is typically small, regularization methods can address this problem and also handle multicollinearity among covariates (Breiman, 1996). The proposed regression models for protein modeling as well as the co-expression analysis are further described in Section 2. In Section 3, the underlying data structure for the gene-protein-models is specified and the

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

main results of the real data application are emphasized in Section 4. Finally, Section 5 concludes with work in progress.

2 Statistical Analysis

The main goal of this work is the investigation of transcriptomics and proteomics relationship and to improve prediction performance of proteins. This is achieved by modeling the protein expressions (PE) as target variable with gene expressions (GE), treatment (T) and duration (D) as covariates. The models that are considered for a single gene-protein-pair j separately are listed in formulas (1) - (4) below. The simplest models (model (1) and (2)) with the pair's gene expression as covariate serve as baseline models. Those models can be treated as ordinary linear regression model, as the PEs can be assumed to be normally distributed. The next step is to include grouped genes (details will follow below) and, eventually, the corresponding proteins as covariates (model (3) and (4)). Here, for a gene-protein-pair j, $\mathcal{I}_j$ is the index set of all G_j genes, which are associated to GE_j, including GE_j itself. When including large amounts of covariates from the grouping process, multicollinearity may become an issue and the problem of the number of parameters exceeding the sample size needs to be addressed. This can be achieved by utilizing regularization techniques, such as LASSO (least absolute shrinkage and selection operator) regression. Here, the minimization problem of the ordinary regression model is penalized by an additional penalty term on the absolute coefficient values (Tibshirani, 1996) and which is implemented in the `R`-package `glmnet` (Friedman et al., 2010). The respective pair's gene GE_j is assumed to always affect the corresponding protein and, therefore, is excluded from the shrinkage process.

$$PE_j = \beta_0 + GE_j \cdot \beta_{GE,j} \tag{1}$$

$$PE_j = \beta_0 + (T,\, D,\, GE_j)^{\mathrm{T}} \cdot \boldsymbol{\beta} \tag{2}$$

$$PE_j = \beta_0 + T \cdot \beta_T + D \cdot \beta_D + \sum_{k \in \mathcal{I}_j} GE_k \, \beta_{GE,k} \tag{3}$$

$$PE_j = \beta_0 + T \cdot \beta_T + D \cdot \beta_D + \sum_{k \in \mathcal{I}_j} GE_i \, \beta_{GE,k} + \sum_{k \in \mathcal{I}_j \setminus j} PE_k \, \beta_{PE,k} \tag{4}$$

These regression model approaches can then be evaluated by comparing certain goodness-of-fit measures. In the following, we use the adjusted R^2 as the proportion of variance that can be explained by a model whilst taking the penalization of the parameter number into account. Additionally, the Akaike information criterion (AIC) is considered as an in-sample measure, which provides a direct trade-off between goodness-of-fit and model

complexity. Moreover, the prediction error as an out-of-sample measure is computed as the root mean square error (RMSE) using the predictions of a test sample based on the model fitted on a training sample. For this, one observation of each realized treatment-duration combination is randomly drawn to form the test sample and the model is fit on the remaining observations. The final prediction error is then the average across d-times randomly drawn train-test-splits.

For gene grouping in gene expression data, weighted co-expression analysis (Zhang and Horvath, 2005) is perfectly suitable and used in the following. Whether two genes i, j are denoted as co-expressed depends on a specified similarity measure s_{ij}, typically the absolute Pearson correlation, and a threshold. Based on the similarity measure, an adjacency function $a_{ij} = |s_{ij}|^{\delta}$ is used, which is specifically suitable for soft thresholding and depends on a power coefficient $\delta \in \mathbb{N}$. Now, for identification of gene groups, hierarchical clustering is applied based on a distance measure d_{ij}, which is constructed with the help of the adjacency function. In particular, it is defined as $d_{ij} := 1 - \omega_{ij}$, where the similarity measure ω_{ij} is the topological overlap of two genes i, j, with

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}},$$

and $l_{ij} = \sum_u a_{iu} a_{uj}$ and $k_i = \sum_u a_{iu}$. For determination of the threshold parameter δ of the adjacency function, a scale-free topology criterion (Zhang and Horvath, 2005) can be applied. Roughly, for increasing power values for δ, a scale-free topology index R^2_{top} and the mean topological-overlap-based connectivity $mean(k)$ with $k = \sum_{j=1}^{n} \omega_{ij}$ are computed. The index R^2_{top} is a measure of how well a grouping process satisfies a scale-free topology and is based on the quadratic correlation between the logarithmic connectivity k and the logarithm of its frequency distribution. For further information we refer to Zhang and Horvath (2005). The power δ is then chosen as the one that maximizes the scale-free topology index R^2_{top}, while keeping a pre-specified level of mean connectivity $mean(k)$.

3 Data Structure

To apply the statistical models to real omics data, gene expression data as well as protein data from 36 mice is available. The data covers the investigation of how liver fibrosis influences lobular zonation. Of the 36 mice, a test sample of 18 mice were induced carbon tetrachloride (CCl_4, 1 g/kg b.w. in olive oil) twice a week for a duration of 2, 6 or 12 months to induce pericentral damage. The remaining 18 mice form the control group and were treated only with olive oil in the same way as the CCl_4-treated mice

over a duration of 0, 6 or 12 months. The samples were captured 6 days after the last injection. Overall, after pre-processing both omics-data, 1,246 clear gene-protein-pairs form joint sub-data sets including both the genes from the gene expression data set and the proteins from the protein data set. It comprises information of the pairs' associated *gene expression values* and *protein expressions* as well as information on the *treatment* (CCl_4/oil) and *duration* ($\in \{0, 2, 6, 12\}$ months). The gene expression values do not contain any missing values. However, not all 36 mice cover full information for the protein data. Protein expressions contain systematical and unsystematical missing values, leading to a maximum of 31 mice with complete information per gene-protein-pair, which is especially relevant for the later regression analyses, as the number of parameters to be estimated can not exceed the number of observations.

4 Main Results

To take an initial look at the gene-protein-data and the relation between gene expressions and protein levels, the Pearson correlations between those variables are briefly investigated. Figure 1 shows the empirical Pearson correlations of all gene-protein-pairs across all mice as well as separately for both treatment groups *oil* and CCl_4. The majority of gene-protein-pairs is positively correlated, but as mentioned in the introduction, by far not all of them are highly correlated and there is also a non-negligible amount of negatively correlated pairs. It is also striking that the empirical correlation overall rises for mice that are induced CCl_4. In contrast, the correlation for *oil*-treated mice seems to be centered around zero.

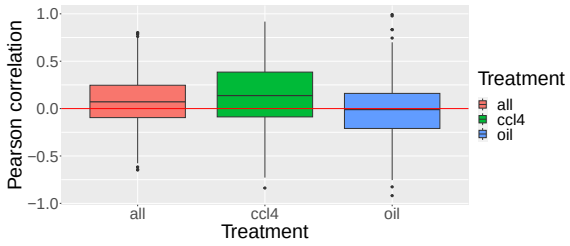


FIGURE 1. Boxplots of the Pearson correlations of all gene-protein-pairs for all mice samples and separately for CCl_4-treated and *oil*-treated mice.

In the next step, the co-expression analysis with a chosen optimal power level of $\delta = 1$ yields a total of 6 gene groups with the smallest group only containing 12 genes and the largest group consisting of 451 genes. To obtain an overview, all group sizes are listed in Table 2. Proceeding to the statistical modeling of the protein data, the goodness-of-fit results are displayed in Table 1. Note that each gene-protein-pair sub-data set is modeled

separately. Thus, the measures are averaged across all pairs. The results of model (1) show that, averaged across all gene-protein-pairs, only a small percentage of variability in the protein expressions can be solely explained by the respective gene expression values. When additionally considering the experiment settings *Treatment (T)* and *Duration (D)* as covariates, the explained variance by model (2) is larger and the AIC also indicates a better model fit, however, the out-of-sample prediction error is roughly the same. This model can be seen as the baseline model for the gene-protein-modeling in this analysis. It also suggests the inclusion of the experiment settings in the further modeling. Moreover, including the information of the co-expression analysis and using the regularized LASSO regression increases the overall quality of the model fits. Finally, when additionally including the protein levels of the grouped gene-protein-pairs as covariates in model (4), it is apparent that both the adjusted R^2 and the prediction error indicate the best fit. In contrast, the AIC on average favors model (3) with only including the grouped genes' expressions as covariates over model (4). However, as proteins might not actually affect each other in general and the results might be influenced by coincidental collinearity issues between the investigated protein and the other predictor variables, the biologically reliable choice might be to select model (3) here. When taking a closer look at the results of model (3) separately for all gene-protein-pairs, it might be of interest, which pairs achieve the highest model fit improvements when including grouping information, i.e. when comparing model (3) to the baseline model (2). Overall, out of the 1,246 pairs, the models of only 6 pairs obtain worse AIC values when considering grouping information. Particularly, the gene-protein-pairs *Glul-P15105, Dnajc3-Q91YW3, Krt1-P04104* and *Gda-Q9R111* reveal the highest improvement among all pairs.

TABLE 1. Model comparison for regression models with target variable *Protein Expression (PE)* based on the mean adjusted R^2, the mean prediction error (RMSE) and the mean AIC.

	adjusted R^2	RMSE	AIC
Model (1)	0.070	1.377	52.095
Model (2)	0.289	1.400	51.094
Model (3)	0.340	1.153	**-79.689**
Model (4)	**0.793**	**0.917**	-73.130

TABLE 2. Group sizes of the 6 groups resulting from the co-expression analysis.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Size	385	208	12	34	451	42

Further Annotations

Note that this project is still work in progress. Thus, the approaches above might be elaborated further and the real data application might be substantiated by a simulation study. Furthermore, the modeling of proteins via multiple genes might be somehow artificial, as variable selection is not directly of interest because, in theory, the protein values mainly depend on the gene counts of the corresponding gene. Thus, this modeling of proteins via grouped genes might violate biological correctness. Altogether, within the presented approach, it is possible to modify the investigation by considering other target variables while keeping the main approach of grouping omics data and include them as covariates. Additionally, this approach could be extended by e.g. considering overlapping groups. One strategy could for example be the implementation of these overlaps in order to construct an extension of the sparse group LASSO (Simon et al., 2013).

Acknowledgments: This work has been supported (in part) by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, Project I5) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation - Project Number 4278-06116).

References

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, **24(6)**, 2350-2383.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33(1)**, 1-22.

Ghallab, A.J., Myllys, M., Holland, C.H., Zaza, A., Murad, W., Hassan, R., Ahmed, Y.A., Abbas, T., Abdelrahim, E.A., Schneider, K.M., Matz-Soja, M., Reinders, J., Gebhardt, R., Berres, M.-L., Hatting, M., Drasdo, D., Saez-Rodriguez, J., Trautwein, C., and Hengstler, J.G. (2019). Influence of Liver Fibrosis on Lobular Zonation *Cells*, **8.12**, 1556.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics*, **22(2)**, 231-245.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58(1)**, 267-288.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, **4(1)**.

A new scalar-on-function generalized additive model for partially observed curves: an application to aneurysm patients

Pavel Hernández-Amaro[1], Maria Durban[1], M. Carmen Aguilera-Morillo[2]

[1] University Carlos III de Madrid, Spain
[2] Universitat Politècnica de València, Spain

E-mail for correspondence: pahernan@est-econ.uc3m.es

Abstract: In this work we present a novel methodology to fit a generalized functional regression model for partially observed functional data avoiding the curves reconstruction and assuming the basis representation of both, the functional coefficient and the functional covariate. The model's coefficients are estimated via Penalized Quasi-likelihood using the mixed model representation of a penalized spline. We test our methodology in a real classification problem with a data set of aneurysm patients.

Keywords: Partially observed functional data, Generalized scalar-on-function regression model, B-splines.

1 Introduction

Functional data analysis is one of the fastest growing fields in statistical analysis. Modern data sets often consist of complex objects, such as functions. Functional data is usually found as discrete and often noisy observations of the true underlying function, measured at different locations in time, space, or other continuum. In most cases it is assumed that all functions are observed over the full extension of their domain. However, in many real data sets, each curve is observed in a subset of the domain, which may even be different for each curve. This type of data is know as partially observed functional data.
In this work we present a new methodology to fit a generalized scalar-on-function regression model to deal with this type of data. The proposed

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

functional model considers each curve only within its observed subset of the domain; also a penalty is added to the estimation of the functional coefficient in order to control its smoothness trough the smoothing parameter. Additionally a basis representation of the functional data and the functional coefficient of the model is assumed. This representation allows us to transform the functional model into a mixed effect model and then estimate directly all the model coefficients including the smoothing parameter. We use B-spline basis for our representations but other suitable basis can be chosen. The performance of the proposed model is tested on a real classification problem.

2 Methodology

Given the following sample data: $\{Y_i, \boldsymbol{C}_i, X_i(t)\}$, $i = 1, \ldots, N, t \in T$, where $\boldsymbol{C}_i$ are the non-functional covariates, $X_i(t)$ is the functional covariate with sample observations $x_{ij} = X(t_{ij})$ and with observation points that falls in a subset of the domain T, i.e., $t_{ij} \in [d_i, T_i] = D_i \subseteq T$. The response variable Y_i follows an exponential family distribution with mean μ_i. The propose model is:

$$\eta_i = g(\mu_i) = \alpha + \boldsymbol{C}_i\boldsymbol{\gamma} + \frac{1}{D_i}\int_{D_i} X_i(t)\beta(t)\,\mathrm{d}t, \quad t \in D_i. \tag{1}$$

This model considers each sample curve $X_i(t)$ only in its domain and hence variable integration limits for each curve are considered.

2.1 Sample estimation

In order to estimate (1) the first step is to consider the basis representation of the functional covariate and the functional coefficient:

$$X_i(t) = \sum_{j=1}^{p} a_{ij}\phi_{ij}(t) = \boldsymbol{\phi}_i^{\mathrm{T}}(t)\boldsymbol{a}_i,$$

$$\beta(t) = \sum_{k=1}^{q} b_k\varphi_k(t) = \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{b},$$

where $\boldsymbol{\phi}_i(t)$, $\boldsymbol{\varphi}(t)$ are the basis used in the representation of the functional covariate $X_i(t)$ and the functional coefficient $\beta(t)$, respectively, with $\boldsymbol{a}_i$ and $\boldsymbol{b}$ the respective basis coefficients. Cubic B-splines basis has been considered in both basis representations.

By assuming these representations, the model (1) is transformed into a multivariate regression model:

$$\begin{aligned}\boldsymbol{\eta} &= \boldsymbol{\alpha} + \boldsymbol{C}\boldsymbol{\gamma} + \tfrac{1}{\boldsymbol{D}}\textstyle\int_{\boldsymbol{D}} X(t)\beta(t)\,\mathrm{d}t \\ &= \boldsymbol{\alpha} + \boldsymbol{C}\boldsymbol{\gamma} + \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{b} = \boldsymbol{B}\boldsymbol{\theta},\end{aligned}$$

with $(\boldsymbol{A})_{N\times Np}$ a block diagonal matrix, which i-th block of the diagonal is $\boldsymbol{a}_i^T$, and $(\boldsymbol{\Psi})_{Np\times q} = (\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_N)^T$, where $\boldsymbol{\Psi}_i = \frac{1}{D_i}\int_{D_i} \boldsymbol{\phi}_i(t)^{\mathrm{T}}\boldsymbol{\varphi}(t)\,\mathrm{d}t$.

In order to correctly calculate the matrix of inner product $\boldsymbol{\Psi}$ the basis $\boldsymbol{\varphi}(t)$ must be carefully constructed. Notice that the domain of this basis corresponds with the full domain T, but the limits of integration of every block matrix of inner product in $\boldsymbol{\Psi}$ are, in principle, smaller than T. Then for the calculation of this inner product matrix the basis $\boldsymbol{\phi}_i(t)$ is being multiplied by a basis resulting of selecting the corresponding knots of the basis $\boldsymbol{\varphi}(t)$ that falls inside the domain D_i, i.e., the number of knots selected of the basis $\boldsymbol{\varphi}(t)$ varies in every inner product matrix $\boldsymbol{\Psi}_i$.

The resulting multivariate regression model falls into the category of generalized linear models and therefore we use the maximum likelihood method in order to estimate the model parameters. Since the functional coefficient has been represented using a B-spline basis, the smoothness of the resulting estimated coefficient is determined by the basis dimension. To avoid the problem of choosing the optimal number of basis functions a penalized likelihood approach is considered. The final penalized likelihood equation is $L_p(\boldsymbol{\theta}, \boldsymbol{y}) = L(\boldsymbol{\theta}, \boldsymbol{y}) - \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\theta}$, where $L(\boldsymbol{\theta}, \boldsymbol{y})$ is the likelihood of $\boldsymbol{Y}$ and $\boldsymbol{P}$ is the penalty term, based on differences of adjacent B-splines coefficients. We take here this second approach, resulting in a penalty matrix $\boldsymbol{P} = \lambda_t(\boldsymbol{D}_q^{\mathrm{T}}\boldsymbol{D}_q)$.

In order to efficiently estimate the smoothing parameters λ_t together with the rest of the parameters in the model, the proposed model is reparametrized as a mixed model. Therefore, we are in the context of Generalized Linear Mixed Models (GLMMs) and the model estimation is carried out by Penalized Quasi-Likelihood (Breslow N. E. et al, 1993). To speed up computations, the SOP(Separation of Overlapping Penalties) algorithm (Rodríguez-Álvarez, M. et al., 2019) has been used.

3 Real data application

The AneuRisk65 data set https://statistics.mox.polimi.it/aneurisk consists of profiles of radius and curvature of the internal carotid artery of 65 subjects suspected to be affected by cerebral aneurysms, the data is shown in Figure 1 where can be seen that the domain where each curve is observed varies across subjects.

The goal is to classify each patient into one of two groups depending on the presence and location of the aneurysms (Stefanucci, M. et al, 2018) and for that the following functional logistic regression model is proposed:

$$\eta = \log(\frac{\pi}{\mathbf{1}-\pi}) = \alpha + \frac{1}{\mathbf{D_1}}\int_{\mathbf{D_1}} X_1(t)\beta_1(t)\,\mathrm{d}t + \frac{1}{\mathbf{D_2}}\int_{\mathbf{D_2}} X_2(t)\beta_2(t)\,\mathrm{d}t,$$

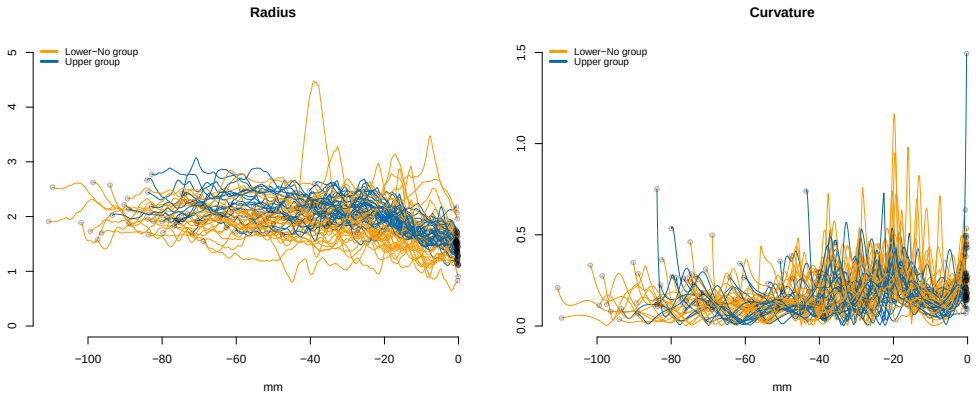


FIGURE 1. Radius (left) and curvature (right) of the internal carotid arteries of 65 subjects. The circles indicate the starting and ending points of each curve. Two different colors are used for subjects in the Upper group (blue) and subjects in the Lower–No group (orange).

where $X_i(t)$ and $\beta_i(t)$, $i = 1, 2$ are the functional covariate and coefficient corresponding with the radius and curvature, respectively. Notice that the previous model presents two functional covariates and no non-functional covariates, but the application of the proposed methodology is straightforward, with the design matrix of the multivariate model being $\boldsymbol{B} = [1|\boldsymbol{B}_1|\boldsymbol{B}_2]$, where $\boldsymbol{B}_i = \boldsymbol{A}_i \cdot \boldsymbol{\Psi}_i$, $i = 1, 2$ and the coefficient vector is $\boldsymbol{\theta} = [\alpha|\boldsymbol{b}_1|\boldsymbol{b}_2]^{\mathrm{T}}$.

As proposed by Stefanucci, M. et al (2018) the data set was analyzed by splitting the full domain T into different portions T_l. These portions of the domain go from the common domain T_1 to the full domain T and each consecutive portion contains the previous one, i.e., $T_1 \subset T_2 \subset \ldots \subset T$. Then, For each of these portions, the proposed methodology was applied.

The errors have been calculated using a Leave-on-out classification, and the optimal cut-off point for the model has been calculated by selecting between the following two criteria the one that minimizes the miss-classification error.

- **Criterion 1: Maximize the sum of the sensitivity and the specificity.**
- **Criterion 2: Maximize the sum of the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV).**

According to Christensen, E. (2009) criterion number 2 is more useful when

dealing with medical problems and shows better results in an out-sampling error measure as the leave-one-out classification. In fact, the best results in our data set had always been achieved when using this criterion. But this criterion is not perfect since it is influenced by the distribution of the data and according to Trevethan, R. (2017) it should not be used when data presents a big proportion of positive cases and a small proportion of negative cases or vice versa (unbalanced data). This problem is not present in our data set since the number of positive cases is 33 and the number of negative cases is 32. The smallest error obtained was 11 and corresponds to the full domain, thes results are shown in Figure 2.

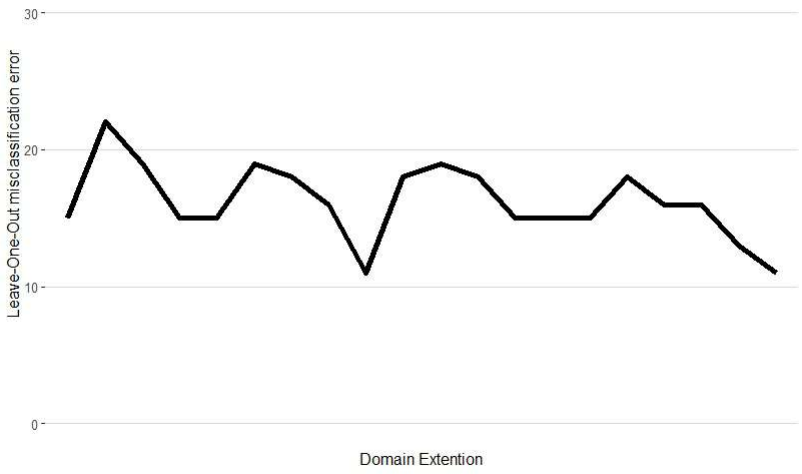


FIGURE 2. Leave-one-out miss-classification error for various domain extensions

Acknowledgments: This work is supported by the grant ID2019-104901RB-I00 from the Spanish Ministry of Science, Innovation and Universities, MCIN/AEI/10.13039/501100011033.

References

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, **88**, 9 – 25.

Christensen, E. (2009). Methodology of Diagnostic Tests in Hepatology. *Annals of Hepatology*, **8(3)**, 177 – 183.

Rodríguez-Álvarez, M. X., Durban, M. and Lee, D-J. and Eilers, P. (2019). On the estimation of variance parameters in non-standard generalised linear mixed models: Application to penalised smoothing. *Statistics and Computing*, **29** , 483 – 500.

Stefanucci, Marco, Sangalli, Laura M. and Brutti, Pierpaolo. (2018, 8). *PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set*. Statistica Neerlandica 72(3), 246–264.

Trevethan, R. (2017). Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health*, **5**, 307.

Detecting heterogeneity of treatment effect between centers in multicenter randomized clinical trials

Sebastiaan Höppner[1], Marc Buyse[1,2,3], Laura Trotta[1]

[1] CluePoints S.A., Louvain-la-Neuve, Belgium
[2] International Drug Development Institute (IDDI), Louvain-la-Neuve, Belgium
[3] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, Belgium

E-mail for correspondence: sebastiaan.hoppner@cluepoints.com

Abstract: Multicenter randomized clinical trials are commonly used in medical research to test the effectiveness of interventions across multiple treatment centers. However, treatment effects may vary between centers due to differences in patient characteristics, clinical practices, or other factors. This variation in treatment effect, known as treatment effect heterogeneity, may affect the validity and generalizability of study findings as it may lead to biased treatment effect estimates. Thus, detecting treatment effect heterogeneity between centers is crucial in the analysis of these studies. Our proposed method involves modeling each individual outcome variable using a generalized linear mixed-effects model. The heterogeneity of a variable's treatment effect is then assessed by estimating a P-value for each center, testing whether the center's treatment effect deviates significantly from the study-wide treatment effect. Each center's collection of P-values across multiple outcome variables is summarized in a single statistic, called the Treatment Effect Inconsistency Score (TEIS). The center's TEIS is a significance probability estimated using a resampling strategy that down-weights highly correlated tests. The database of a large randomized clinical trial with known fraud was analyzed with the aim to detect the fraudulent center having an atypical treatment effect for multiple variables.

Keywords: Treatment effect; Multicenter clinical trial; Generalized linear mixed-effects model.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

1 Testing for treatment effect consistency

Consider a multicenter randomized clinical trial with N_c centers and center j contains n_j patients where $j = 1, \ldots, N_c$ indexes centers. Suppose outcome data are collected for each patient on N_v outcome variables such as a response to treatment, a toxicity to treatment, etc. Let $Y_{ij}^{(k)}$ denote the kth outcome for patient i in center j where $i = 1, \ldots, n_j$ indexes patients and $k = 1, \ldots, N_v$ indexes variables. Variable T_{ij} is the treatment group indicator of the ith patient with $T = 0$ for the control treatment and $T = 1$ for the experimental treatment.

First, the study-wide treatment effect on each outcome variable $Y^{(k)}$ is estimated by fitting a generalized linear mixed-effects model with link function $g(\cdot)$, e.g. linear or logistic, depending on the type of variable. The regression model includes a fixed treatment effect and a random center effect to account for variability in the outcome among centers:

$$g\left(Y_{ij}^{(k)}\right) = \alpha + \beta T_{ij} + s_j + \varepsilon_{ij}, \quad s_j \sim N\left(0, \sigma_c^2\right), \quad \varepsilon_{ij} \sim N\left(0, \sigma_p^2\right)$$

for each variable $k = 1, \ldots N_v$. By fitting this model, we obtain estimates for the mean outcome of the control group (α), the study-wide treatment effect on the kth outcome (β), and the random effect (s_j) for center j which is normally distributed with variance σ_c^2, independently from the residual errors $\varepsilon_{ij} \sim N\left(0, \sigma_p^2\right)$.

Next, the treatment effect te_j is measured in each individual center j for each outcome $Y^{(k)}$. For example, if $Y^{(k)}$ is a continuous variable, we compute the sample mean of the control group ($m_{j,0}$), resp. treatment group ($m_{j,1}$), in center j and measure the treatment effect in center j as the difference between the means: $te_j = m_{j,1} - m_{j,0}$. If $Y^{(k)} \in \{0, 1\}$ is a binary variable, we compute the sample odds of $Y^{(k)} = 1$ for the control group ($o_{j,0}$), resp. treatment group ($o_{j,1}$), in center j and measure the treatment effect in center j as the log odds ratio: $te_j = \log\left(o_{j,1}/o_{j,0}\right)$. If the treatment effect on an outcome variable would be consistent across the centers, one expects $te_j = \beta$ for each center $j = 1, \ldots, N_c$.

We test in particular for centers that have a *weaker* treatment effect on variable $Y^{(k)}$ compared to the study-wide treatment effect through the P-values of the one-tailed test, taking into account the direction of the study-wide treatment effect (i.e. the sign of β)

$$H_0 : sign(\beta) \cdot te_j \geqslant sign(\beta) \cdot \beta \quad \text{vs} \quad H_1 : sign(\beta) \cdot te_j < sign(\beta) \cdot \beta$$

for each center $j = 1, \ldots, N_c$ and each outcome $k = 1, \ldots, N_v$.

The consistency of treatment effect on the N_v outcome variables across the N_c centers is more efficiently assessed through a summary score for each center using the tests performed on all outcome variables. The Treatment Effect Inconsistency Score (TEIS) is calculated for each center j ($j = 1, \ldots, N_c$) (Trotta, 2019):

1. Let p_{jk} be the P-value associated with center j for the kth test ($k = 1, \ldots, N_v$). First, the score S_j for center j is calculated as

$$S_j = \frac{1}{\sum_{k=1}^{N_v} w_k} \sum_{k=1}^{N_v} w_k \log p_{jk}$$

   where the weights w_k account for the correlation between the tests.

2. TEIS of center j is the significance probability P_j, on a log-scale, assigned to score S_j using a resampling method: $TEIS_j = -\log_{10}(P_j)$

A TEIS of $1.3 = -\log_{10}(0.05)$ or larger corresponds to an overall P-value (P_j) less than 0.05. As such, the treatment effect in a center is identified as atypical if its TEIS > 1.3, considering the tests for all outcome variables.

2 Multicenter clinical trial with known fraud

We used patient-level clinical data from a published trial to demonstrate how our method identifies centers with atypical treatment effect. In one of the 59 centers in the trial, it was later found that most of the 219 patients randomized in that center had never received the study therapies. (Hoeksema et al., 2000). The dataset includes 1758 placebo patients and 1760 treatment patients. We analyzed the treatment effect on 3 continuous variables and 7 binary variables (e.g. "Did the patient have a stroke during the trial? Yes or no"). The goal was to find out if our method could identify the known fraudulent center. First, we apply our test to each of the 10 variabes with the aim of detecting centers with weaker treatment effect compared to

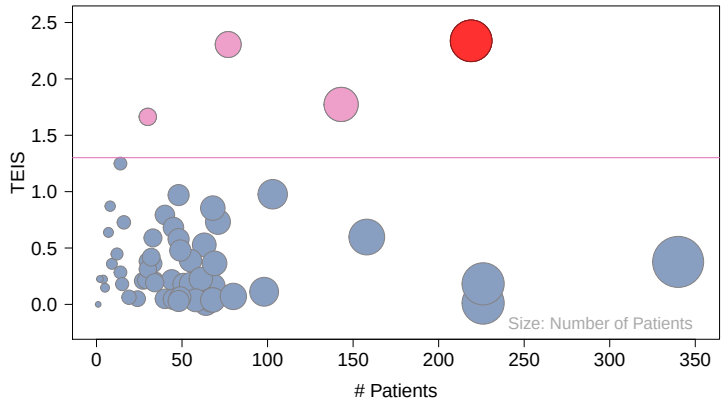


FIGURE 1. Bubble plot showing 4 centers with a Treatment Effect Inconsistency Score (TEIS) > 1.3 (overall P-value < 0.05). Center with known fraud in red. The size of each bubble is proportional to the number of patients in the center.

the study-wide treatment effect. As such, a P-value is estimated for each of the 59 centers for each of the 10 variables. Each center's set of 10 P-values is then summarized in its Treatment Effect Inconsistency Score (TEIS). Figure 1 shows the "bubble plot" with each bubble positioned according to the number of patients (x-axis) and TEIS (y-axis) of a center. Of the 59 centers, 4 are identified as having an atypical treatment effect (TEIS > 1.3, magenta bubbles), including the known fraudulent center (red bubble) having the highest TEIS $= 2.34$ (overall P-value of 0.0046). The P-values of the fraudulent center are presented in Table 1 showing that the treatment effect is significantly weaker on erythrocytes in particular.

TABLE 1. P-values of the fraudulent center.

Erythrocytes	Hematocrit	Hemoglobin	Adverse event	Stroke
0.00048	0.063	0.196	0.314	0.439
TIA	Complaints	Gastric pain	Headache	Bleeding
0.786	0.219	0.015	0.098	0.064

Figure 2 shows the estimated treatment effect (te_j) on erythrocytes (y-axis) versus the number of patients (x-axis) in each center. The figure shows 3 centers with an atypical weak treatment effect on erythrocytes (P-value < 0.05, magenta bubbles), including the fraudulent center (red bubble).

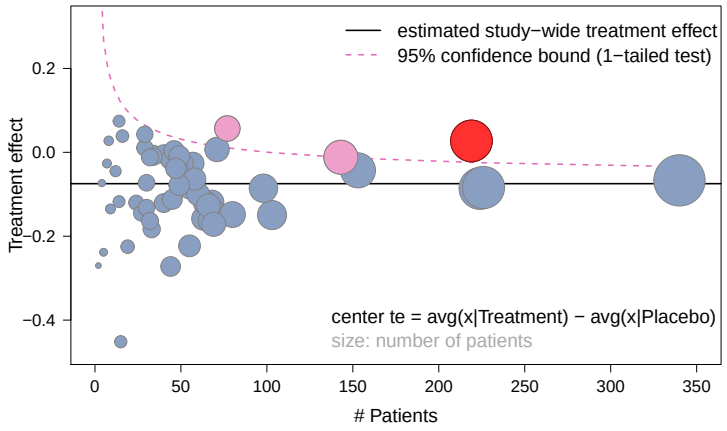


FIGURE 2. Treatment effect on erythrocytes vs. center size.

References

Hoeksema, H.L, et al. (2000). Fraud in a pharmaceutical trial. *The Lancet*, **356**(9243):1773.

Trotta, L., et al. (2019). Detection of atypical data in multicenter clinical trials using unsupervised statistical monitoring. *Clinical Trials*, **16**(5):512 – 522.

Rate of return to education of compliers: Estimation based on Rubin Causal models

Caizhu Huang[1], Jierui Du[2], Claudia Di Caterina[3]

[1] Department of Statistical Sciences, University of Padova, 35121, Padova, Italy
[2] School of Economics and Statistics, Guangzhou University, China
[3] Department of Economics, University of Verona, Italy

E-mail for correspondence: caizhu.huang@phd.unipd.it

Abstract: Randomized experiments cannot determine the impact of migrant work on the rate of return to education due to the inability to randomly assign rural populations to rural and urban households. Under Rubin causal models, we evaluate the effect of migrant work on the rate of return to education of compliers by estimating the complier average causal effect (CACE) parameter. Our analysis focuses on data from rural and migrant residents in China in 2013, and we construct estimators of unknown parameters and the Mincer earnings function using a linear combination of polynomial spline functions. The empirical results indicate that migrant work increases the rate of return to education of compliers by 3.57%. These results provide a scientific evaluation of the social and economic value of migrant work during the economic transition period, from a human capital perspective.

Keywords: migrant work; rate of return to education; Complier Average Causal Effect; partial linear models; gender differences.

1 Introduction

Estimating the rate of return to education is a fundamental task in educational economics. The rate of return to education is a measure of the future net economic remuneration for an educated individual caused by one extra year of schooling. However, since the rural population cannot be randomly assigned to rural households and rural-to-urban migrant households, it is not possible to perform a randomized experiment to obtain the effect of migrant work on the rate of return to education. The Complier Average Causal Effect (CACE) parameter obtained under the principal stratification framework is a popular method to solve this kind of problem (Frangakis and Rubin, 2002). The Mincer earnings function (Mincer, 1974)

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

is the most widely used model for estimating the rate of return to education (Polacheck, 2008).

Based on Rubin causal models, empirically evaluates the impact of migrant work on the rate of return to education using a dataset from the 2013 wave of the Chinese Household Income Project (CHIP) survey. In the Mincer earnings function, we replace the linear relationship between work experience and the log of income with an unknown smooth function, which can be approximated by a polynomial spline. Our results show that migrant work can increase the rate of return to education of the compliers by 3.57%. At the micro level, our research provides scientific evidences for the need of educational investments in favor of the Chinese rural population.

2 Notation and Assumptions

The Mincer earnings function (Mincer, 1974) is a single-equation model that explains wage income as a function of schooling and work experience. We here consider the model as

$$\log(1+E) = \beta_0 + \beta_1 S + g(Exper) + \varepsilon,$$

where E represents earnings (in CNY), S indicates the years of schooling. $Exper$ stands for the work experience computed as age $- S - 6$ (Du et al., 2023), and ε is an unobserved random error with mean 0 and variance σ^2. For the ith individual, let D_i denote the treatment received. Specifically, $D_i = 1$ if the ith individual is a migrant resident, while $D_i = 0$ if the ith individual is a rural resident; and Z_i denote the randomized treatment assignment. Based on China's geography and a preliminary data screening, individuals living closer to Beijing are supposed to be more likely to move for labor. Thus, exploiting the information from the Chinese administrative division, we set $Z_i = 0$ if the ith individual's province belongs to the Northeast or North China regions, while we set $Z_i = 1$ if the ith individual's province is in East China, Central China, South China, Southwest China or Northwest China. Assuming that n individuals are independent, a random sample $\{(D_i, Z_i, S_i, Exper_i, E_i), i = 1, \ldots, n\}$ can be obtained. For simplicity of notation, we shall use $X = (S, Exper)^T$ and $Y = \log(1+E)$. Moreover, let $D_i(z)$ and $Y_i(z)$ denote the potential treatment received and the potential outcome for the ith individual under treatment $Z = z$, respectively.

We refer to the "principal stratification" framework proposed by Angrist et al. (1996) and Frangakis and Rubin (2002) to analyze causal effects. Let U_i be the compliance status of the ith unit, defined as follows: $U_i = c$ if $D_i(0) = 0$ and $D_i(1) = 1$; $U_i = n$ if $D_i(0) = 0$ and $D_i(1) = 0$; $U_i = a$ if $D_i(0) = 1$ and $D_i(1) = 1$; and $U_i = d$ if $D_i(0) = 1$ and $D_i(1) = 0$. The values c, n, a, and d stand for complier, never-taker, always-taker, and defier, respectively. Here we consider the CACE(x), which equals

$$\text{CACE}(x) = E\{Y(1) - Y(0) | U = c, X = x\}$$

where $Y(1)$ represents the potential outcome if $Z = 1$, and $Y(0)$ represents the potential outcome if $Z = 0$.
In order to guarantee the CACE(x) identifiable, some sufficient conditions on the latent variables are needed. There are seven basic assumptions coming from Angrist et al. (1996); Zigler and Belin (2011) and Chen et al. (2015). We are not specified here.
To estimate CACE(x), we proposed following three steps procedure given the random sample $\{(D_i, Z_i, X_i, Y_i), i = 1, \cdots, n\}$:

Step 1: Obtain $P(Z = 1)$, $P(U = n|X = x)$ and $P(U = a|X = x)$

If $n_z = \sharp\{i : Z_i = 1\}$, then $P(Z = 1) = n_1/n$. Following to Frangakis and Rubin (2002), Barnard et al. (2003) and Zigler and Belin (2011), we are able to obtain $P(U = u|X = x), u = n, a, c$, where

$$\Psi_n(X) = P(U = n|X = x) = 1 - \Phi(\boldsymbol{\delta}_n^\top \tilde{X}),$$
$$\Psi_a(X) = P(U = a|X = x) = \{1 - \Psi_n(X)\}\Phi(\boldsymbol{\delta}_a^\top \tilde{X}) = \Phi(\boldsymbol{\delta}_n^\top \tilde{X})\Phi(\boldsymbol{\delta}_a^\top \tilde{X}),$$
$$\Psi_c(X) = P(U = c|X = x) = 1 - \Psi_n(X) - \Psi_a(X),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Since year of schooling and work experience are both non-negative and large in value, we consider the logarithm effect and let $\tilde{X}_i = \{1, \log S_i, \log(Exper_i)\}^\top$.

Step 2: Obtain $(\hat{\boldsymbol{\beta}}_n, \hat{g}_n, \hat{\boldsymbol{\sigma}}_n^2)$ and $(\hat{\boldsymbol{\beta}}_a, \hat{g}_a, \hat{\boldsymbol{\sigma}}_a^2)$

The B-spline approximation of $g(Exper)$ can be expressed as

$$\tilde{g}(Exper) = \sum_{j=1}^{N_n} \gamma_j B_j(Exper) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N_n} \gamma_j B_j(Exper_i).$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_{N_n})^\top$ is the N_n-dimensional vector of coefficients, given that the number of knots N_n satisfies $n^{1/4} \ll N_n \ll n^{1/2}$. Henceforth, we denote $\tilde{g}$(Exper) by $\boldsymbol{\gamma}\top\mathbf{B}$(Exper). According to the compound exclusion restrictions, we denote $p_n(y|x; \boldsymbol{\beta}_n, \boldsymbol{\gamma}_n, \sigma_n^2) = p_{1n}(y|x; \boldsymbol{\beta}_{1n}, \boldsymbol{\gamma}_{1n}, \sigma_{1n}^2) = p_{0n}(y|x; \boldsymbol{\beta}_{0n}, \boldsymbol{\gamma}_{0n}, \sigma_{0n}^2)$ and compute the likelihood function for the never-taker with $(Z_i, D_i) = (1, 0)$ as

$$L_{10}(\boldsymbol{\beta}_n, \boldsymbol{\gamma}_n, \sigma_n^2) = \prod_{i:(Z_i,D_i)=(1,0)} p_n(y|x; \boldsymbol{\beta}_n, \boldsymbol{\gamma}_n, \sigma_n^2),$$

and find the estimators $\hat{\boldsymbol{\beta}}_n$, $\hat{g}_n$ and $\hat{\sigma}_n^2$. Likewise, we have $p_a(y|x; \boldsymbol{\beta}_a, \boldsymbol{\gamma}_a, \sigma_a^2)$ $= p_{1a}(y|x; \boldsymbol{\beta}_{1a}, \boldsymbol{\gamma}_{1a}, \sigma_{1a}^2) = p_{0a}(y|x; \boldsymbol{\beta}_{0a}, \boldsymbol{\gamma}_{0a}, \sigma_{0a}^2)$ and can obtain the estimator $\hat{\boldsymbol{\beta}}_a$, $\hat{g}_a$ and $\hat{\sigma}_a^2$ by focusing on the always-taker with $(Z_i, D_i) = (0, 1)$.

Step 3: Obtain $(\hat{\boldsymbol{\beta}}_{1c}, \hat{g}_{1c}, \hat{\boldsymbol{\sigma}}^2_{1c})$ and $(\hat{\boldsymbol{\beta}}_{0c}, \hat{g}_{0c}, \hat{\boldsymbol{\sigma}}^2_{0c})$

Finally, we maximize the likelihood function for the complier with $(Z_i, D_i) = (1, 1)$

$$L_{11}(\boldsymbol{\beta}_{1c}, \boldsymbol{\gamma}_{1c}, \sigma^2_{1c}) = \prod_{i:(Z_i,D_i)=(1,1)} \{\Psi_a(X; \hat{\boldsymbol{\delta}}_n, \hat{\boldsymbol{\delta}}_a) p_a(y|x; \hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n, \hat{\sigma}^2_n) + \Psi_c(X; \hat{\boldsymbol{\delta}}_n, \hat{\boldsymbol{\delta}}_a) p_{1c}(y|x; \boldsymbol{\beta}_{1c}, \boldsymbol{\gamma}_{1c}, \sigma^2_{1c})\},$$

to compute the estimators $\hat{\boldsymbol{\beta}}_{1c}$, $\hat{g}_{1c}$ and $\hat{\sigma}^2_{1c}$. Similarly, we get $\hat{\boldsymbol{\beta}}_{0c}$, $\hat{g}_{0c}$, $\hat{\sigma}^2_{0c}$ by focusing on $(Z_i, D_i) = (0, 0)$. Therefore, the CACE $\left(x; \hat{\boldsymbol{\beta}}_c, \hat{g}_c, \hat{\sigma}_c\right)$ with $\hat{\boldsymbol{\beta}}_c = (\hat{\boldsymbol{\beta}}_{0c}, \hat{\boldsymbol{\beta}}_{1c})^T = (\hat{\beta}_{1c,0} - \hat{\beta}_{0c,0}, \hat{\beta}_{1c,1} - \hat{\beta}_{0c,1})^T$ and σ^2_c unknown parameters can be estimated by

$$\begin{aligned} \text{CACE}\left(x; \hat{\boldsymbol{\beta}}_c, \hat{g}_c, \hat{\sigma}_c\right) &= \hat{\boldsymbol{\beta}}_{0c} + \hat{\boldsymbol{\beta}}_{1c}\text{S} + (\hat{\boldsymbol{\gamma}}^\top_{1c} - \hat{\boldsymbol{\gamma}}^\top_{0c})\mathbf{B}(\text{Exper}), \\ \hat{\sigma}^2_c &= \hat{\sigma}^2_{1c} + \hat{\sigma}^2_{0c}, \end{aligned}$$

respectively.

3 Data Analysis

TABLE 1. Descriptive statistics of the data.

Variable	Mean		Standard deviation		Minimum		Maximum	
	$D=1$	$D=0$	$D=1$	$D=0$	$D=1$	$D=0$	$D=1$	$D=0$
Logarithmic annual income	10.256	9.904	0.835	0.834	0	0	12.739	13.592
Years of schooling	9.565	8.78	2.908	2.784	0	0	19	20
Experience	19.32	20.7	10.108	11.745	4	0	44	44
Sample size	1237	15245	1237	15245	1237	15245	1237	15245

We focus here on the subsamples of migrant and rural individuals of the CHIP. In this study, rural data are based on a sample with 39,408 individuals and migrant data on a sample with 2,839 individuals. After cleaning the data, the records used for our analysis were described in Table 1, which shows the basic statistical summaries of the variables involved in the sample, where D = 1 identifies migrant units and D = 0 identifies rural units. Partial results obtained from the analysis of the CHIP data based on the proposed estimation method show that the estimated rate of return to education of the always-takers is 3.60% whereas that of the never-takers is almost three times greater, i.e. 9.10%. The estimate of the rate of return to education is 6.62% for the migrant compliers and 3.05% for the rural compliers, so that migrant work serves to increase the rate of return to education of the compliers by 3.57%.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, **91**, 444 – 455.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference *Biometrics*, **58**, 21 – 29.

Mincer, J. (1974). *Schooling, Experience and Earnings*. New York: Columbia University Press.

Zigler, C. M. and Belin, T. R. (2011). The potential for bias in principal causal effect estimation when treatment received depends on a key covariate. *The Annals of Applied Statistics*, **5**, 1876 – 1892.

Understanding the role of conditional residual distances from simulated envelopes in half normal plots

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 involves models that account for overdispersion. Several overdispersion models are contained in the generalized linear modelling framework, being extensions of the basic Poisson model. It is common to assess goodness-of-fit graphically by using half-normal plots with a simulated envelope. The envelope is such that under a well-fitting model one would expect very few points to lie outside of the envelope. However, very similar graphs may be obtained for closely related models. This paper tries to evaluate the influence of the residual points falling outside the envelope and the contribution of these points to the construction of a numerical summary for half-normal plots.

Keywords: Generalized linear models; Goodness-of-fit; Half-normal plots

1 Introduction

Ecological studies include research on the wide variety of flora and fauna on Earth. It is common to record observations in ecological studies as counts, e.g. number of species or number of animals. It is not viable to model counts using Gaussian linear regression models, since they are not a suitable method to examine non negative discrete data. Typically we analyse count data using the Poisson model or extended versions of it, to account for either under- or overdispersion. It is important, then, to check whether model assumptions are valid, and whether the observed data are a plausible realisation of the fitted distribution. Half-normal plots with a simulated

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

envelope are a useful tool in this case. Here, we propose summary statistics based on half-normal plots to aid model comparison.

1.1 Half-normal plots with a simulated envelope

Half-normal plots with a simulated envelope are created by plotting ordered residuals, in absolute value, versus the expected order statistics of a half-normal distribution:

$$\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right) \tag{1}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the Gaussian cumulative distribution function, i is the $i-$th order statistic, $1 \leq i \leq n$, and n is the sample size.
The simulated envelope is constructed by (i) simulating 99 or more response variables using the same distribution, design matrix, and fitted coefficients; (ii) re-fitting the same model to each simulated sample; (iii) calculating the same type of residuals, in absolute value and in order; and (iv) computing desired percentiles for each order statistics (usually 2.5% and 97.5%). The median is also computed, and shown as a dashed line in the plot.
Figure 1 shows half-normal plots with a simulated envelope for the normal, Poisson and negative binomial models fitted to two simulated datasets; the first from a Poisson distribution, and the second from the negative binomial distribution. We see that the normal model does not fit the data well in both instances, whereas the negative binomial model presents good performance for both simulated datasets. Our question is: is it possible to differentiate the performance of the Poisson vs. negative binomial model for the first simulated dataset, where both fit the data well? This is expected, since the negative binomial is an extension of the Poisson model, however in this case parsimony would lead one to select the Poisson model to analyse the data.

2 Methodology

We construct a statistic based on distances from the residual points to parts of the envelope $\mathcal{E}_i = \{x \in \mathcal{R} | x \in (l_i, u_i)\}$, namely the envelope median m_i, upper (u_i) and lower (l_i) bounds. The statistic is given by:

$$d = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (r_i - m_i)^2 g(b_i)^{I(r_i \in \mathcal{E}_i)},$$

where r_i is the $i-$th ordered residual and $g(b_i)$ is a function of the distance of the residual point to the boundary of the envelope:

$$b_i = \begin{cases} r_i - u_i, \text{ if } r_i > u_i \\ l_i - r_i, \text{ if } r_i < l_i \end{cases}$$

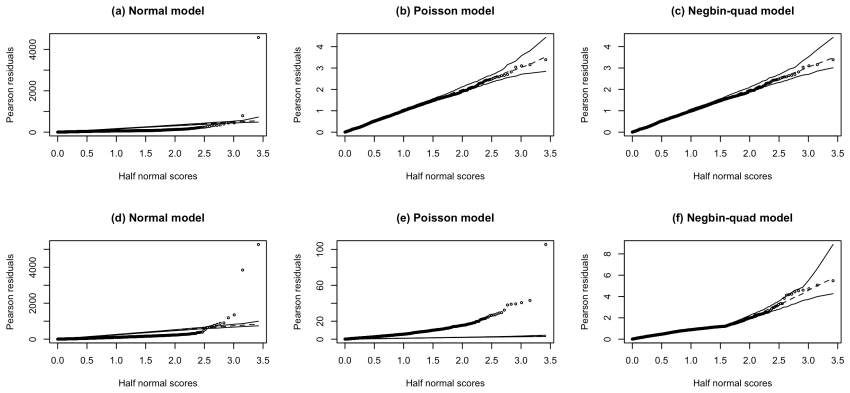


FIGURE 1. Half-normal plots with a simulated envelope for different models fitted to data simulated from a Poisson distribution (a) Normal model, (b) Poisson model and (c) Negative binomial; and from a negative binomial model with quadratic variance with a dispersion value of 1.5, (d) Normal model (e) Poisson model and (f) Negative binomial model with quadratic variance.

The indicator function $I(r_i \in \mathcal{E}_i)$ is equal to 1 if the residual point is contained in the envelope and equal to 0 otherwise, therefore the penalty function g only influences the metric if the point is outside of the envelope. We tested five different variations of $g(b)$:

- constant/no penalty: $g(b) = 1$,
- unlimited linear increase: $g(b) = \alpha + \gamma b$,
- saturated increase (ratio): $g(b) = \dfrac{\alpha + \gamma_1 b}{1 + \gamma_2 b}$,
- saturated increase (logistic): $g(b) = \dfrac{\alpha + \gamma}{1 + \exp -\delta + (b - \eta)}$ and
- saturated increase (hyperbolic tangent): $g(b) = \alpha + \gamma \tanh \delta b$.

The hyper-parameters α, β, γ, γ_1, γ_2, η and δ are assumed to be known and fixed. The penalties are introduced to differentiate, for instance, residual points that are close to either u_i or l_i, but inside the envelope (and therefore expected under the fitted model), from points barely outside of the envelope, which should be more penalised, since that would not be expected under the fitted model most of the time.

2.1 Simulation study

We carried out a simulation study with 1,000 simulated samples from each of three sample sizes (20, 50, and 100) and five parent models (Pois-

son, negative binomial with a quadratic variance function with strong and mild overdispersion, negative binomial with a linear variance function with strong and mild overdispersion). We fitted three models to each simulated sample (Poisson and negative binomial with quadratic and linear variance functions), produced a half-normal plot with a simulated envelope for the Pearson residuals and computed d.

3 Results and discussion

We found no influence of the different penalty functions $g(b)$ on the distance measure (Figure 2). This pattern is seen across all parent models considered. When the Poisson model is the parent distribution, there is no significant difference between the distance values calculated for all the fitted models, as these models behave similar to Poisson when the dispersion parameter is zero (Figure 2(a)). When simulating from overdispersed parent distributions (Figure 2(b)), we found that d shows that the Poisson model is not preferable.

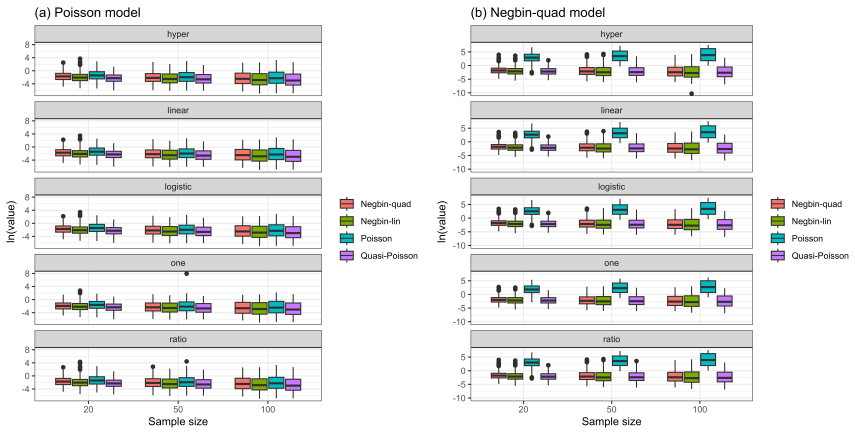


FIGURE 2. Natural logarithm of d for the five penalty functions and three sample sizes, for the (a) Poisson and (b) negative binomial with quadratic variance function and strong overdispersion parent models fitted to 1,000 simulated samples.

Acknowledgments: Special Thanks for the financial support of Science Foundation Ireland under Grant number 18/CRT/6049 for funding this project.

References

Moral, R.A., Hinde, J., Demétrio, C.G.B. (2017) *Half-normal plots and overdispersed models in R: the hnp package* Journal of Statistical Software 81(10).

Targeted bias reduction 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: An approach for targeted bias reduction of parameters of interest within a statistical model is proposed, and applied to the setting of generalised additive models, in which only parameters corresponding to linear terms are of immediate interest for bias reduction. The approach means the estimation method for the functional parameters can remain the same, allowing a more natural and efficient process than reducing the bias of all parameters. The method is tested for the case of a binomial generalised additive model via a simulation study.

Keywords: Adjusted score; bias reduction; generalised additive models; penalised likelihood; smoothing parameter.

1 Targeted bias reduction

1.1 Adjusted score equations

Within statistical research, it is often the case that an estimator of an unknown parameter $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ is biased, that is the expected value of the estimator is not equal to the parameter. As a result, bias reduction of estimators, with the aim of improving inference from statistical models, has attracted a lot of research, and there is a range of methods developed for this purpose, such as those reviewed in Kosmidis (2014). An example of a typically biased estimator is the commonly used maximum likelihood estimator, which under regularity conditions from Cox and Hinkley (1974), has bias with asymptotic order $O(n^{-1})$. As $n \to \infty$, the bias vanishes, but for finite samples the bias may be substantial. The maximum likelihood estimator is found by solving the score equations

$$S(\theta) = \nabla_\theta \ell(\theta) = 0_p$$

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

where $\ell(\theta)$ is the log-likelihood for θ, as long as the observed information matrix $I(\theta) = -\nabla_\theta \nabla_\theta^{\mathrm{T}} \ell(\theta)$ is positive definite when computed at the maximum likelihood estimator $\hat{\theta}$. Firth (1993) showed that an alternative estimator of θ with $O(n^{-2})$ bias may be found by solving the adjusted score equations

$$S^*(\theta) = S(\theta) + A(\theta) = 0_p. \tag{1}$$

The adjustments $A(\theta)$ may take various forms, and are $O_p(1)$ as $n \to \infty$. One specific form given in Firth (1993) gives the t-th element of the vector of adjustments as

$$A_t(\theta) = \frac{1}{2}\mathrm{trace}[F(\theta)^{-1}\{P_t(\theta) + Q_t(\theta)\}]$$

where $P_t(\theta) = E_\theta\{S(\theta)S(\theta)^{\mathrm{T}}S_t(\theta)\}$ and $Q_t(\theta) = -E_\theta\{I(\theta)S_t(\theta)\}$, with $F(\theta) = E_\theta\{I(\theta)\}$ being the expected information matrix.

1.2 Targeted adjustments

Consider a model where the aim is to estimate a set of parameters $\theta = (\alpha^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$, where α has dimension p_1, and β has dimension $p_2 = p - p_1$. The adjusted score equations for reduced bias estimates of θ can be decomposed in to each set of parameters as

$$S_\alpha(\hat{\alpha}, \hat{\beta}) + A_\alpha(\hat{\alpha}, \hat{\beta}) = 0_{p_1} \quad \text{and} \quad S_\beta(\hat{\alpha}, \hat{\beta}) + A_\beta(\hat{\alpha}, \hat{\beta}) = 0_{p_2},$$

where 0_ν is a vector of ν zeros. Suppose further that only the α parameters are of interest, and we wish to compute reduced bias estimates of α. One way to do this is to solve the adjusted score equations (1), but an alternative is to only adjust the score equations S_α. By letting $\bar{A}$ denote these targeted adjustments, and following a similar derivation to Firth (1993), we obtain

$$\bar{A}_\alpha = A_\alpha + \{I^{\alpha\alpha}\}^{-1} I^{\alpha\beta} A_\beta, \tag{2}$$

while S_β is not adjusted, that is $\bar{A}_\beta = 0$. We denote $I^{\alpha\alpha}$ and $I^{\alpha\beta}$ to be the (α, α) and (α, β) blocks of $I(\theta)^{-1}$ respectively.

2 Generalised additive models

Following Wood (2017), a generalised additive model (GAM) takes the form

$$g(\mu_i) = Z_i^{\mathrm{T}}\alpha + \sum_j f_j(x_{ji})$$

for $i \in \{1, \ldots, n\}$. We have that $\mu_i = E(Y_i)$, where Y_i is a response variable with an exponential family distribution, $Y_i \sim EF(\mu_i, \phi)$. The f_j are

smooth functions, each typically represented by a corresponding basis expansion $f_j(x) = \sum_{m=1}^{k} b_m^j(x)\beta_{jm}$, where $b_m^j(x)$ is the mth basis function for function $f_j(x)$. Therefore, we can construct a design matrix X^j for each f_j such that $X_{im}^j = b_m^j(x_i)$. After applying identifiability constraints, this leads to an overall design matrix $\tilde{X}$ for the functional part of the model, and a design matrix $X = (Z|\tilde{X})$ for the whole model, giving a generalised linear model (GLM) structure

$$g(\mu_i) = X_i\theta, \quad Y_i \sim EF(\mu_i, \phi),$$

where $\theta = (\alpha^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$. Maximising the likelihood $\ell(\theta)$ for this model would typically lead to overfitting, so parameter estimation is by maximising a penalised likelihood

$$\ell_p(\theta) = \ell(\theta) - \frac{1}{2\phi}\beta^{\mathrm{T}} S \beta,$$

where $S = \sum_j \lambda_j S_j$. 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).

Suppose that we wish to compute reduced bias estimates of θ. The α parameters are linear regression parameters and so reducing their bias is of interest. However, to directly reduce the bias of the β parameters is not of immediate interest, as they are representations of functions rather than individually interpretable parameters. Hence, we can apply the targeted adjustments of Section 1 in this setting. If we take $\ell_p(\theta)$ as our target to be maximised, we can apply the targeted adjustments by constructing adjustments (2), using the corresponding penalised score function $\nabla_\theta \ell_p(\theta)$.

3 A small simulation study

A binomial GAM is considered. In this setting, $Y_i \sim \text{Binomial}(m_i, \psi_i)$, and the penalised score is

$$S(\theta) = \begin{bmatrix} \nabla_\alpha \ell(\theta) \\ \nabla_\beta \ell(\theta) \end{bmatrix} = \begin{bmatrix} Z^{\mathrm{T}}(y - \mu) \\ \tilde{X}^{\mathrm{T}}(y - \mu) - S\beta \end{bmatrix}, \tag{3}$$

where $\mu_i = m_i\psi_i$, $X = (Z|\tilde{X})$ is the design matrix for the whole model and S is the overall penalty matrix. We consider the following model:

$$\log\left(\frac{\psi_i}{1 - \psi_i}\right) = \alpha_0 + x_{1i}\alpha_1 + 2\sin(\pi x_{2i}),$$

where x_1 and x_2 are independent Unif$(0, 1)$ random variables. We perform 1000 repetitions of estimating parameters by solving the original score equations (3), and the targeted adjusted score equations for $n \in \{100, 200\}$ and $\lambda \in \{0.1, 0.2, 0.8, 1.6, 3.2\}$, using the `nleqslv` package in `R`. Applying the full bias reducing adjustments required significantly more computing time than for the targeted adjustments, due to slower convergence of the optimisation algorithm, while providing similar results. Therefore, we omit the full adjustments here.

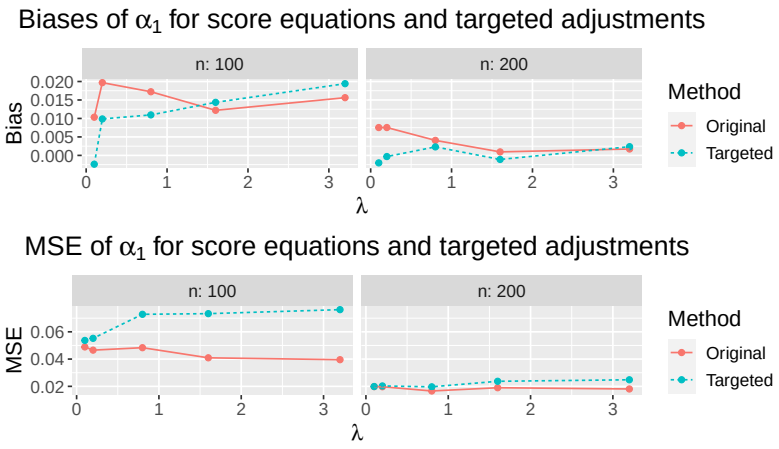


FIGURE 1. Estimated biases and mean squared errors when solving the original score equations, and score equations with targeted adjustments.

Observe that we have some evidence of bias reduction, particularly at smaller values of λ. We observe a small increase in mean squared error, and difficulty in achieving convergence for smaller n, where bias is expected to be higher. This is likely due to the numerical scheme used, therefore we are developing an iterative scheme incorporating the PIRLS approach, with the aim of increased stability, and examining the theoretical properties of the targeted bias reduction approach, in terms of the magnitude of the smoothing parameter λ. Overall, the targeted bias reduction approach shows promise and work is ongoing to develop it.

References

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall Ltd.

Firth, D. (1993). Bias reduction of maximum likeliood estimates. *Biometrika*, **80(1)**, 27 – 38.

Kosmidis, I. (2014). Bias in parametric estimation: reduction and useful side-effects. *WIREs Computational Statistics*, **6(3)**, 185 – 196.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman & Hall / CRC Texts in Statistical Science.

A novel gradient boosting framework for generalised additive mixed models

Lars Knieper[1], Elisabeth Bergherr[1], Torsten Hothorn[2], Nadia Müller-Voggel[3], Colin Griesbach[1]

[1] Chair of Spatial Data Science and Statistical Learning, Georg-August-University, Germany
[2] Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland
[3] Center for Biomagnetism Erlangen, University of Erlangen-Nurnberg, Germany

E-mail for correspondence: `lars.knieper@uni-goettingen.de`

Abstract: In this work we propose a novel gradient boosting scheme for generalised additive mixed models within the well-known *mboost* framework for model-based gradient boosting. The newly developed routine overcomes existing drawbacks like irregular selection properties and biased parameter estimation for the random structure caused by cluster constant covariates while simultaneously preserving the flexibility provided by *mboost*.

Keywords: Mixed Models; Gradient Boosting; R-Package

1 Overview

Mixed models are widely used for modelling longitudinal or clustered data by capturing within-cluster correlations using cluster-specific random effects. While usually fitted based on the penalised likelihood, model-based boosting (Bühlmann and Hothorn, 2007) offers a fast and intuitive alternative which additionally enables variable selection and stable performance in high dimensional data. For this purpose the well-known `R`-package *mboost* was equipped with a random effects base-learner in order to estimate generalised additive mixed models within the framework of component-wise gradient boosting (Kneib *et al.*, 2009). However, this approach tends to produce biased estimates in the presence of cluster-constant covariates and in addition lacks any parameter estimation for the random components.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

In this work we embed previous efforts to overcome these issues (Griesbach *et al.*, 2021) into the *mboost* framework (Hothorn *et al.*, 2022) and therefore gain a powerful boosting approach which enables well performing estimation of flexible mixed models based on gradient boosting.

2 Methodology

2.1 Generalised Additive Mixed Models

Generalised additive mixed models (GAMMs) allow the effect of each covariate $\mathbf{x}_r$, with $r = 1, \ldots, p$, to be modelled by a possibly non-linear function $f_r(\cdot)$. Therefore, the predictor $\boldsymbol{\eta}$ is defined as

$$\eta_{ij} = f_1(x_{ij1}) + \cdots + f_p(x_{ijp}) + \mathbf{z}_{ij}^{\top}\boldsymbol{\gamma}_i,$$

with clusters $i = 1, \ldots, n$ and single observations $j = 1, \ldots, n_i$. The influence of covariates $\mathbf{z}_{ij}$ in the random structure is assumed to be linear with random effects $\gamma_i \sim \mathcal{N}^{\otimes q}(0, \mathbf{Q})$, where covariance $\mathbf{Q}(\vartheta)$ depends on the unknown parameter vector ϑ. The distribution of the response variable $\mathbf{y}$ determines the necessary response function $h(\cdot)$ that achieves $\mathbb{E}(\mathbf{y}) = h(\boldsymbol{\eta})$. Resulting from the Laplace approximation the models' penalised likelihood has the following form

$$\ell^{pen}(\beta, \gamma, \vartheta) = \sum_{i=1}^{n} \log\left(f(\mathbf{y}_i|\beta, \gamma, \vartheta)\right) - \frac{1}{2}\sum_{i=1}^{n} \gamma_i^{\top}\mathbf{Q}^{-1}\gamma_i. \tag{1}$$

2.2 *mermboost* Algorithm

The *mermboost* approach essentially resembles an extension of conventional gradient boosting with nuisance parameters where now the complete random structure including random effects $\boldsymbol{\gamma}$ as well as variance components $\mathbf{Q}$ are treated as nuisance parameters and updated accordingly after every fixed effects boosting cycle. Algorithm 1 displays the *mermboost* procedure which is essentially a wrapper around *mboost* applying a novel family object with an enhanced nuisance component.

In the first step **a)** of each iteration a regular component-wise boosting update for the fixed effects is performed. Afterwards, step **b)** obtains updates for the random structure by maximizing the corresponding penalised likelihood (1) with the current fit computed in step **a)** as an offset. The updates of the random effects $\boldsymbol{\gamma}$ in step **b)** are treated with a correction mechanism as suggested by Griesbach et al. (2021) to prevent bias in their estimation, which is present in *mboost*. This separation of a component-wise procedure for fixed effects and an iterative update of the random effects eliminates competition between them and therefore, ensures an adjustment of cluster specific effects in every boosting iteration.

As this algorithm uses *mboost* the extensive variety of distribution families and base-learners of *mboost* becomes immediately available for *mermboost* enabling a highly customisable estimation approach to a broad class of mixed models via gradient boosting.

Algorithm 1 (*mermboost*)

- **Initialize** predictor $\hat{\boldsymbol{\eta}}^{[0]}$ and specify base-learners $bl_1(\cdot), \ldots, bl_p(\cdot)$. Choose step length ν and number of total iterations m_{stop}.
- **for** $m = 1$ to m_{stop} **do**

  a) fixed effects: Compute the negative gradient $\mathbf{u}^{[m]}$ of the current model, determine the best performing base-learner fit $\widehat{bl}_{r_*}$ corresponding to $\mathbf{u}^{[m]}$ and update

  $$\hat{\boldsymbol{\eta}}^{[m]} = \hat{\boldsymbol{\eta}}^{[m-1]} + \nu \widehat{bl}_{r_*}.$$

  b) random structure: Receive current estimates $\hat{\boldsymbol{\gamma}}^{[m]}$ and $\hat{\mathbf{Q}}^{[m]}$ by maximizing the penalised likelihood (1) with $\hat{\boldsymbol{\eta}}^{[m]}$ as offset.

  end for

- **Determine** the best performing stopping iteration m_* based on a pre-chosen criteria (e.g. cross-validation).

  Return model fit $\hat{\boldsymbol{\eta}}^{[m_*]}$ as well as $\hat{\boldsymbol{\gamma}}^{[m_*]}$ and $\hat{\mathbf{Q}}^{[m_*]}$.

---

3 Simulations

To evaluate the performance of the new algorithm *mermboost* in comparison to usual *mboost* estimations with regard to parameter estimation and variable selection properties several simulations got conducted. Table 1 demonstrates exemplary results of a random intercept case, so that $\mathbf{Q} = \tau^2$. Since *mboost* does not give an actual estimate for $\mathbf{Q}$ and therefore for τ, the standard deviation of the random effects is used as a proxy, which is emphasised by italic values.
Simulations with a Poisson distributed response reveal strongly improved performance of *mermboost* compared to *mboost* regarding accuracy of estimates, i.e. the mean squared error mse_β of fixed linear coefficients β as well as the mean squared error mse_τ of the random component τ.
Even though a significant shrinkage is observed for all parameter estimates with a binomial distributed response *mermboost* still outperforms *mboost* concerning mse_β. Contrarily, for the mean squared error of the random component, mse_τ, a decreased performance is indicated. Figure 1 illustrates

TABLE 1. Simulation results (mse_β, mse_τ and f.p.r.) for Poisson and binomial data for the exemplary case of $\tau = 2$ and varying dimensions p.

		mboost			mermboost		
data	p	mse_β	mse_τ	f.p.r.	mse_β	mse_τ	f.p.r.
Poisson	25	4.804	*1.890*	**0.000**	**3.421**	**0.087**	0.500
	50	4.882	*1.933*	**0.000**	**3.523**	**0.094**	0.326
	100	4.966	*1.958*	**0.000**	**3.612**	**0.096**	0.208
binomial	25	3.899	***0.718***	**0.310**	**3.446**	1.136	0.429
	50	4.024	***0.903***	**0.239**	**3.617**	1.316	0.348
	100	4.224	***1.186***	**0.182**	**3.997**	1.570	0.219

mermboost's significant shrinkage as well as *mboost*'s similar estimates for different values of τ. The latter is due to the competition between random and fixed effects in *mboost* and estimates strongly depend on the number of times the random effects' base-learner gets picked. Hence, for a large τ (e.g. $\tau = 3$) *mermboost* actually outperforms *mboost* concerning mse_τ. However, correlations between the true random effects and the estimated ones are higher for *mermboost* compared to *mboost*'s estimates for all cases. This is a result from erasing the structural bias in cluster constant fixed effects, which *mboost* tries to correct with the random effects. Nevertheless, large shrinkage is a drawback of this method, so that it will be looked into a restricted maximum likelihood approach for the random components as Tutz and Groll (2010) achieved substantially improved metrics using it. As a result of the mentioned bias, the cross-validation of *mboost* leads to much smaller false positive rates (f.p.r.) than *mermboost* for both distributions.

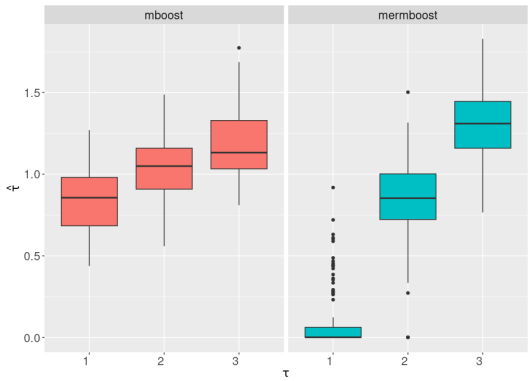


FIGURE 1. Estimations of $\tau = \{1, 2, 3\}$ by both boosting packages for binomial data ($p = 50$ is held constant here).

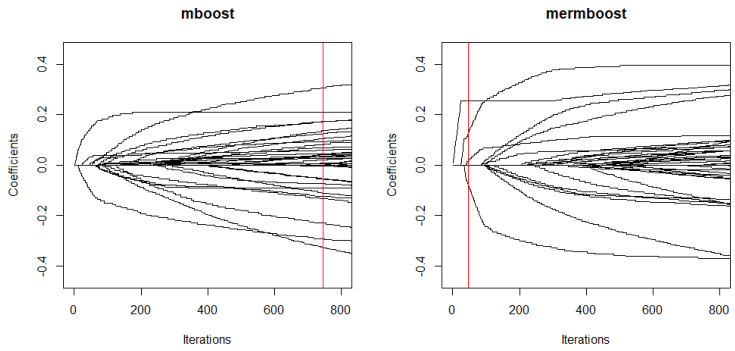


FIGURE 2. Coefficient paths of both boosting alternatives for the data example.

4 Application

We consider clustered data of 20 patients with 3656 observations overall where patients' brain activity got measured while alternating between the states of silence and a played sound. Previous neurophysiological research suggests that alpha and gamma waves show reactions to conscious perception e.g. during a played sound. By using the source points of the brain and their spatial dependence 4813 covariates of brain waves are candidates to explain whether a sound was played to a patient or not. This high dimensional clustered logit model is fitted by the old and novel approach of *mboost*. In contrast to *mermboost* the old *mboost* approach does not detect any random effects. Figure 2 demonstrates that this detection leads to a drastic influence on the optimal number of iterations (746 vs. 48, vertical red line) and consequently, on the number of picked covariates, which is 34 vs. 4. Table 2 demonstrates magnitudes of coefficients differ much from another but the four brain regions chosen by *mermboost* are also picked by *mboost*. The signs of the coefficients confirm the expectations of previous research in the way that alpha activity is high when the brain awaits cognitive perception and gamma activity can be observed during cognitive processes. Still, spatial information in form of coordinates are are yet to be received and might increase the model's performance significantly.

5 Summary

The new proposed *mermboost* algorithm incorporates a correction step as suggested by Griesbach et al. (2021) and a separation of estimating fixed and random effects. The latter ensures a removal of competition between fixed and random effects. This might be seen as a drawback since now the random effects are pre-specified instead of being picked by a statistical

TABLE 2. Picked brain regions and their coefficents estimated by both boosting approaches. (Direction in the name indicates the site of the brain. Colon indicates an interaction of neighboured regions.)

brain region	*mboost*	*mermboost*
122right-alpha	-0.291	-0.078
9left-gamma	0.210	0.131
117right-gamma:123right-gamma	0.060	0.021
149right-gamma:150right-gamma	0.281	0.255

learning algorithm. However, both adjustments result not only in unbiased estimates for cluster constant fixed effects but also in unbiased random effects with a reasonable estimate of their covariance. While simulated data with a Poisson distributed response give convincing results, for binomial data a large shrinkage is observed especially in the random structure. A restricted maximum likelihood approach might help to overcome this shrinkage as Tutz and Groll (2010) observed much better results by this within a similar framework. Furthermore, it might be of interest to look into scenarios with almost cluster constant covariates, where a similar bias would be expected but is currently not accounted for.
Currently, *mermboost* is available as an add-on `R`-package for *mboost* but to enhance its usability it is planned to integrate it in *mboost* as an own family.

References

Bühlmann, P. and Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, **22(4)**, 477 – 505.

Griesbach, C., Säfken, B., and Bergherr, E. (2021). Gradient boosting for linear mixed models. *The International Journal of Biostatistics*, **17(2)**, 317 – 329.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2022). mboost: Model-Based Boosting. *R-Package version 2.9-7.*

Kneib, T., Hothorn, T. and Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics*, **65(2)**, 197 – 215.

Tutz, G. and Groll, A. (2010). Generalized Linear Mixed Models Based on Boosting. In: Statistical Modelling and Regression Structures. Kneib, T., Tutz, G. (eds) Physica-Verlag HD, 477 – 505.

Interval-censored covariates in regression models

Klaus Langohr[1], Andrea Toloba López-Egea[1], Guadalupe Gómez Melis[1]

[1] Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain

E-mail for correspondence: klaus.langohr@upc.edu

Abstract: Interval-censored time-to-event data are common whenever the event of interest is a silent event that cannot be observed directly. Here, we present an estimation method for generalized linear models with interval-censored covariates that is applied to data from a metabolomic study.

Keywords: Interval-censored covariates; Generalized linear models; Metabolomic studies; Residual analysis.

1 Introduction

Interval censoring is typically encountered in the analysis of times to a silent event. In these cases the time that the event occurs, for instance, the moment of an infection with a certain virus, cannot be observed exactly. Methods to analyse such data have been extensively studied; see, e.g., Gómez et al. (2009). However, scientific literature on regression models with an interval-censored covariate is scarce because times to an event of interest are, most often, rather a study's response than one of the explanatory variables.

Gómez et al. (2003) presented a linear regression model with an interval-censored covariate in the context of a clinical trial for HIV-infected persons and proposed an Expectation-Maximization (EM)-type algorithm, the so-called GEL (Gómez–Espinal–Lagakos) algorithm, to jointly estimate the model parameters and the marginal distribution of the interval-censored covariate. This algorithm was implemented in R by Langohr and Gómez (2014). More recently, Morrison et al. (2022) adapted the GEL technique

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

to accommodate left-truncated interval-censored data, and Gómez et al. (2022) applied the algorithm to generalized linear models.
In this work, we present the GEL algorithm for generalized linear models and illustrate the method with a gamma regression model applied to data from a metabolomic study. In this study, one of the explanatory variable is the sum of compound concentrations that cannot be quantified exactly under the quantification limits (LoQ) or detection limits (LoD) of its components. Residuals for such models will be sketched and are, presently, under study.

2 Generalized linear models with an interval-censored covariate

2.1 Model expression and parameter estimation

The generalized linear model with an interval-censored covariate can be expressed as follows:

$$\mu = E(Y|X, Z) = g^{-1}(\alpha + \boldsymbol{\beta}' \cdot \boldsymbol{X} + \gamma \cdot Z), \tag{1}$$

where Y is a continuous or discrete response variable that belongs to the k-parameter exponential family, g is the link function, $\boldsymbol{X}$ is a p-dimensional vector of covariates, and Z is an interval-censored variable with distribution function given by F_Z. The observed intervals of Z are denoted by $[Z_l, Z_r]$. Typical examples for the distribution of Y are the binomial, Poisson, and gamma distributions leading to the logistic, Poisson, and gamma regression models, respectively.
Given an independent sample $(Y_i, \boldsymbol{X}_i, Z_{l_i}, Z_{r_i})$, $i = 1, \ldots, n$, and assuming noninformative censoring (Oller et al., 2004), the observed intervals can be treated as fixed in advanced and the likelihood is then proportional to

$$L(\boldsymbol{\theta}, F_Z) = \prod_{i=1}^{n} \int_{z_{l_i}}^{z_{r_i}} f_{Y|\boldsymbol{X},Z}(y_i|\boldsymbol{x_i}, z; \boldsymbol{\theta})\, dF_Z(z),$$

where $f_{Y|\boldsymbol{X},Z}(y|\boldsymbol{x}, z; \boldsymbol{\theta})$ denotes either the conditional density or the probability function of Y given $\boldsymbol{X}$ and Z depending on whether Y is a continuous or discrete random variable. The vector $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}', \gamma, \boldsymbol{\tau}')'$ contains the parameters of model (1) and $\boldsymbol{\tau}$ stands for the vector of extra parameters of the exponential family.
For the sake of the maximization of the corresponding log-likelihood function, we assume that Z is discrete with support $S = \{s_1, \ldots, s_m\}$ and corresponding probability masses $\boldsymbol{w} = \{w_1, \ldots, w_m\}$. Hence, the expression of the log-likelihood function is given by

$$l(\boldsymbol{\theta}, \boldsymbol{w}) = \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{m} \eta_j^i f_{Y|\boldsymbol{X},Z}(y_i|\boldsymbol{x_i}, s_j, \boldsymbol{\theta}) w_j\Big), \tag{2}$$

where $\eta_j^i = \mathbf{1}\{s_j \in [z_{l_i}, z_{r_i}]\}$ indicates whether the support point s_j is included in the observed interval of the i^{th} individual or not.
The GEL algorithm for the estimation of $\alpha, \boldsymbol{\beta}', \gamma$, and $\boldsymbol{\tau}'$ in presence of the nuisance parameter $\boldsymbol{w}$ consists of the following two-step algorithm and both steps are alternated until joint convergence is achieved:

1. Given an estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{w}}$ is obtained by solving the self-consistent equations:

$$w_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\eta_{ij} f_{Y|\boldsymbol{X},Z}(y_i \mid \boldsymbol{x_i}, s_j; \hat{\boldsymbol{\theta}}) w_j}{\sum_{k=1}^{m} \eta_{ik} f_{Y|\boldsymbol{X},Z}(y_i \mid \boldsymbol{x_i}, s_j; \hat{\boldsymbol{\theta}}) w_k}, \quad j = 1, \ldots, m.$$

2. Given an estimate of $\boldsymbol{w}$, the log-likelihood function (2) is maximized with respect to the regression parameters $\alpha, \boldsymbol{\beta}', \gamma$, and $\boldsymbol{\tau}'$.

2.2 Gamma regression model

Using the log as link function, the expression of the gamma regression model is

$$\log(\mu) = \log(\mathrm{E}(Y|\boldsymbol{X}, Z)) = \alpha + \boldsymbol{\beta}' \cdot \boldsymbol{X} + \gamma \cdot Z. \tag{3}$$

In this case, the conditional density function of Y given $\boldsymbol{X}$ and Z to be plugged into the log-likelihood function (2) is given by

$$f_{Y|\boldsymbol{X},Z}(y|\boldsymbol{x}, z; \boldsymbol{\theta}) = \frac{\nu^\nu}{(e^{\alpha+\boldsymbol{\beta}'\cdot\boldsymbol{X}+\gamma\cdot Z})^\nu \Gamma(\nu)} \exp\left\{-\frac{\nu}{e^{\alpha+\boldsymbol{\beta}'\cdot\boldsymbol{X}+\gamma\cdot Z}} \cdot y + (\nu-1)\cdot \log(y)\right\},$$

where ν is the shape parameter of the gamma distribution. The estimation of the parameters by means of the GEL algorithm has been implemented by the authors in R (`https://github.com/klongear/ICbook`).
An important aspect of our current research are goodness-of-fit techniques for this model. The idea is to adapt the model's deviance and the Pearson's chi-squared statistic to the presence of an interval-censored covariate. Both statistics depend on the values of the model's covariates and we propose to impute the unknown values of Z by the respective expected means given the observed intervals under the estimated Turnbull distribution $\hat{F}_Z$ (Gómez et al., 2022).

3 Application to the data of a metabolomic study

The gamma regression model (3) is used to model the glucose level of the participants of the PREDIMED-Plus trial, a multicenter, randomized, primary prevention field trial of cardiovascular disease in an older population with metabolic syndrome (Marhuenda-Muñoz et al., 2019).

The explanatory variables of interest is the sum of the α-carotenoids. Since none of the carotenoid compounds, C_j, $j = m$, can be quantified exactly under its quantification limit (LoQ), and is not even detectable under its detection limit (LoD), the observed concentrations are either left-censored $[0, \text{LoD}_j)$, interval-censored $[\text{LoD}_j, \text{LoQ}_j)$, or exactly determined $[C_j, C_j]$; see Figure 1 for an illustration. As a consequence, the sum over the compounds, $Z = \sum_{j=1}^{m}$, is an interval-censored covariate. Notice that different from most applications with interval-censored data, here, Z is not a time-to-event variable.

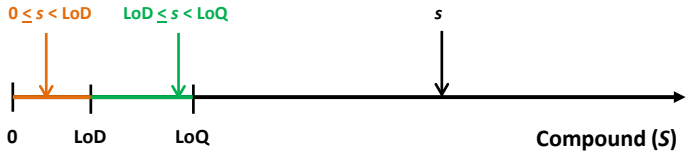


FIGURE 1. Possible observations of compound concentrations: a concentration might be lower than the limit of detection (LoD), lie between the LoD and the limit of quantification (LoQ), or be quantified exactly.

The gamma regression model is adjusted for age and energy intake and the estimated parameter estimate obtained with the GEL algorithm is $-0.0026\frac{mg/dl}{\mu\text{mol}/L}$ with a standard error of $0.0022\frac{mg/dl}{\mu\text{mol}/L}$. Hence, based on the data at hand, it cannot be claimed that glucose level depends on the sum of the α-carotenoids.

Acknowledgments: This work was funded by the *Ministerio de Ciencia e Innovación* (Spain) [PID2019-104830RB-I00/ DOI (AEI): 10.13039/501100011033] and by the *Agència de Gestió d'Ajuts Universitaris i de Recerca* of the *Generalitat de Catalunya* (Spain) [2021 SGR 01421].

References

Gómez, G., Calle, M., Oller, R., and Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modelling*, **9**, 259 – 297.

Gómez, G., Espinal, A., and Lagakos, S.W. (2003). Inference for a linear regression model with an interval-censored covariate. *Statistics in Medicine*, **22**, 409 – 425.

Gómez, G., Marhuenda-Muñoz, M., and Langohr, K. (2022). Regression analysis with interval-censored covariates. Application to liquid chromatography. In: Sun J., Chen DG. (eds). *Emerging topics in modeling interval-censored survival data.* ICSA Book Series in Statistics. Springer, Cham.

Langohr, K. and Gómez, G. (2014). Estimation and residual analysis with R for a linear regression model with an interval-censored covariate. *Biometrical Journal*, **56**, 867 – 885.

Marhuenda-Muñoz, M., Domínguez-López, I., Langohr, K. et al. (2022). Circulating carotenoids are associated with favorable lipid and fatty acid profiles in an older population at high cardiovascular risk. *Frontiers in Nutrition*, **9**, 967967.

Morrison, D., Laeyendecker, O., and Brookmeyer, R.] (2022). Regression with interval-censored covariates: Application to cross-sectional incidence estimation. *Biometrics*, **78**, 908 – 921.

Oller, R., Gómez, G., and Calle, M. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood. *The Canadian Journal of Statistics*, **32**, 315 – 325.

Bayesian regularisation for tail index regression

M.W. Lee[1], M. de Carvalho[1], D. Paulin[1], S. Pereira[2], R. Trigo[2] C. Da Camara[2]

[1] School of Mathematics, University of Edinburgh, United Kingdom
[2] Faculdade de Ciências and CEAUL, Universidade de Lisboa, Portugal

E-mail for correspondence: johnny.myungwon.lee@ed.ac.uk

Abstract: We propose a novel approach for modelling the extreme values via a Bayesian regularisation that learns about a tail index regression framework. Our method is based on a conditional Pareto-type specification which is regularised with a shrinkage prior. To validate the performance of the proposed approach a battery of numerical experiments was conducted, and an illustration is given on extreme wildfires in Portugal.

Keywords: Bayesian Regularisation; ℓ_p-penalty; Tail Index Regression; Heavy-Tailed Response; Conditional Pareto-type Distribution.

1 Introduction

In this paper, we extend the Tail Index Regression (Wang & Tsai, 2009) to a Bayesian regularisation framework that characterises the extreme behaviour of a response variable that follows a conditional Pareto-type tail specification.

From a Bayesian perspective, each regression coefficient follows an independent and identically distributed shrinkage prior that behaves equivalently to the ℓ_p-type penalty regularisation. This aligns with the structure of the heavy-tailed distribution where certain covariates are determined as key factors of the extremeness. As a result, our approach entails a regularisation on a fully semiparametric framework by concentrating on learning about the regression coefficients that achieve a relatively sparse structure. Our contribution has important implications, particularly to the research in modelling extreme wildfires—such as the devastating 2017 Portugal wildfire (Turco et al., 2019)—and on the identification of their underlying drivers.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Bayesian Regularisation for Tail Index Regression

Our starting point for modelling is the following conditional Pareto-type tail specification that stems from Beirlant et al. (2004, Ch. 9):

$$P(Y > y \mid \boldsymbol{X} = \boldsymbol{x}) \equiv 1 - F(y \mid \boldsymbol{x}) = y^{-\alpha(\boldsymbol{x})}\mathcal{L}(y \mid \boldsymbol{x}). \tag{1}$$

Here, $\alpha(\boldsymbol{x}) = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$, is a covariate-adjusted tail index, with the observation $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ and regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$; in addition, $\mathcal{L}(y \mid \boldsymbol{x})$ is a covariate-adjusted slowly varying function, that is, $\mathcal{L}(yt \mid \boldsymbol{x})/\mathcal{L}(y \mid \boldsymbol{x}) \to 1$, as $y \to \infty$, for all $t > 0$. The specification in (1), allows for the heavy tail behaviour—as captured by the tail index—to depend on covariates. We follow Wang & Tsai (2009) and consider the Hall's (1982) class of covariate-adjusted slowly-varying functions given by,

$$\mathcal{L}(y \mid \boldsymbol{x}) = c_0(\boldsymbol{x}) + c_1(\boldsymbol{x})y^{-\theta(\boldsymbol{x})} + O(y^{\theta(\boldsymbol{x})}), \tag{2}$$

where $c_0(\boldsymbol{x}), c_1(\boldsymbol{x})$ and $\theta(\boldsymbol{x}) > 0$. Hence, $\mathcal{L}(y \mid \boldsymbol{x}) \to c_0(\boldsymbol{x})$ and $\partial\mathcal{L}(y \mid \boldsymbol{x})/\partial y \to 0$, as $y \to \infty$.

To regularise the above described tail index regression model, we resort to shrinkage priors. Given the space constraint we will focus on the Laplace prior, but other variants of the approach can be readily constructed by considering other shrinkage penalties, as illustrated in Fig. 1.

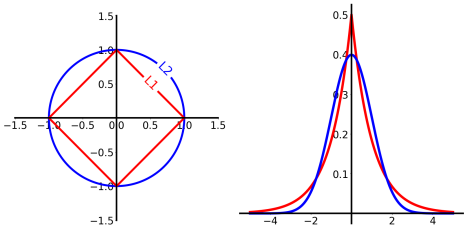


FIGURE 1. Comparison of the geometry of a unit ball induced by Laplace (red) and Normal (blue) priors depicting ℓ_1 and ℓ_2 penalty regularisation, respectively.

Concretely, in terms of the Bayesian Lasso (Park and Casella, 2008) version of our approach for (1), we learn about $\boldsymbol{\beta}$ from a random sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim F(\boldsymbol{x}, y)$. In Bayesian context, the posterior mode from a Laplace prior corresponds to maximising the constraints of the objective function with a ℓ_1 penalty. Thus, the sparsity-inducing regularisation shrinks the coefficients of less influential variables and results in the sparse structure distribution. Then, the resulting posterior density is,

$$p(\boldsymbol{\beta} \mid \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n) \propto \exp\left\{1/\lambda \sum_{j=1}^{p} |\beta_j|\right\} L(\boldsymbol{\beta}) \tag{3}$$

where L is the approximated likelihood that follows from (2),

$$L(\boldsymbol{\beta}) \approx \prod_{i=1}^{n} f(y_i \mid \boldsymbol{x}_i) \approx \prod_{i=1}^{n} \alpha(\boldsymbol{x}_i)(y_i/u)^{-\alpha(\boldsymbol{x}_i)} y_i^{-1}, \tag{4}$$

for some large threshold u, with $y_i > u$, and where $f = \mathrm{d}F/\mathrm{d}y$. The priors for the regression parameters, $\boldsymbol{\beta}$ of tail index, $\alpha(\boldsymbol{x})$ are then defined as

$$\beta_j \mid \lambda \sim \text{Laplace}(\lambda), \quad \lambda \sim \text{Gamma}(a, b), \tag{5}$$

with $a, b > 0$, where an uninformative Gamma prior was chosen as the hyperprior for λ. Since the posterior has no closed-form expression, we resort to Markov Chain Monte Carlo (MCMC) methods for sampling.

3 Simulation Study

To assess the performance of the proposed method, we consider:
Scenario A: Conditional Pareto, i.e., $\mathcal{L}(y|\boldsymbol{x}) \propto 1$, with

$$\boldsymbol{\beta} = c(0.2, 0, 0.8, 0, 0, -0.1, 0, 0, 0, -0.4)^T.$$

Scenario B: Conditional Burr, i.e., $\mathcal{L}(y|\boldsymbol{x}) \propto (y^{-c(\boldsymbol{x})}+1)^{-2}$, where $\alpha(\boldsymbol{x}) = c(\boldsymbol{x})$ with

$$\boldsymbol{\beta} = c(0.1, 0.5, 0, 0, -0.9, -0.5, 0, 0.4, 0, 0)^T.$$

Scenario C: Conditional F, i.e., $\mathcal{L}(y|\boldsymbol{x}) \propto (y^{k_1/2-1}(k_1 + k_2(\boldsymbol{x})y)^{-(k_1+k_2(\boldsymbol{x}))2})$, where $\alpha(\boldsymbol{x}) = k_2(\boldsymbol{x})/2$ with

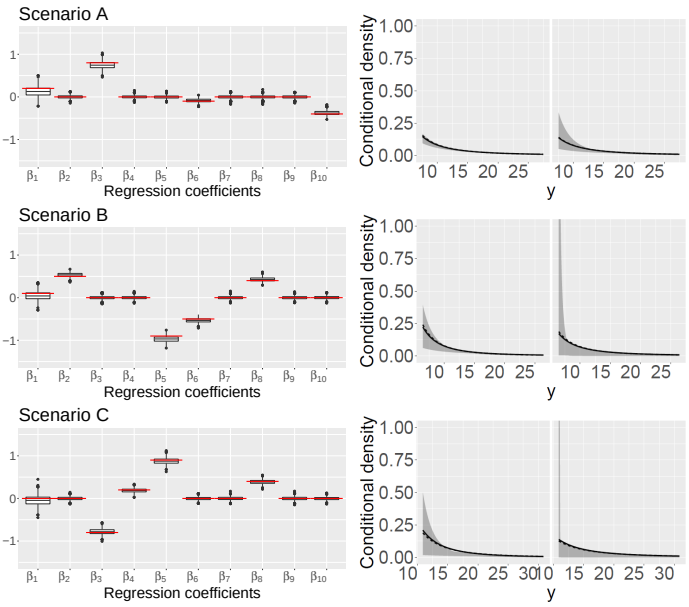
$$\boldsymbol{\beta} = c(0, 0, -0.8, 0.2, 0.9, 0, 0, 0.4, 0, 0)^T.$$

We retrieved the 90% quantile of 5000 random samples from (1) and repeated the study 250 times for Monte Carlo simulation. An uninformative Gamma prior, $\Gamma(0.1, 0.1)$ was employed, and in terms of MCMC we took 10,000 burn-in iterations and collected 20,000 samples. From Fig. 2 (top), it can be observed that each posterior mean approximates the true value well, hence suggesting a good performance of our method.

4 Real Data Application

We illustrate the proposed method on data of *Instituto Dom Luiz* that consists the daily burn area of forest fires between 1980 and 2019 in Portugal. We examine the following potential drivers for the same period of time: southerly flow (SF), westerly flow (WF), total flow (F), southerly shear vorticity (ZS), westerly shear vorticity (ZW), total shear vorticity (Z), and direction of flow (DF). We filtered 731 observations out of 14,609 and standardised their covariates. We used a Normal prior, $N(0, 100^2)$ for the intercepts and the same setup as in Section 3. Fig. 2 (bottom left) suggests WF, DF and SF are the drivers of the extreme forest fires. Fig. 2 (bottom right) depicts the corresponding randomised quantile residuals against the theoretical standard normal quantiles and it evidences a good fit.

Simulation



Real Data

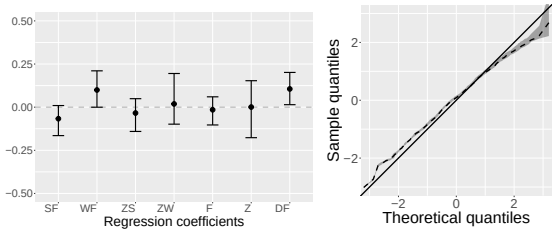


FIGURE 2. Results from simulation study (Top) and data illustration (Bottom).

References

Beirlant, J. , Goegebeur, Y., Segers, J. & Teugels, J. L. (2004). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons.

Hall, P. (1982). On Some Simple Estimates of an Exponent of Regular Variation. *Journal of the Royal Statistical Society: Series B*, **44(1)**, 37 – 42.

Park, T. & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, **103(482)**, 681 – 686.

Turco, M., Jerez, S., Augusto, S., Tarín-Carrasco, P., Ratola, N., Jiménez-Guerrero, P., & Trigo, R. M. (2019). Climate Drivers of the 2017 Devastating Fires in Portugal. *Scientific reports*, **9(1)**, 13886.

Wang, H. & Tsai, C.L. (2009). Tail Index Regression. *Journal of the American Statistical Association*, **104(487)**, 1233 – 1240.

Best subset selection for principal components analysis and partial least square models using continuous optimization

Benoit Liquet[1,2], Sarat Moka[3], Samuel Muller[1]

[1] School of Mathematical and Physical Sciences, Macquarie University, Sydney, Australia
[2] Laboratoire de Mathématiques et de leurs Applications, Université de Pau et des Pays de l'Adour, Pau, France
[3] School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia

E-mail for correspondence: benoit.liquet-weiland@mq.edu.au

Abstract: Choosing the most important variables in supervised and unsupervised learning is a difficult task, especially when dealing with high-dimensional data where the number of variables far exceeds the number of observations. In this study, we focus on two popular multivariate statistical methods - principal component analysis (PCA) and partial least squares (PLS) - both of which are linear dimensionality reduction techniques used in a variety of fields such as genomics, biology, environmental science, and engineering. Both PCA and PLS generate new variables, known as principal components, that are combinations of the original variables. However, interpreting these components can be challenging when working with large numbers of variables. To address this issue, we propose a method that incorporates the best subset selection approach into the PCA and PLS frameworks using a continuous optimization algorithm. Our empirical results demonstrate the effectiveness of our method in identifying the most relevant variables. We illustrate the use of our algorithm on two real datasets - one analyzed using PCA and the other using PLS.

Keywords: Best Subset Selection; Continuous Optimization; Partial Least square; Principal Component; Sparsity

1 Introduction

Identifying the most relevant variables is a difficult task, particularly in high dimensional contexts where there are typically many more variables than

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

observations. Analyzing each variable individually can be time-consuming, and presenting results using graphs or numerical measures may not provide sufficient insight as either too many features are visualized, or the summary information may be inconclusive. To address this, multivariate statistical methods such as principal component analysis (PCA) and partial least squares (PLS) are commonly used. These methods are well-established linear dimensionality reduction techniques that are particularly useful for analyzing data with a large number of variables. By constructing new variables (known as principal components) that are linear combinations of the original variables, PCA and PLS can help identify the most important variables and simplify the analysis of complex datasets.

This paper proposes a new method for identifying components that are based on the most relevant variables. Specifically, we approach the challenge of defining sparse components as a "best subset selection" (BSS) problem, where the objective is to find the best subset of k variables for constructing the components. BSS has been extensively studied in the context of linear regression, with existing methods offering solutions beyond exhaustive search, such as the Furnival Wilson algorithm, which becomes impractical when the number of variables exceeds 30. To address this, we propose an approach for BSS in PCA and PLS models that is based on a continuous optimization algorithm recently developed for BSS in linear regression.

Our approach frames BSS for PCA and PLS as continuous optimization algorithms that can leverage standard continuous optimization techniques such as gradient descent to explore a large set of subsets.

In this short paper we only present the problem of best subset selection (BSS) for the PLS model with univariate response, which is the simpler optimization problem to solve. This particular PLS model is known as PLS1. During our oral presentation, we will present BSS for the multivariate case of PLS, called PLS2. We will further show that the BSS for PCA can be easily derived from the BSS for PLS2. A simulation study will be also presented during our talk where we highlight the ability of our algorithms to recover best subsets in PCA and PLS models. Finally, we will present applications of our algorithm for two different real datasets.

2 Best Subset Selection for PLS with Univariate Response

The first component pair for the PLS model for the data matrices X of dimension $n \times p$ and univariate response $\mathbf{y} \in \Re^n$ is obtained by solving,

$$\max_{\mathbf{u}\in\Re^p,\ \|\mathbf{u}\|=1} \operatorname{cov}(X\mathbf{u}, \mathbf{y}) = \max_{\mathbf{u}\in\Re^p,\ \|\mathbf{u}\|=1} \frac{\langle X\mathbf{u},\ \mathbf{y}\rangle}{n}. \tag{1}$$

where $\operatorname{cov}(\cdot,\cdot)$ is the sampling covariance operator.

We now consider the BSS framework for constructing the first component score. Finding the optimal solution $\mathbf{u}^* \in \Re^p$ of (1) is given by

$$\mathbf{u}^* = \frac{X^\top \mathbf{y}}{\|X^\top \mathbf{y}\|}. \tag{2}$$

Now suppose we want to introduce sparsity, in the sense that the new optimization problem is

$$\max_{\mathbf{u}_{[\mathbf{s}]} \in \Re^k,\ \|\mathbf{u}_{[\mathbf{s}]}\|=1} \frac{\langle X_{[\mathbf{s}]}\mathbf{u}_{[\mathbf{s}]},\, \mathbf{y}\rangle}{n}, \quad \text{subject to} \quad \mathbf{s} \in \{0,1\}^p,\ |\mathbf{s}| \le k, \tag{3}$$

where $X_{[\mathbf{s}]}$ is the matrix constructed from X by removing all its columns with indices j where $s_j = 0$, k is the sparsity parameter that represents the subset size, and $|\mathbf{s}|$ denotes the number of ones in the binary vector $\mathbf{s}$. Observe that for any fixed binary vector $\mathbf{s}$, the optimal solution is $\mathbf{u}^*_{[\mathbf{s}]} = X^\top_{[\mathbf{s}]}\mathbf{y} / \left(\|X^\top_{[\mathbf{s}]}\mathbf{y}\|\right)$.
Thus, the optimization problem (3) can be expressed as

$$\max_{\mathbf{s}\in\{0,1\}^p} \frac{\langle X_{[\mathbf{s}]}\mathbf{u}^*_{[\mathbf{s}]},\, \mathbf{y}\rangle}{n}, \quad \text{subject to} \quad |\mathbf{s}| \le k,$$

Since,

$$\langle X_{[\mathbf{s}]}\mathbf{u}^*_{[\mathbf{s}]},\, \mathbf{y}\rangle = \left(\mathbf{u}^*_{[\mathbf{s}]}\right)^\top X^\top_{[\mathbf{s}]}\mathbf{y} = \frac{\|X^\top_{[\mathbf{s}]}\mathbf{y}\|^2}{\|X^\top_{[\mathbf{s}]}\mathbf{y}\|} = \|X^\top_{[\mathbf{s}]}\mathbf{y}\|,$$

we can express (3) as

$$\min_{\mathbf{s}\in\{0,1\}^p} \left[-\frac{\|X^\top_{[\mathbf{s}]}\mathbf{y}\|}{n}\right], \quad \text{subject to} \quad |\mathbf{s}| \le k. \tag{4}$$

This problem defines the best subset selection for PLS1. However, solving this problem is NP-hard, and hence, we consider, by exploiting the same idea as in Moka et al. (2022), a relaxation of (4) given by

$$\min_{\mathbf{t}\in[0,1]^p} \left[-\frac{\|X^\top_{\mathbf{t}}\mathbf{y}\|}{n}\right], \quad \text{subject to} \quad \sum_{j=1}^{p} t_j \le k, \tag{5}$$

where $\mathbf{t} = (t_1, \ldots, t_p)^\top$, with each $t_j \in [0,1]$, and $X_{\mathbf{t}}$ is obtained from X by multiplying its j-th column with t_j for every $j = 1, \ldots, p$. Since minimizing $-\|X^\top_{\mathbf{t}}\mathbf{y}\|$ is equivalent to minimizing $-\|X^\top_{\mathbf{t}}\mathbf{y}\|^2$, to simplify the gradient expression later, we rewrite (5) as

$$\min_{\mathbf{t}\in[0,1]^p} \left[-\frac{\|X^\top_{\mathbf{t}}\mathbf{y}\|^2}{n^2}\right], \quad \text{subject to} \quad \sum_{j=1}^{p} t_j \le k. \tag{6}$$

Note that the optimization problem in (4) is defined using $X_{[\mathbf{s}]}$ constructed by removing columns from the design matrix X (and hence $X_{[\mathbf{s}]}$ and X are of different sizes) while $X_{\mathbf{t}}$ in optimization problem in (5) is constructed by multiplying the j-th column of X by t_j for every j. Thus, both $X_{\mathbf{t}}$ and X are of the same size. This construction allows us to define our new estimator of the loading vector $\mathbf{u_t}$ for all $\mathbf{t} \in [0,1]^p$ while guaranteeing that

$$\|X_{\mathbf{t}}^\top \mathbf{y}\| = \|X_{[\mathbf{s}]}^\top \mathbf{y}\|, \quad \text{for } \mathbf{t} = \mathbf{s},$$

at the corner points $\mathbf{s}$ of the hypercube $[0,1]^p$. This construction also guarantees that the new objective function $-\frac{\|X_{\mathbf{t}}^\top \mathbf{y}\|^2}{n^2}$ is smooth over the hypercube as illustrated in Figure 1.

Finally, instead of solving (6), we consider $f_\lambda^{\text{PLS1}}(\mathbf{t}) = -\frac{\|X_{\mathbf{t}}^\top \mathbf{y}\|^2}{n^2} + \lambda \sum_{j=1}^p t_j$, and solve

$$\min_{\mathbf{t} \in [0,1]^p} f_\lambda^{\text{PLS1}}(\mathbf{t}), \tag{7}$$

using a continuous optimization method, such as basic gradient descent or Adam (as shown in the example of Figure 1). To execute such a gradient descent algorithm, we use the gradient expression given by

$$\nabla f_\lambda^{\text{PLS1}}(\mathbf{t}) = \lambda I - \frac{2}{n^2} \left(\mathbf{t} \odot X^\top \mathbf{y} \odot X^\top \mathbf{y} \right), \tag{8}$$

where I is the identity matrix and $\odot$ is the element-wise product operator. Exploiting the continuity of the new objective function enables gradient descent algorithms to explore a huge space of models while converging in a few iterations to identify the best subset. By increasing the value of λ, we can increase the sparsity of the solution of the optimization problem (5), because the penalty $\lambda \sum_{j=1}^p t_j$ encourages sparsity in $\mathbf{t}$ (see Figure 1).

References

Moka, S., Liquet, B., Zhu, H., and Muller, S. (2022). COMBSS: Best Subset Selection via Continuous Optimization. doi: 10.48550/ARXIV.2205.02617 .

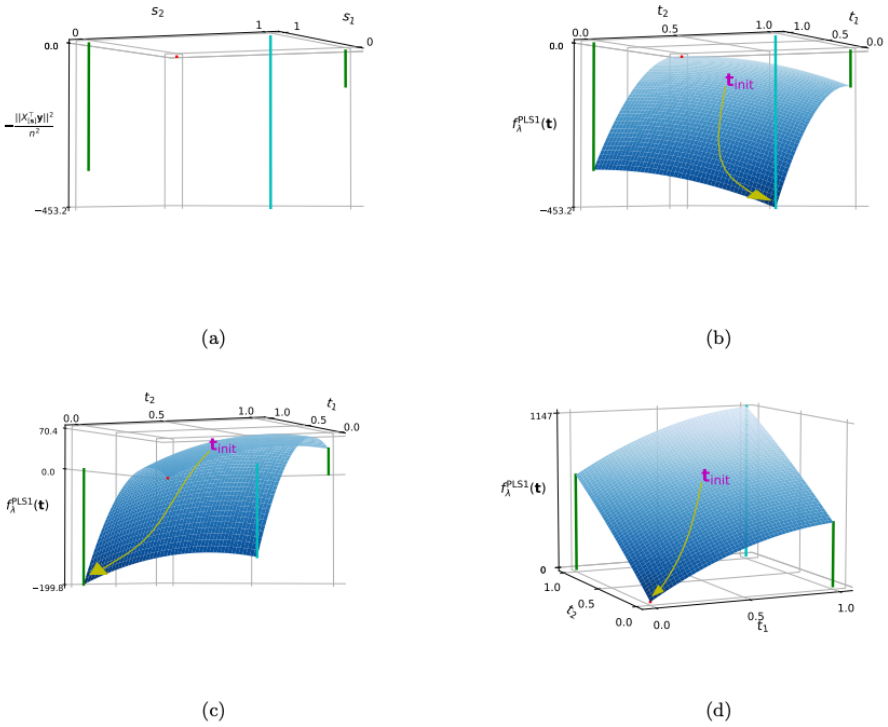


FIGURE 1. Illustration of the workings of our continuous optimization method using basic gradient descent for an example data with $p = 2$. Plot (a) shows the objective function of the PLS model with univariate response at binary points $\mathbf{s} \in \{0,1\}^2$. Observe that the best subsets correspond to $k = 0$, $k = 1$, and $k = 2$ are $(1,1)^\top$, $(0,1)^\top$, and $(0,0)^\top$, respectively. Plots (b) - (d) show the objective function of our optimization method (7) for different values of the parameter λ. In each of these three plots, the curve (in yellow) shows the execution of basic gradient descent algorithm that, starting at the initial point $\mathbf{t}_{\text{init}} = (0.5, 0.5)^\top$, converges towards the best subsets of sizes $0, 1$, and 2.

The consequences of not completing the generational cohort in estimating age-at-menopause

Rui Martins[1,2], Bruno de Sousa[3], Thomas Kneib[4], Maike Hohberg[5], Nadja Klein[6], Elisa Duarte[3], Vítor Rodrigues[7]

[1] Faculdade de Ciências da Universidade de Lisboa (FCUL), Portugal
[2] Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)
[3] Center for Research in Neuropsychology and Cognitive and Behavioral Intervention (CINEICC), University of Coimbra, Portugal
[4] University of Göettingen, Chair of Statistics, Humboldtallee 3, 37073 Goettingen, Germany
[5] Department of Medical Statistics, University Medical Center Göettingen, 37073 Göttingen, Germany
[6] Chair of Uncertainty Quantification and Statistical Learning, Research Center for Trustworthy Data Science and Security; Dep. Stat. Technische Universität Dortmund, Germany
[7] Faculty of Medicine, University of Coimbra, Portugal

E-mail for correspondence: rmmartins@fc.ul.pt

Abstract: When studying age-at-menopause of a particular generation cohort of women the approach where women without an observed menopause are deleted from the study is not advisable because they might convey different informations for the analysis namely about the so called period effect. Generally, the deleted are the youngest who have not yet reached menopause.
The context is a Portuguese breast cancer screening programme in the period 1990–2010 where a late menopause is considered a risk factor. Our aim is to recover missing menopause ages by comparing methods for handling missing (or incomplete) data.
Two imputation approaches are considered: (i) multiple imputation based on a truncated distribution but ignoring the mechanism of missingness; (ii) a bivariate copula-based imputation that simultaneously handles the age-at-menopause and the missing mechanism.
There are contradictory results in current research about whether age-at-menopause is increasing or decreasing in Western countries. We show that both imputation methods unveiled an increasing trend of age at menopause when

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

viewed as a function of the birth year for the youngest generation. This trend is hidden if we model only women with an observed age-at-menopause.

Keywords: menopause; incomplete data; copula regression; GJRM; gamlss.

1 Introduction

Age-at-menopause has an important role in the research about risk factors for breast cancer. However, it is a variable prone to incompleteness, because the time when women participate in a breast cancer screening program overlaps the time when women are most likely to enter menopause. Therefore, the younger women of the generation cohort under analysis tend to have missing information on age-at-menopause. Not recovering those values can lead to wrong conclusions because the parameters estimates for the most recent years will tend to be dominated by these young women.
The question of whether missing values of a variable are related to the underlying value itself allows for classifying the missing data mechanism into three categories (Rubin 1976): missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).
We frame the issue of imputing age-at-menopause as a missing data problem since we consider this measure as a covariate in a potential subsequent risk cancer analysis. We therefore ask the same question as in a classical missing value setting: Is the missing mechanism informative or not? Note that recovering the values for age-at-menopause as the dependent variable could also be treated as a censoring or prediction problem but is not the focus of this work.
To test how different strategies to impute missing ages-at-menopause for the youngest women influence the analysis of time- and spatial-trends of that variable, we will analyse the case of a breast cancer screening program in central Portugal. Exploratory analyses show the presence of a geographical pattern of the missing data and a close relation with a woman's year of birth (a.k.a. period effect), implying, at least, a violation of the MAR assumption. Additionally, there is a high percentage of missing values in the variable of interest (23.6%), which precludes an analysis by simply deleting those individuals.
To achieve the goals defined above, we will consider two statistical modelling approaches with the aid of two `R` packages, namely `GJRM` – Generalised Joint Regression Modelling (Marra and Radice, 2017) and `gamlss` – Generalised Additive Models for Location, Scale and Shape (Rigby and Stasinopoulos, 2005). The `GJRM` package allows us to deal simultaneously with two response variables while their specific marginal distributions are conveniently expressed in a joint manner by means of a copula function that binds them together. In this way, we will be able to define a joint distribution for both the process that governs the probability that a woman has not yet reached menopause and for the age-at-menopause itself. The

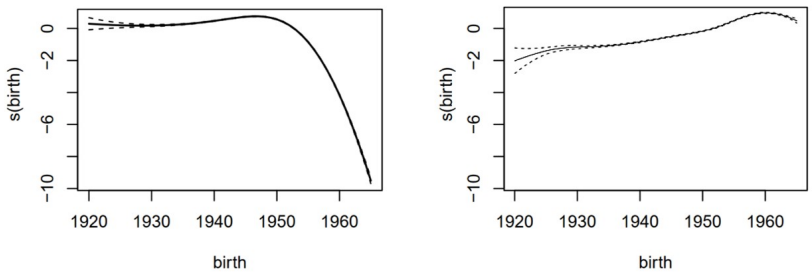


FIGURE 1. Birth year (flexible) effect if only considering women with an observed menopause (left panel). Birth year (flexible) effect after the missing menopause ages have been replaced with the imputations via a truncated Weibull distribution at the screening age (right panel).

`gamlss` package adopts a method for the imputations which is very flexible and allows imputations from truncated distributions.

2 Data

The dataset that we are working with has of 278 282 women between 1990 and 2010. At the age of 45, all women in each of the 78 municipalities in central Portugal are invited to have a free screening mammogram and every two years thereafter until the age of 69. This region roughly represents 25% of the Portuguese population. At the time of the last screening, 65 765 women (23.6%) stated they had not yet reached menopause (missing information). The variables included in the dataset are: (i) binary characteristics provided by the variables `pregnancy`, `breastfeeding` and the use of `oral contraceptives`; (ii) quantitative information carried by the continuous variables are `age at menopause`, `age at menarche`, `year of birth` and `age at the last screening`; (iii) demographic information given by the `municipality purchasing power index`; and (iv) spatial information embodied in neighbourhood structure of the `municipality of residence`.

2.1 Methodology

The primary goal of this work is to draw inferences about the distribution of the age-at-menopause, Y_i, $i = 1, 2, \ldots, n$ given a set of observed covariates, $\mathbf{v}_i$, by considering the primary analysis model $[Y_i \mid \mathbf{v}_i]$. The most popular approach would be to estimate its parameters using only the observed Y_i's, yet estimates from such an analysis would be less efficient than they would be if we had observed Y for every individual. Recovering information via an imputation technique, e.g. multiple imputation (MI), should allow to retrieve some of the information about Y that is not available.

We discuss two different approaches for dealing with missing menopause ages. One considers the data as MNAR and therefore we jointly model the missing data mechanism and the response variable of interest via a bivariate copula. The other considers an MAR data structure and thus only the statistical process of the age-at-menopause is modelled.
The imputations will be obtained by sampling from an approximation to the posterior predictive distribution of the missing data given modelling assumptions and the observed data,

$$f(\mathcal{Y}_{\text{mis}} \mid \mathcal{Y}_{\text{obs}}, \mathbf{v}_i) \approx \int f(\mathcal{Y}_{\text{mis}} \mid \mathbf{\Phi}, \mathbf{v}_i)\, \tilde{f}(\mathbf{\Phi} \mid \mathcal{Y}_{\text{obs}}, \mathbf{v}_i)\, \mathrm{d}\mathbf{\Phi}, \tag{1}$$

where $\mathcal{Y}_{\text{obs}}$ represents the observed menopause ages and $\mathcal{Y}_{\text{mis}}$ the unobserved ones; $\tilde{f}(\mathbf{\Phi} \mid \mathcal{Y}_{\text{obs}}, \mathbf{v}_i)$ is the approximated posterior distribution of all the parameters combined in the vector $\mathbf{\Phi}$.

3 Results

An imputation procedure for the missing ages-at-menopause is required if the study aims at analysing the trend of a variable in a setting that includes a cohort of women where the majority has already reached menopause and only a small part has not yet. This is always the case when we have a cohort whose age range includes the more likely age to reach the menopause. In settings, where either all women have already reached menopause, or neither woman is in menopause yet, there is no need to resort to any imputation procedure. From a statistical point of view, the first situation only requires the specification of an analysis model. The second situation cannot be inferred because we do not have information to predict individual menopause, unless we assume that they have the same characteristics as the older cohorts but then we would not be able to study the temporal trends across cohorts.
With a dataset similar to the one that we worked with, not imputing the missing ages-at-menopause means that we will have to wait for all women belonging to the youngest cohorts to reach menopause in order to be able to assess the temporal trends of the menopause for that specific cohort of women. When fulfilling a dataset with imputations made in a proper way, we can model the temporal trends of the age-at-menopause immediately. This means that, in terms of public health, we will be studying the phenomenon of menopause without delays. The naive approach of simply delete the women without an observed menopause leads to biased results (Figure 1 - Left panel).
Finally, we would like to emphasize that age-at-menopause is increasing in the central region of Portugal as a function of the birth year (Figure 1 - Right panel).

Acknowledgments: This work was partially funded by Fundação para a Ciência e a Tecnologia (FCT) through the projects INIC-DAAD – DAAD 441.00, UIDB/ 00006/ 2020 and POCI/ 01/ 0145/ FEDER/ 029443 – SHSADReM – Addressing Social and Health Challenges through new developments in Structured Additive Distributional Regression Models. Nadja Klein acknowledges support through the Emmy Noether grant KL 3037/1-1 of the German research foundation (DFG).

References

Marra G, Radice R. (2017) Bivariate copula additive models for location, scale and shape. *Comput Stat Data An*, **112**, 99 – 113.

Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape (with discussion) *J R Stat Soc Ser C Appl Stat*, **54**(3), 507 – 554.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**(3):581 – 92

Information retrieval models with GPT-3: Techniques for improving ranking performance through text enhancement

Kenan M. Matawie[1], Sargon Hasso[2]

[1] Western Sydney University, Australia
[2] Loyola University Chicago, UL, Northbrook, IL USA

E-mail for correspondence: k.matawie@westernsydney.edu.au

Abstract: This paper discusses techniques for improving the ranking performance of information retrieval models through text enhancement using GPT-3's Large Language Model (LLM). Our goal is to demonstrate how the relevance of retrieved documents can be improved by ingesting and indexing better quality corpus data in the Solr search engine. We describe the methodology used in our research and present an analysis and evaluation of our test results. Our conclusion is that using GPT-3 to generate higher quality documents can enhance the relevance of retrieved documents in information retrieval models. This provides another alternative for evaluating retrieval models using test collections made available to the retrieval research community at large.

Keywords: GPT-3; Relevance; Non-Parametric Analysis; Information Retrieval Models; Large Language Model

1 Introduction

In our previous work, Hasso, S., Matawi, K. (2022), we discussed how semantically enriched query alternatives improve the score and rank of search results in Information Retrieval Models. The improvement was statistically analysed using TREC data and Solr full-text search platform, Apache Software Foundation (2021). The scoring functions such as BM25 is generally used to measure the improvement. Mean Average Precision (mAP) measure was used to compare different configurations of the search engine. We used a scheme to transform a query into an alternative form, in addition to its original form, and combined it with other factors, such as term boosting and word embedding models, to produce different configurations as input

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

to the Solr search engine. The purpose of that research paper was to answer the question of what factors influence the relevance of retrieved documents and how to validate the differences statistically. In this paper we aim to show how the relevance of a retrieved document can be improved if a better quality corpus data is ingested and indexed by Solr search engine. We will demonstrate how we used GPT-3, OpenAI (2021), to generate better quality documents.
In section 2 we survey some of the related research and give some background information. In section 3, we briefly explain the scoring Model, ranking, and the evaluation model. In section 4, we describe a methodology we used in this research and describe our experimental design. An analysis and evaluation of the test results are discussed in section 5. We provide summary in section 6.

2 Related Work

The TREC test collections and evaluation software are available to the retrieval research community at large, so organisations can evaluate their own retrieval systems at any time, Harman, D. (2005). In TREC evaluation, systems are given a document set and a set of information needs called topics. The model produces a ranked list of documents per topic where each list is ordered by decreasing likelihood that the document matches the information need. Not all documents are judged for each topic, collection builders sample the collection so that a small fraction of the entire document set is judged for a topic but (most of) the relevant documents are nonetheless identified. The relevance judgment for each topic is provided in a file called "qrels", query relevance, that we use to evaluate the performance of any retrieval model.

3 Evaluation Models

Let d be a document and q be a query. For each query term i, let f_{id} be the frequency of term i in document d. Let n_i be the number of documents containing term i, and let N be the total number of documents in the corpus.

$$S_{dq} = \frac{\sum_{i=1}^{n} w_i(k+1)f_{id}}{f_{id} + k(1-b+b|d|/avgdl)} log(\frac{N-n_i+0.5}{(n_i+0.5})$$

Where, S is the Score, k and b are tuning parameters, typically set to k=1.2 and b=0.75 based on empirical studies, $avgdl$ is the average document length in the corpus, $|d|$ is the length of document d, w_i is a weighting factor for term i, which is typically calculated using the inverse document frequency (IDF) scheme:

$$w_i = log((N-n_i+0.5)/(n_i+0.5))$$

The BM25 formula, Jones, et al. (2000), calculates a score for each document d based on the relevance of its content to the query q, taking into account factors such as term frequency, document length, and term specificity.
Mean Average Precision mAP is a metric used to evaluate the effectiveness of a ranking model, such as a search engine, in returning relevant results for a given query. The mAP is calculated by averaging the precision of the top k documents returned for a set of queries, where k is a fixed number. The formula for calculating the mAP is:

$$mAP = \frac{1}{Q}\sum_{i=1}^{Q}(\frac{1}{K}\sum_{j=1}^{K}P_j)$$

where: Q is the total number of queries, K is the number of documents to be considered for each query, P_j is the precision at position j in the ranked list of documents for the ith query, defined as:

$$P_j = (number\ of\ relevant\ documents\ in\ top\ j)/j$$

The mAP metric provides an average measure of precision across all queries, taking into account the rank of relevant documents in the result list. A higher mAP value indicates a better ranking model, as it means that more relevant documents are being returned at the top of the result list for a given query.

4 Text Enhancement Technique and Experimental Design

Just as we demonstrated in the Query expansion techniques, we also set out to experiment and explore how a better quality text document improves relevancy of the result sets returned by the model. ChatGPT is a large language model developed by OpenAI, based on the GPT-3.5 architecture. It is designed to understand and generate human-like text in response to user input, OpenAI (2021). Its generative human-like text capability was a driving force for us to test the responses obtained when we submitted the queries we used to test the performance of the retrieval model implemented by Solr search application. So, the process follows these steps:

1. select all the queries from the TREC AVI collection
2. submit each query to chatGPT and obtain a response
3. process all the collected responses as individual documents and index them via Solr search application

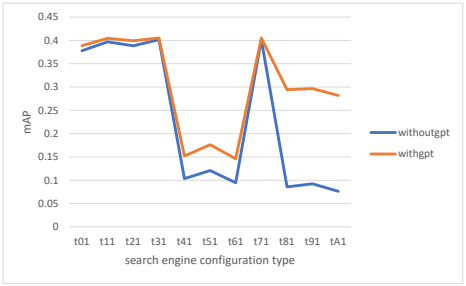


FIGURE 1. Performance of Retrieval Model with and without GPT-3 generated documents

4. run the test harness we developed from our previous work, Hasso, S., Matawie, K. (2022), to generate trec-metrics. During retrieval, Solr uses specific IR model for ranking.

5. analyse the results

Our assumption is that if chatGPT-3 (this is the user interface to GPT-3) language model is giving us a reasonably good response based on its vast knowledge base, those responses would be good and relevant documents when thrown in the mix of the remaining TREC AVI test collection and subjected to the same queries in our Solr-based search application. To test this assumption, we modified the above mentioned "qrels" file to include documents retrieved from chatGPT-3 as relevant documents. Our hypothesis was that if our assumption is correct, the retrieved results from the search engine should also include documents generated by chatGPT-3 and, at least in theory, should outrank or, at least, be in the top list of all other relevant documents. This experiment would also demonstrate whether the documents generated by chatGPT-3 are resilient to the number of query and search engine configurations, 4 and 11, respectively. We used all of the 225 queries provided in the TREC collection. Every response obtained from chatGPT-3 was given an identifier associated with the query that generated the response. This way, we are able to track those documents that came from the TREC collection and those that were generated by chatGPT-3.

5 Analysis and Evaluation

At this stage of the research, we were only interested in testing the quality of the responses returned by chatGPT-3 to a sample of the queries

from TREC collection relevant judgment. We used our Solr engine application experiment from our last research to generate the data and evaluate the relevancy of documents retrieved compared to the standard pre-judged documents. With the assumption that we made in section 4, indeed all the documents generated by chatGPT-3 were ranked in the top 10 returned results by the IR model. This proves, to some extent, that chatGPT-3 can generate documents with a quality that equals or rivals the human-provided relevant judgement. We also made the assumption that each generated response was relevant to the query that chatGPT-3 responded to. In reality, the same document may be relevant to several queries and in fact the AVI TREC collection exhibits this behavior. Figure 1 shows the performance of the IR model using mAP measure as computed by TREC evaluation software. This topic presents a significant opportunity for extensive investigation and analysis, with the potential to supplant human judgment with that of AI.

We conducted a frequency analysis on GPT documents within the corpus we had for a single out of tested 11 configurations, considering their ranks in all the queries. Our analysis revealed that GPT documents consistently achieved a top 10 rank in all queries. The GPT document's frequency of being in the top 10 ranks was further examined using non-parametric Sign and Friedman tests across randomly selected queries samples, and it was found that they achieved this high rank in most of the sampled queries ($p < 0.05$). These findings provide strong evidence suggesting that GPT documents are statistically significant, consistently outperforming other documents and across multiple queries.

6 Conclusion

We have developed an experimental methodology using the Solr search engine to generate and evaluate test results using TREC's evaluation metrics. The Solr search engine was configured and queries were expanded using different techniques. We used the same experimental platform to evaluate the quality of documents generated by GPT-3 large language model-based platform. Responses to our AVI's topics, queries, obtained from GPT-3 were added to the TREC AVI collection. We used the TREC test collections and evaluation software to evaluate the retrieval model implemented by Solr search engine. We observed that for all query list, the documents generated by GPT-3 were among the top 10 ranked documents returned by our retrieval model. The results so far showed that changing either the search engine configurations or the query rewrites did not have observable impact on the ranking of those documents. In addition to the analysis provided in section 5, advanced statistical modeling including document ranks and scores can be further researched.

References

Apache Software Foundation (2021) *The Apache Lucene.* http://lucene.apache. Accessed Feb 2022.

OpenAI (2021). GPT-3.5 [Computer software]. Accessed Mar 30, 2023 from https://openai.com/blog/gpt-3-5/

Harman, D. (2005) The TREC Test Collections, TREC: Experiment and Evaluation in Information Retrieval, 2005, (Accessed March 30, 2023

Hasso, S., Matawi, K. (2022). Experimental Results and Comparisons of Semantically Enriched Query Alternatives in Information Retrieval Models. In *International Workshop on Statistical Modelling(IWSM2022)-Italy*, Pages 462-466.

Jones, S. K., Walker, S. and Robertson, S. E. (2000) A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. In *Information Processing and Management*, pages 779-840.

Analysis of climatological drivers of low-flow events in hydrological Bavaria using large ensemble climate projections

Theresa Meier[1], Nikita Paschan[1], Andrea Böhnisch[2], Henri Funk[1,2], Alexander Sasse[2], Helmut Küchenhoff[1]

[1] Department of Statistics, LMU Munich, Germany
[2] Department of Geography, LMU Munich, Germany

E-mail for correspondence: Theresa.Meier@campus.lmu.de

Abstract: As the world faces the dire reality of climate change, hydrological droughts have become a major concern, with devastating consequences for nature and humans. In Bavarian rivers, low-flow events have occurred more frequently. Therefore, this research project aims to quantify the primary drivers for these events. In climatology, large ensemble climate projections of meteorological and hydrological variables are commonly used to understand the effects of climate change and to make possible predictions. Using ten different realizations, a logistic regression is applied to analyse the data, evaluate the effect sizes and predict low-flow scenarios under changing climate conditions. Furthermore, a K-means clustering algorithm is applied to detect spatial patterns. The analysis reveals large regional differences between the effects and significance of drivers such as precipitation, soil water, snow storage and temperature on the emergence of low-flow in "hydrological Bavaria". For more extreme climate conditions, a partially severe increase of low-flow events can be detected.

Keywords: Climate Modelling; Applied Statistics; Scenario Analysis.

1 Introduction

Changing climate is not only inducing extreme weather patterns but also affects hydrology. In recent years, Bavaria faced more frequent and intense low-flow events. Those events change nature and animal habitats, cause damage to infrastructure and economy and impact the water supply (Marx et. al. 2018). This project contributes to a better understanding of the

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

climatological and hydrological drivers of low-flows in different catchments of "hydrological Bavaria" and their assessment of future climate change.

2 Data

The data are composed of ten hydrological simulations of the WaSiM model (Willkofer et al. (2020)), each of which is driven by a member of the single model initial condition large ensemble CRCM5-LE (Leduc et al. (2019)). The differences in the corresponding realizations between members are induced by perturbations of the initial conditions of the members, yet they are homogeneous in their distributional characteristics. This approach allows natural variability to be taken into account. The resulting time series data covers three-hourly data from 1990 until 2020 which is aggregated into daily data for this analysis.

Hydrological Bavaria (see Figure 1) is divided into 98 catchments with virtual gauges that act as measuring stations. The data set provides regional catchment averages of hydrological and meteorological variables such as precipitation, temperature, snow storage and soil water. A day is classified as a low-flow event when drainage falls below the season-, catchment- and member-specific NM7Q for at least 3 days in a row. The seasons refer to the hydrological half-year, with summer covering the months from May to October and winter the months from November to March.

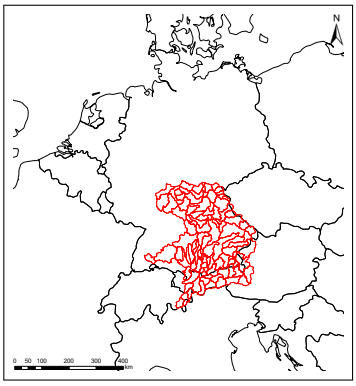


FIGURE 1. Map of hydrological Bavaria.

3 Methods

The occurrence of low-flow events is explained by logistic regressions with drivers such as precipitation, temperature, soil water and snow storage as covariates, as well as interaction terms between temperature and precipitation, soil water and snow storage respectively. Due to the data situation at hand, some challenges arise. To account for a time lag in the drainage-driver relationship, each covariate is included as a simple moving average, the length of which is determined by the inertia of the driver on drainage. Thus, the rolling window for soil water is set to 60 days, for snow storage to 30 days and for the rest of the covariates to 7 days. Since the effects of drivers differ between hydrological half-years, separate models are fitted for

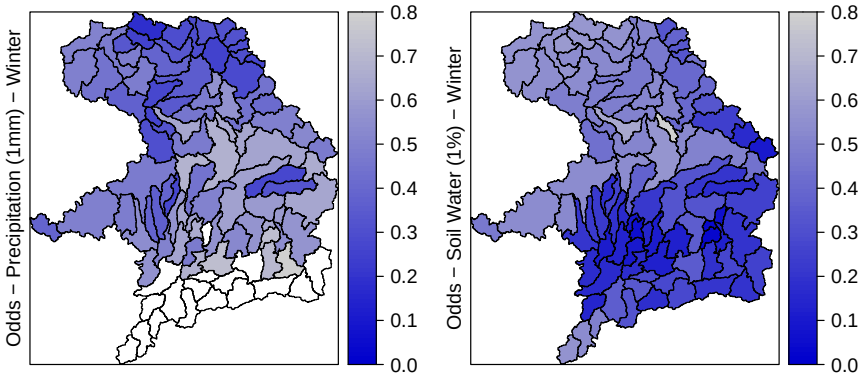


FIGURE 2. Effects for a 1 mm increase of precipitation in winter (left) and for a 1 % increase of soil water in winter (right) on the odds of low-flow events per catchment. Non-significant effects are displayed in white.

each season. The members are taken into account by means of individual logit models and in order to make the effects comparable, the coefficients are averaged over the 10 members. The Bonferroni correction is therefore applied to the assessment of the significance of the effects. Since the catchment "Altmühl-Aha" is not subject to any low-flow events in 9 out of 10 members in summer, this catchment is excluded from the modelling process. Altogether, fitting one model for each combination of catchment, member and season leads to a total of 1959 logistic models. To group catchments according to drivers, a K-means clustering algorithm with Euclidean distance is applied to the coefficients averaged over members of each season. Using the elbow method the optimal number of clusters is set to 4.

4 Results

To assess the goodness of fit, the AUC is determined for each model using test sets consisting of members not used for training. This mostly yields values greater than 0.9, indicating a very good fit. The member-averaged effect sizes are interpreted in terms of odds. Despite natural variability, a comparison of member-specific coefficients leads to very similar results.

4.1 Effects

Analysing the effects of drivers on low-flow events reveals striking regional differences in magnitude and significance. For instance, Figure 2 indicates that in each catchment, an increase in precipitation or soil water decreases the odds of low-flow in winter. Due to a large fraction of snow fall on precipitation in the Alpine regions, no significant effects at a corrected significance level are observable. The north of hydrological Bavaria is strongly

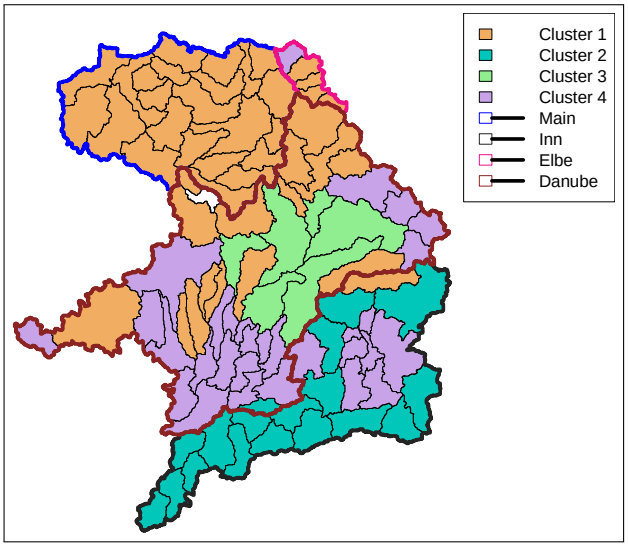


FIGURE 3. Clusters for coefficients in summer and river basins.

influenced by precipitation, while the south is more affected by soil water. These effects are significant, whereas temperature and snow storage show hardly any significant effects. Different combinations of effect sizes and directions between north and south indicate regional differences in the driving dynamics of low-flows.

4.2 Clustering

The clustering groups catchments according to the direction and magnitude of the effects (see 4.1). Subsequently, a cluster is a group of catchments with similar low-flow driving dynamics. For demonstration purposes, only the summer coefficient results are shown in Figure 3. The cluster sizes for the summer coefficients vary, ranging from 6 to 43 catchments within a cluster. The clustering reveals regional similarities in the low-flow driving process. Catchments that are regionally close to each other tend to have similar effects, while more distant catchments are unlikely to exhibit similar patterns. The catchments belonging to clusters 1 (orange) and 4 (purple) are scattered throughout hydrological Bavaria, while the other two clusters (blue and green) are more regionally contiguous. The northern part of hydrological Bavaria, comprising the Main and Elbe river basins, is dominated by the largest cluster 1, which is characterised by a low temperature effect and a lack of snow storage. The latter is due to the fact that its catchments have no snow storage in summer. In contrast, the smallest cluster (cluster 3), located in the centre and entailing some catchments of the Danube, shows the highest mean effect of snow storage and temperature.

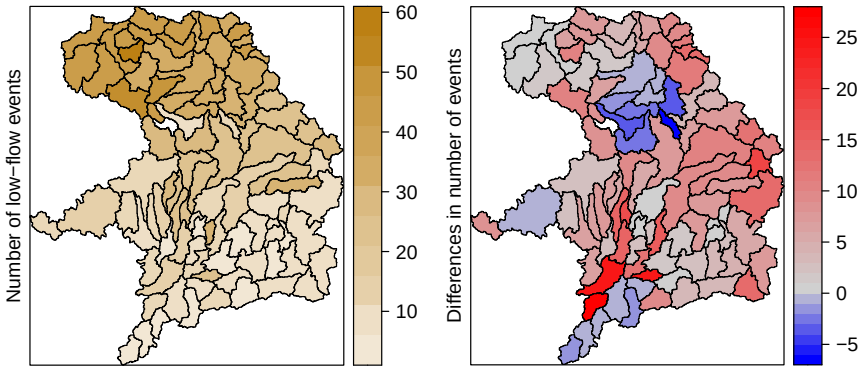


FIGURE 4. Number of predicted low-flows in summer 2010 for unmodified data (left) and differences in the number of days for the climate scenario (right).

Meanwhile, the Alpine region, where the catchments of the river Inn are located, is grouped in cluster 2 and shows the highest mean effect size for precipitation and the second highest effect size for bottom water. It stands alone in having a negative mean temperature coefficient, indicating that an increase in temperature leads to a decrease in the probability of low-flow events. Cluster 4, the second largest cluster with 31 catchments, is scattered across the heart of hydrological Bavaria and shows the largest mean effect of soil water.

4.3 Climate Scenario

In order to analyse the impacts of potential changes of the current climate, e.g. a 3 ˚C rise in temperature, a 50 % reduction in precipitation and the absence of snow storage, the fitted models are now utilized to predict the number of low-flow events for the original and climate scenario data in summer as an example. Due to substantial considerations including a ROC analysis, the threshold for predictions in Figure 4 is set to 0.4. While fewer low-flow events are observed in central and south-western catchments compared to predictions for unmodified data, the expected number of events increases in large parts of hydrological Bavaria, in some catchments even drastically to up to 27 days more. Overall, these hypothetical climate changes result in an additional 519 low-flow days in the summer season across all catchments.

5 Conclusion and Outlook

By using logistic regression, including rolling averages and interactions in the drivers, a framework can be created that facilitates an understanding of the process of low-flow emergence in hydrological Bavaria and allows a comparison between catchments. The analysis exhibits differences in effect size and direction of the drivers by region and season. Clustering of the coefficients derived for summer reveals that the north can be characterized by low temperature and snow storage effects, while the south is dominated by stronger effects of precipitation and soil water. A comparison of predictions for original and more extreme climate data shows a partly drastic increase in low-flow. This analysis compares the 10 members separately, however, a mixed model including members as random effects could be applied to the whole data set. Ongoing research is extending the presented approach by introducing non-linear effects and more flexible time lag structures for more detailed modelling of drivers in individual catchments.

References

Leduc, M. et al. (2019). The ClimEx project: A 50-member ensemble of climate change projections at 12-km resolution over Europe and northeastern North America with the Canadian Regional Climate Model (CRCM5). *Journal of Applied Meteorology and Climatology*, 58(4), 663-693.

Marx, A. et al. (2018). Climate change alters low flows in Europe under global warming of 1.5, 2, and 3 C. *Hydrology and Earth System Sciences*, 22, 1017–1032.

Willkofer, F. et al. (2020). *A Holistic Modelling Approach for the Estimation of Return Levels of Peak Flows in Bavaria.* Water 12(9).

Modeling women's football scores with bivariate distributions from the Sarmanov family

Rouven Michels[1], Marius Ötting[1], Dimitris Karlis[2]

[1] Bielefeld University, Germany
[2] Athens University of Economics and Business, Greece

E-mail for correspondence: r.michels@uni-bielefeld.de

Abstract: For modelling the number of goals in football, the model by Dixon and Coles (1997) has found tremendous impact. By extending the classical double Poisson model such that the probabilities for 0-0, 1-0, 0-1 and 1-1 can be changed, this model is widely considered as the standard model for football scores. We show that this model is also a special case of a multiplicative model known as the Sarmanov family. Within this family we explore further bivariate distributions and fit these extended models to women's football data, as previous models have been applied to men's football only. However, the scores in women's football are different to those of men's football. We find that an extended Sarmanov model emerges as the most promising model for women's football scores.

Keywords: Bivariate distribution, Correlation, Sarmanov family, Football scores

1 Introduction

Modelling football score is of interest for many people such as sports bettors, analysts, and teams' coaches to make informed decisions about the game and for a better understanding of the sport. In the academic literature, Maher et. al (1982) were the first to investigate the joint appearance of goals in men's football. Since then, a variety of model extensions have been proposed, e.g. bivariate Poisson models by Dixon and Coles (1997), Karlis and Ntzoufras (2003) and Groll et. al (2018).
Among these, the model by Dixon and Coles (1997) has been considered as one of the most widely used. In their work, the authors model correlation between the number of goals by shifting probabilities between the scores 0-0, 0-1, 1-0 and 1-1 as they observed more 0-0s and 1-1s in real data of

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

men's football scores compared to what would have been expected under independence. The empirical properties of the other scores matched the assumption of independence. However, such assumptions are not realistic when modelling women's football scores.
To find an appropriate statistical model for women's football scores, we extend the Dixon and Coles model. In particular, we show that it is a special case of the Sarmanov family (Sarmanov, 1966). Within this family, we demonstrate how to shift probabilities for scores with more than one goal and we allow for marginals other than the Poisson distribution. Finally, we apply our proposed models to women's football data from the top leagues in Europe for the seasons 2011/12–2018/19 and 2021/22.

2 Extending the Dixon & Coles model

In this section, we first show that the model by Dixon and Coles (DC) is a special case of the Sarmanov family. Afterwards, we extend this model based on the properties of the Sarmanov family.

2.1 Dixon and Coles model as a member of Sarmanov Family

The Sarmanov family, introduced by Sarmanov (1966), assembles bivariate probability distributions which can be constructed by different probability mass functions $P_i(x_i)$, and bounded non-constant functions $q_i(x_i)$, $i = 1, 2$. If these functions fulfill the condition $\sum_{x_i=-\infty}^{\infty} q_i(x_i)P_i(x_i) = 0$, we can define a joint pmf

$$P(X_1 = x_1, X_2 = x_2) = P_1(x_1)P_2(x_2)[1 + \omega q_1(x_1)q_2(x_2)], \tag{1}$$

with $\omega q_1(x_1)q_2(x_2)$ specifying the dependence of X_1 and X_2. For $\omega = 0$, the variables X_1 and X_2 are independent, i.e., the model collapses to a simple double Poisson model. The correlation between X_1 and X_2 is given by

$$\rho = \frac{\omega u_1 u_2}{\sigma_1 \sigma_2}$$

where $u_i = E[X_i q_i(X_i)]$, for $i = 1, 2$ (Bermúdez and Karlis, 2021).
By setting $\omega = -\tilde{\omega}$ and selecting the functions $q_1(x_1)$ and $q_2(x_2)$ as

$$q_{dc}(x_i) = \begin{cases} -\lambda_i & \text{if } x_i = 0 \\ 1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i > 1, \end{cases}$$

where λ_i is the mean of the Poisson-distributed random variable X_i, we end up with the well-known bivariate model by Dixon and Coles (1997).

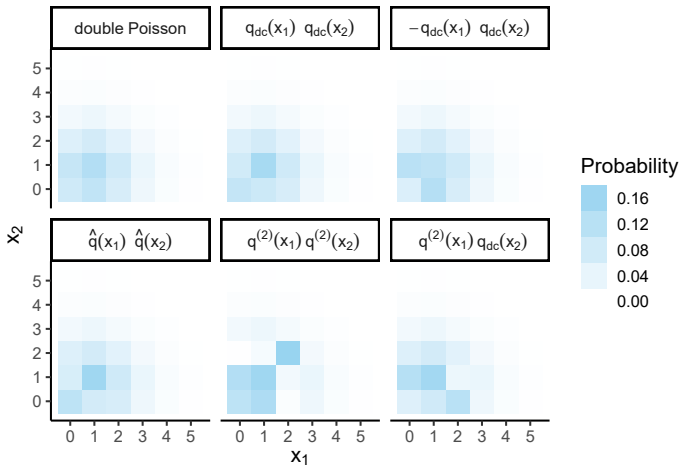


FIGURE 1. This figure displays different bivariate distributions with Poisson marginals based on the Sarmanov family. For the Poisson means we consider 1.3 and 1.2 for x_1 and x_2, respectively.

2.2 Extended models

Within the Sarmanov family, we can extend the DC model by choosing $q-$functions other than q_{dc}. In particular, by using

$$\hat{q}(x_i) = \begin{cases} -\lambda_i^2 & \text{if } x_i = 0 \\ \lambda_i & \text{if } x_i = 1 \\ 0 & \text{if } x_i > 1 \end{cases} \quad \text{or} \quad q^{(s)}(x_i) = \begin{cases} -x_i!\lambda^{s-x_i} & \text{if } x_i < s \\ ss! & \text{if } x_i = s \\ 0 & \text{if } x_i > s, \end{cases}$$

we can change the intensity of the weights ($\hat{q}$), e.g., to inflate the probability of draws even more than in the classical DC model, or to shift probabilities to scores greater than one ($q^{(s)}$) for an arbitrary $s \in \mathbb{N}$. In Figure 1, we compare the probabilities under the different proposed models. In particular, we the selected $q-$functions can also differ across the two teams, as demonstrated in the bottom right panel that. Sticking together different $q-$functions enables us to develop even more powerful bivariate distributions. For data exhibiting overdispersion, the assumption of Poisson marginals might not be well suited. We thus further allow for other marginals, e.g., the negative binomial distribution. For this distribution, a candidate $q-$function is

$$q_{nb}(x_i) = \begin{cases} -\phi_i \left(\frac{\mu_i}{\phi_i + \mu_i}\right) & \text{if } x_i = 0 \\ 1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i > 1, \end{cases}$$

where μ_i is the mean and ϕ_i the overdispersion parameter of X_i, for $i = 1, 2$.

Another characteristic of the $q-$ functions considered so far is that they weight only a finite number of pairs. Following Ting Lee (1996), we can use

$$q_{Sar}(x_i) = \exp(-x_i) - L_i(1),$$

$x_i \in \mathbb{N}_0$, to shift probabilities across the entire support. Here, $L_i(1)$ is the value of the Laplace transform of the marginal distribution at $s = 1$, that is $L_i(s) = E\left(e^{-sX_i}\right) = \sum_{x_i=0}^{\infty} \exp(-sx_i)P(x_i)$ with $P(\cdot)$ denoting the pmf of the $i-$th marginal distribution.
Moreover, for X_i being negative binomial distributed, we can generalize q_{Sar} and construct an **A**lternative **N**egative binomial **S**armanov (ANS) distribution by considering

$$q_{ANS}(x_i) \;=\; \left(\phi_i/(\phi_i + \mu_i)\right)^{x_i} - c_i,$$

for $c_i \;=\; \left(\frac{\phi_i}{\phi_i+\mu_i}\right)^{\phi_i} \left[1 - \left(1 - \frac{\phi_i}{\phi_i+\mu_i}\right) \frac{\phi_i}{\phi_i+\mu_i}\right]^{-\phi_i}$, $x_i \in \mathbb{N}_0$, $i = 1, 2$.

When comparing the Sarmanov model with the Laplace transform and negative binomial marginals with the ANS model, it can be seen that the latter model places more emphasis on clear wins rather than scoreless draws and close wins.

3 Application

To demonstrate the feasibility of the proposed models and their usefulness for practical applications, we fit the presented models to women's football scores from the seasons 2011/12-2018/19 and 2021/22 of the first women leagues in England, Germany, France and Spain. We observe several peculiarities in the data which are different to men's football scores: First, 0-0s are clearly *under*represented in women's football while clear wins with one team conceding no goals are *over*represented. The Chi-squared test rejects the null hypothesis of independence for all leagues except the English one. Second, the data exhibit overdispersion. Third, we find a substantial negative correlation in all leagues (-0.269 (England), -0.352 (Germany), -0.395 (France), and -0.263 (Spain)). When fitting the models proposed in Section 2 to our data, we include team-specific effects and a home-team effect. Table 1 summarises the AIC results for the different models and leagues. We find that the extended DC model emerges as the most promising model for the English league according to the AIC. This is most likely caused by only a minor amount of overdispersion in the data compared to the other leagues. Except for England, the ANS model provides the best model fit to women's football data as it is able to capture overdispersion while still being flexible enough to model the underrepresentation of scoreless draws and the overrepresentation of high wins. These two properties cannot be captured by the classical Dixon & Coles model which renders is not suitable for modelling women's football scores.

TABLE 1. The table displays the AICs for the models fitted. Bold values indicate the models preferred by the AIC.

	England	Germany	France	Spain
double Poisson	4016.23	7348.87	7110.91	13529.97
double NB	4017.61	7334.03	7104.77	13518.99
D&C Poisson	4018.07	7350.77	7112.88	13531.52
D&C Poisson with $\hat{q}$	4018.10	7350.86	7112.68	13531.48
D&C Poisson with $q^{(2)}$	**4014.14**	7350.87	7112.86	13531.95
D&C NB with q_{nb}	4019.39	7335.94	7106.74	13520.57
Sarmanov Poisson	4016.31	7340.10	7112.18	13529.48
Sarmanov NB	4017.59	7324.81	7105.95	13518.47
ANS	4018.12	**7321.30**	**7102.81**	**13515.88**

To check the adequacy of the model we fit the ANS model to two-thirds of the 2021/22 season in Germany, simulate the remaining matches based on the fitted model and calculate the final points. We find that the predicted 95% confidence intervals (obtained via the Monte Carlo simulations) include the true final points for each team (see Figure 2).

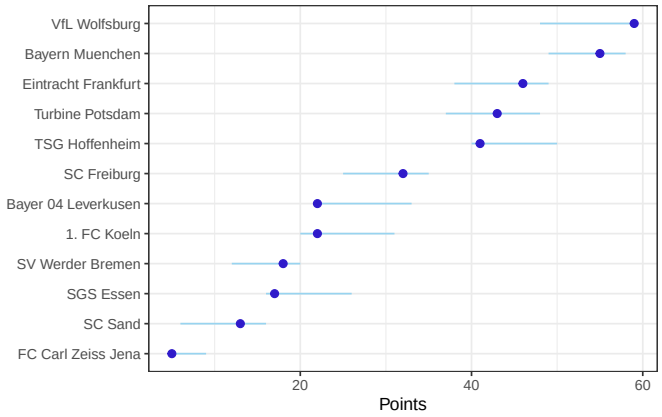


FIGURE 2. The plot displays the true final points (dark blue points) and the 95% confidence intervals (light blue lines) of the simulated final tables of the German Frauen-Bundesliga under the ANS model.

4 Discussion

This paper presents several extensions of the Dixon and Coles model, which is a widely used model for predicting football scores. Our extensions are

based on the Sarmanov family. We find that (in contrast to the DC model) the ANS model, which changes probabilities for a wider range of values and uses negative binomial marginals, is well suited for modeling and predicting women's football scores. In the future, to further improve predictions of women's scores it could be helpful to consider, e.g., the teams' recent performances. In this context, one possible approach is to adopt the method proposed by Dixon and Coles (1997), which involves giving less weight to matches that occurred further in the past. Alternatively, one could consider using latent states to model a team's form, similar to the approach considered by Ötting et al. (2021).

References

Bermúdez, L. and Karlis, D. (2021). Multivariate INAR(1) regression models based on the Sarmarnov distribution *Mathematics*, **9**(5):505.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **46**, 265–280.

Groll, A., Kneib, T., Mayr, A. and Schauberger, G. (2018). On the dependency of soccer scores–a sparse bivariate Poisson model for the UEFA European Football Championship 2016. *Journal of Quantitative Analysis in Sports*, **14**, 65–79.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models *Journal of the Royal Statistical Society: Series D (The Statistician)*, **52**, 381–393.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, **36**, 109–118.

Ötting, M., Langrock, L. and Maruotti, A. (2021). A copula-based multivariate hidden Markov model for modelling momentum in football. *AStA Advances in Statistical Analysis*, **107**, 9-27

Sarmanov, O. V. (1966). Generalized normal correlation and two-dimensional Fréchet classes. *In Doklady Akademii Nauk*, **vol. 168**, 32–35. Russian Academy of Sciences.

Ting Lee, M.-L. (1996). Properties and applications of the Sarmanov family of bivariate distributions. *Statistics and Computing*, **30**(5):1419-1432

Using measures of effect size and decision trees for variable selection

Annette Möller[1], Ann Cathrice George[2], Jürgen Groß[3]

[1] Bielefeld University, Germany
[2] Institute for Quality Assurance of the Austrian School System (IQS), Austria
[3] University of Hildesheim, Germany

E-mail for correspondence: annette.moeller@uni-bielefeld.de

Abstract: When analysing educational data possible challenges include the selection of useful predictor variables for a (multilevel) regression model as well as coping with extremely large sample sizes. This work extends a previous study and proposes to assist variable selection and quantification of relevance of these variables in a regression model with tree-based methods as well as effect size measures. The respective effect size measures were generalized to be also applicable in the setting of mixed linear models (multilevel models). When predicting math competencies of 4th grade students in Austria, this novel procedure yields an improved model fit as well as relevant insights about factors that influence math competencies. Future work will investigate the potential of the proposed procedure when applied to high-dimensional settings with several hundred variables.

Keywords: Effect Size; Variable Selection; Regression Model; Decision Trees; Math Competency.

1 Introduction

A recent project investigated the use of classification trees as a tool to assist variable selection for predicting a binary response (Möller et al., 2022). The study revealed that the classification tree based variable selection is able to improve the performance of (multilevel) regression models, which originate from a theoretical educational approach. Furthermore the variables identified as relevant by the classification trees were in accordance with educational findings. In this study the approach of Möller et al. (2022) is refined: a stepwise variable selection procedure is presented, which takes into account the knowledge from educational theory, the variable selection from decision trees and effect size as a measure of relevance for predictors.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

The new procedure is presented by the example of predicting students' mathematical competency based on a restricted set of variables. Further research will reveal the full potential of the proposed procedure when applied to high-dimensional settings consisting of several hundred variables.

2 Educational Research Questions and Data

The data analyzed in this study is obtained from the Austrian educational standards test in mathematics for fourth graders in 2018 (BIFIE, 2019). The Austrian standards testing is mandatory and leads to a complete survey of 73,780 students in 4,925 classes and 2,961 schools. The data includes an overall score for students' competencies in mathematics (`PVM4`), measured on a continuous scale with mean 551.4868 and standard deviation 98.5153. Additional background information of students, teachers, parents and schools is collected via so-called context questionnaires (BIFIE, 2018). This study includes a subset of 39 variables (plus an anonymized ID variable for the class and school).
Starting point of the case study on the proposed selection procedure is a benchmark model to predict the math competency (`PVM4`) of students. This model includes determinants of school performance, such as gender (`geschlecht`), migration background (`mig`), mother tongue (`spr-ndeu`), social status , (`sozstat`) and degree of urbanization at students' place of residence (`urban`). These variables are determinants which are known from educational theory to influence (math) competence of students (Brühwiler and Helmke, 2018). In a stepwise procedure the benchmark model is enlarged by adding further variables from the data set to improve the fit of the model to the given data. The additional variables are also analyzed in terms of their relevance within the model as well as in terms of educational knowledge possibly gained from the extended model.

3 Measures of Effect Size

A widely known and applied measure is Cohens's d (Cohen, 1988). It is an effect size measure for the two-sample t test with equal variances for the null hypothesis $H_0 : \mu_1 = \mu_2$ versus the alternative $H_1 : \mu_1 \neq \mu_2$. It is related to the test statistic t by the formula $d = t\sqrt{\frac{n_1+n_2}{n_1 n_2}}$, where n_1 and n_2 are the sample sizes in the two groups. According to Cohen, values $|d| = 0.2$, $|d| = 0.5$ and $|d| = 0.8$ indicate a small, medium and large effect, respectively.
A corresponding measure when the variable of interest (response) depends on further independent variables is given by Cohen's f^2, which is based on the F test of a linear hypothesis. It measures the effect size of one or multiple independent variables given a second set of independent variables.

The measure can be computed as $f^2 = \frac{R^2 - R_0^2}{1-R^2}$, where R^2 is the coefficient of determination from the complete model with both sets of variables and R_0^2 is the coefficient of determination in the reduced model containing only the second set of variables. Cohen suggests values $f^2 = 0.02$, $f^2 = 0.15$ and $f^2 = 0.35$ for a small, medium and large effect, respectively. Groß and Möller (2023a) propose a generalization d^* of Cohen's d for a grouped variable that additionally depends on sets of independent variables. The generalized version d^* has an exact relationship with the measure f^2. Furthermore, Groß and Möller (2023b) consider an adaptation of f^2 for a linear mixed model (multilevel model) containing additional random effects. The authors outline that f^2 can be obtained in a unifying framework applicable for the classical fixed effect linear model as well for a random effects model. In both cases f^2 can be computed from an F statistic for a linear hypothesis. However, in case of a random effects model the F statistic is defined subject to the estimate of the covariance matrix of the response, which depends also on the covariance structure of the random effects. Given an estimate of the covariance, f^2 can directly be obtained as defined above, based on the coefficient of determination R^2.

4 Application to Prediction of Math Competency

For predicting the math competency `PVM4` the benchmark model

$$M_1 : \texttt{PVM4} \sim \texttt{geschlecht} + \texttt{mig} + \texttt{sozstat} + \texttt{spr-ndeu} + \texttt{urban}$$

contains only predictors chosen by educational theory. A predictor is subsequently added to the model, yielding a nested model sequence. The candidate variables to be added to the current model are split variables in a regression tree grown on the full data set. They were added to the current model in the order they appeared in the tree: the variable `ma-sk-mean` was chosen for the very first split, `maueb-wh-nahi` was chosen in the second split, and `bil-hoechst` in the third one.

In the first step the mathematical self-concept (`ma-sk-mean`) is added to M_1, yielding the model M_2. Then the number of hours for private tutoring (`maueb-wh-nahi`) is added to M_2, yielding model M_3. In the final step the parents aspiration of the highest education their children will achieve (`bil-hoechst`) is added to M_3, yielding the largest model M_4. The corresponding multilevel models additionally contain a random intercept on school level.

Table 1 shows the (adjusted) coefficient of determination R^2 for each model, and Cohen's f^2 for the additional variable given the predictors in the preceding model – for the fixed effect model and the respective one with additional random intercept. It is clearly visible that adding tree predictors increases the fit to the data in terms of R^2 for both, the pure fixed effect and the random intercept model.

TABLE 1. (Adjusted) Coefficient of Determination R^2 of the four models as well as Cohen's f^2 of fixed effect model and model with additional random intercept on school level, for the additional predictor added in each subsequent model.

Model (additional predictor)	R^2	R^2_{random}	f^2	f^2_{random}
M_1	0.2412	0.1938	-	-
M_2 (`ma-sk-mean`)	0.4023	0.3906	0.2695	0.3257
M_3 (`maueb-wh-nahi`)	0.4267	0.4166	0.0426	0.0446
M_4 (`bil-hoechst`)	0.4448	0.4367	0.0325	0.0356

Adding only the first variable `ma-sk-mean` already increases the fit substantially compared to the benchmark model. The effect size of `ma-sk-mean` is between medium and high. The effect sizes of the other two variables are small. In a variable selection procedure based on f^2 this suggests to add `ma-sk-mean` to the model, but refrain from adding the other two, although the F-test identifies them as highly significant with p-value ≈ 0. Adding a random intercept to the model yields an even more pronounced effect, for each of the added variables.

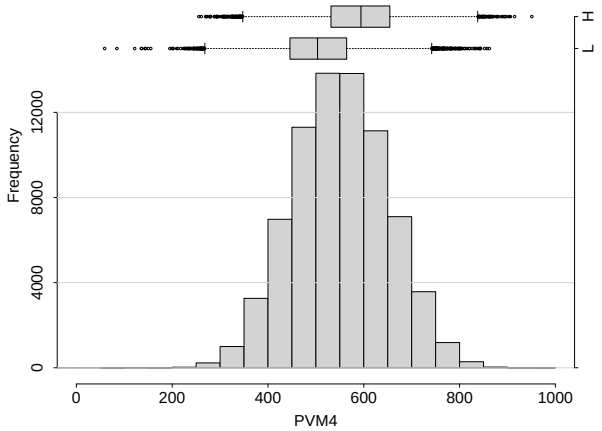


FIGURE 1. Histogram of PVM4 and boxplots of PVM4 in the two groups of `ma-sk-mean`, with `L` (low) denoting values smaller than the mean and `H` (high) denoting values larger than the mean.

Above observations illustrate clearly that the strength of an effect of a variable can be influenced by the presence of additional (fixed or random) variables. If for example an additional continuous or a grouping variable (as e.g. represented by a random or a fixed effect) contains information about

the response that the other variables in the model cannot fully capture, adding this new variable peels out the effect of the variable of interest even clearer. Although the actual value of f^2 (for both types of models) depends on the order in which the variables are added, the effect of adding `ma-sk-mean` is always largest, regardless in which step it is added to the current model.

As Cohen's d is more popular among educational scientists than Cohen's f^2, an alternative approach to investigate the relevance of `ma-sk-mean` is its dichotomization so that d or the generalized version d^* can be computed. When using as cut point for the two resulting classes the mean of the variable in the data set this results in the effect size $d = 0.9799$, that is, the two groups resulting from dichotomization have a strong effect on the math competency. The generalized Cohen's d^* allows to assess the effect size of above dichotomized variable `ma-sk-mean` given additional predictors, in this case the ones in the original model M_1, yielding a value $d^* = 0.8558$. This also indicates a strong (but slightly smaller) effect of the binary variable given the educational predictors. Figure 1 shows the empirical distribution of `PVM4` together with boxplots of the two groups resulting from the dichotomization. The boxplots indicate a substantial difference between the two groups `L` (values lower than the mean) and `H` (values larger than the mean), which is confirmed by the small p-value of the respective t-test and the large effect according to Cohen's d or d^*.

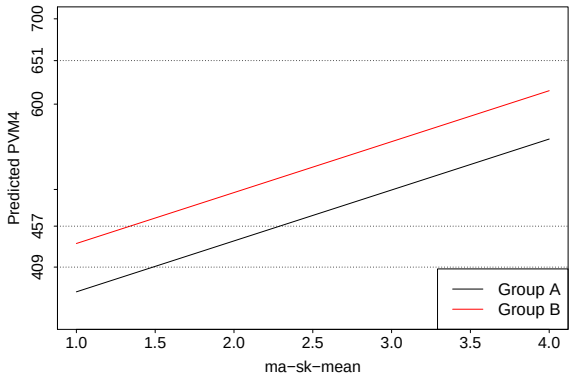


FIGURE 2. Predicted `PVM4` by the regression model M_2 across the range of `ma-sk-mean` with other predictors fixed at two sets of values.

The impact of `ma-sk-mean` on math competency is not contradictory to existing educational theories (Ahrens et al., 2016). Nonetheless, this variable would not be chosen based on an educational model building approach.

Figure 2 shows the predicted values of `PVM4` with model M_2 for the range of possible values of `ma-sk-mean`, while the other predictors are held constant at two sets of values, each yielding a predicted regression line. The two sets

of predictor values represent students with disadvantageous (group A) and advantageous context variables (group B). In both groups the mathematical self-concept has a strong influence on the level of math competence. Especially for the group of students with disadvantageous context factors an increase of the self-concept from value 1 to value 4 leads to an increase of two levels in math competence (see BIFIE, 2018) from "standards not achieved" (below bottom dashed horizontal line) to "standards achieved" (between middle and top dashed horizontal line). Thus, it is indeed beneficial to include this variable as predictor for `PVM4`.

Acknowledgments: We thank the Federal Institute for Quality Assurance of the Austrian School System for providing the data via the research data library (https://iqs.gv.at/fdb).

References

Arens, A. K., Marsh, H. W., Pekrun, R., Lichtenfeld, S., Murayama, K. and vom Hofe, R. (2016). Math Self-Concept, Grades, and Achievement Test Scores: Long-Term Reciprocal Effects Across Five Waves and Three Achievement Tracks. *Journal of Educational Psychology*, **109**, 621 – 634.

BIFIE (2018). *Context questionnaires of educational standards test 2018 in mathematics for grade four*. Bundesinstitut BIFIE Austria.

BIFIE (2019). *Standardüberprüfung 2018. Mathematik, 4. Schulstufe. Bundesergebnisbericht.*. Bundesinstitut BIFIE Austria.

Brühwiler, C. and Helmke, A. (2018). Determinanten der Schulleistung. In: *D. H. Rost, J. R. Sparfeldt and S. R. Buch (Hrsg.), Handwörterbuch Pädagogische Psychologie.*, Beltz Psychologie Verlags Union, 78 – 92.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Erlbaum Associates.

Groß, J. and Möller, A. (2023a). A Note on Cohen's d From a Partitioned Linear Regression Model. *Journal of Statistical Theory and Practice*, **17**, 22, DOI: https://doi.org/10.1007/s42519-023-00323-w.

Groß, J. and Möller, A. (2023b). Effect Size Estimation in Linear Mixed Models. arXiv:2302.14580.

Möller, A. George, A.C. and Groß, J. (2022). Predicting and explaining school transition rates in Austria with classification trees. *International Journal of Research and Method in Education*, DOI:10.1080/1743727X.2022.2128744.

A comparison of time series forecasting models on industrial process data

Jack Moore[1], Jamie Wilson[1], Norma Bargary[1], Kevin Burke[1]

[1] University of Limerick, Ireland

E-mail for correspondence: Jack.D.Moore@ul.ie

Abstract: The goal of this paper is to investigate the performance of various time series models on the basis of their ability to predict future data, in the context of Industrial processes. The processes in question were explored in terms of both daily and hourly data. The motivation for this choice being the varying predictions that may be required in an industrial setting. We considered several models, including SARIMA, exponential smoothing and TBATS. We used simulated data based off real-world industry data to train our models. This data had a strong weekly pattern, but no significant trend. We then compared these models based on their RMSE values produced in relation to another period of test data. For our daily data, we found that the TBATS model generally outperformed the SARIMA and Exponential Smoothing models, however, there were exceptions. When considering hourly data, the only viable model out the ones we considered was the TBATS model. The results obtained during this research can be used to inform decisions in relation to industries seeking to create accurate predictions for various processes. However, further research should be used to explore the limitations of the models examined in this paper.

Keywords: Seasonal ARIMA; Exponential Smoothing; TBATS.

1 Introduction

The goal of this paper is to explore the types of time series models that may be applied in an industry setting. As a time series can be defined as a series of values occurring at successive times, such a definition could be applied to many industry processes. Therefore modelling and predicting such data could lead to benefits, such as detecting unwanted trends or unusual deviations from previous data. Such models include ARIMA, Exponential Smoothing, and TBATS, with TBATS being a special case of our Exponential Smoothing models. In section 2, we will explore these models in

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

greater detail before using them to model and predict data that was simulated based off real-world industry data. This data had a weekly seasonality and no significant trend. We also considered this data as both hourly and daily, to demonstrate the flexibility of such models. The accuracy of these predictions was then compared via RMSE values in section 3, through multiple simulations of our data, using functions from packages 'forecast' and 'TSA' in R. This is followed by a brief conclusion of our results in section 4.

2 Exploration of Time Series Models

Within this section, we will explore the time series models we wish to compare. The data we considered for our comparison had a weekly pattern, and no significant trend.

The SARIMA model can be represented as combination of a standard ARIMA model with a Seasonal ARIMA model, ARIMA(p,d,q)X(P,D,Q)s, where p and P represent the number of lag components we include in the model, d and D represent the number of times the data was differenced, q and Q represent the the number of components of the Moving Average model we include, and s represents the period for our seasonal difference:

$$\phi(B)\Phi(B)(1-B)^d(1-B^s)^D Y_t = \theta(B)\Theta(B)e_t$$

Note: B represents the backward shift operator.

There is no one single general equation of exponential smoothing models, as the models which fall under this category all differ to some extent. The common theme being that the models predicts future values based off some weighted combination of past values. Two models did appear to fit the data more frequently than the other models in this category. A simple exponential smoothing model with additive errors, and an exponential smoothing model with additive errors and additive seasonality, however, other choices did appear. Starting with our simple exponential smoothing model with additive errors and additive seasonality:

$$y_t = l_{t-1} + s_{t-m} + \epsilon_t$$
$$l_t = l_{t-1} + \alpha\epsilon_t$$
$$s_t = s_{t-m} + \gamma\epsilon_t$$

where l_t is the level at time t, s_t is our seasonal component, α and γ are the smoothing parameters and m is our period.

For our simple exponential smoothing model with additive errors, we simply remove our seasonality component.

The last model we consider is the TBATS model, which builds on the above exponential smoothing models by incorporating Fourier terms into the model. This change leads to certain benefits, one of which is the ability to handle larger seasonal periods.

2.1 Daily Data

For our daily data, all the models we considered were suitable, as we were dealing with a weekly pattern. This allowed us to explore the performance across each of these models.

2.2 Hourly Data

In the case of hourly data, the weekly pattern turned into a period of 168. As such, our previous models were not suitable. Therefore, the TBATS model was the only viable option. Due to this, we will not draw comparisons between the RMSE values. However, we will still compare the forecasts this model produces to the real data obtained from the following week.

3 Comparison of Time Series models

Now that we have discussed the models we wish to use, and the context in which we could use such models, we can compare their performance in relation to daily data, via RMSE values.

TABLE 1. Example RMSE values

Model	RMSE values
SARIMA	0.297
ETS	0.273
TBATS	0.267

In Table 1, we have 3 RMSE values, obtained from one simulation of our data. As we can see, the TBATS has performed the best, followed by exponential smoothing and SARIMA. This pattern persisted throughout the majority of the simulations we ran. Nonetheless, we see in Figure 1 and Figure 2 that the ARIMA and ETS models were still able to give accurate predictions for the next week of data, despite being outperformed by the TBATS model.

For our hourly data, as mentioned in section 2, the SARIMA and Exponential Smoothing models were not suitable for this model. As such, we cannot compare the RMSE values from the TBATS to gauge how well it performed. However, we can still compare our predictions with the actual data.

As we see in Figure 3, the TBATS model has managed to accurately predict the next week of data. This result was repeatedly obtained through multiple simulations.

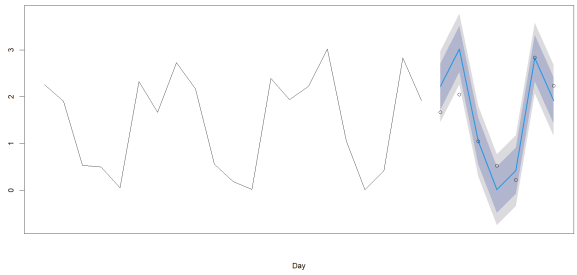


FIGURE 1. ARIMA predictions Vs Test Data

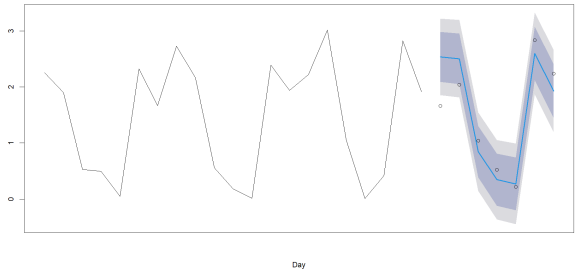


FIGURE 2. ETS predictions Vs Test Data

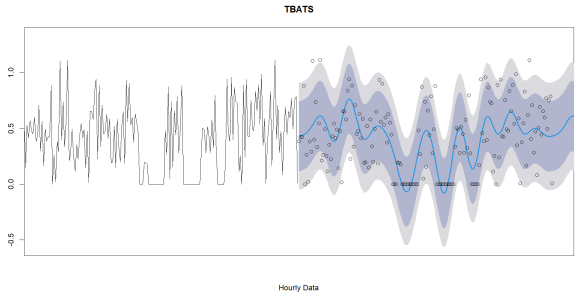


FIGURE 3. TBATS Prediction Vs. Test Data

4 Conclusion

In regards to daily data, we found that TBATS model typically performed the best, followed by exponential smoothing models, and SARIMA models. However, we found that all the models we considered were able to give reliable predictions. It is also worth noting that the TBATS model will not always be the most suitable model for predicting industry processes.

When considering other processes within this research, the SARIMA model became the ideal choice of model. As such, this paper has demonstrated the benefits of applying such time series models to industry processes, however, further research should be used to explore the limitations of such models. Another limitation of this short paper is the lack of exploration in relation to models which can model large periods, such as the 168 period of hourly weekly data. Other models which could be explored in this regard may include dynamic regression models, and GP models. Another avenue of exploration within this area would be the comparison of performance from the individual models listed within this paper, to different combinations of those same models. Although such models typically lead to more accurate predictions, the time required to create such predicitons may not be ideal in an industry setting.

References

Hyndman, R.J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Comparing trial and variable association in contingency table data using multinomial models for clustered data

Darcy Steeg Morris[1], Andrew M. Raim[1]

[1] U.S. Census Bureau, Center for Statistical Research and Methodology, USA

E-mail for correspondence: darcy.steeg.morris@census.gov

Abstract: Multinomial models are used for analysis of contingency table data with probability specifications designed to allow dependence between variables. These models, however, assume that there is no association between underlying trials, which is likely violated in clustered data. We assess the flexibility of alternative categorical data models that relax the assumption of trial independence. The conceptual probability mechanisms for trial association in a couple of categorical data distributions that allow for data dispersion are discussed. Through analysis of simulated data we explore the utility of modeling trial association possibly as a substitute for higher-order variable dependencies.

Keywords: categorical data analysis; count data; clustered data; data dispersion.

1 Introduction

Analysis of multiple categorical outcome variables is often done through modeling of contingency table data. These aggregated cross-tabulations of qualitative information can be assumed to arise from a variety of sampling mechanisms such as Poisson, multinomial, and product multinomial (Agresti, 2012). In part due to computational and notational simplicity, a Poisson loglinear model is commonly fit to the table cell counts with the mean parameter structured to allow for relationships between variables. The Poisson sampling mechanism specifies that the cell counts are independent and that the number of trials populating the table is random. Variable relationships estimated with a multinomial model will be the same as for the Poisson model, however, multinomial sampling assumes that there are a fixed number of underlying independent trials.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

We are interested in analyzing contingency table data using extensions of the multinomial distribution that allow trial association. Using these extended multinomial distributions, we empirically explore two sources of cell dependencies and their relation to each other: the (in)dependence structure between the variables measured through the probability parameters and the association between trials measured through an association parameter.

The Census Bureau collects household and person characteristics that are often categorical: for example race, sex, and age in the Decennial Census. Characteristics within a cluster – e.g. a group of people in a household or a group of households in a low-level geography – may be related through either observed or unobserved information. Observed characteristics that distinguish clusters can be incorporated in traditional contingency table analysis as an additional dimension in the table (i.e. through specification in the multinomial probabilities); however unobserved characteristics would be ignored with the assumptions that the trials – e.g. people or households – are independent. We are interested in flexible models that allow for the possibility of association through clustering that is not directly observed. Such models may offer improvement over traditional joint imputation models for categorical variables in the presence of missing data.

2 Contingency Table Notation and Assumptions

Consider two nominal categorical variables X_1 and X_2 with I and J categories, respectively. The corresponding two-way contingency table counts the occurrence of each of IxJ possible combinations of X_1 and X_2. Table 1 shows the 3x3 table that cross-classifies $n = \sum_{i=1}^{3}\sum_{j=1}^{3} n_{ij}$ trials.

TABLE 1. Two-Way Table Structure with $I = J = 3$.

		X_2		
		1	2	3
X_1	1	n_{11}	n_{12}	n_{13}
	2	n_{21}	n_{22}	n_{23}
	3	n_{31}	n_{32}	n_{33}

Let $\pi_{ij} = P(X_1 = i, X_2 = j)$ be the joint probability that a trial occurs in cell (i, j); $\pi_{i+} = P(X_1 = i) = \sum_{j=1}^{J} \pi_{ij}$ be the marginal probability for X_1; and $\pi_{+j} = P(X_2 = j) = \sum_{i=1}^{I} \pi_{ij}$ be the marginal probability for X_2. The joint probability mass function for the table of counts assuming multinomial sampling with fixed sample size n is

$$P(\mathbf{X} = \mathbf{x}) = \frac{n!}{n_{11}! \ldots n_{IJ}!} \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{n_{ij}},$$

where we let $\mathbf{X} = (n_{11}, \ldots, n_{1J}, \ldots, n_{I1}, \ldots, n_{IJ})$ be the collection of all IxJ cell counts n_{ij}. Dependence between variables is captured through assumptions on the structure of the joint probabilities. For example, the independence model defines each cell probability as the product of the two marginal probabilities: $\pi_{ij} = \pi_{i+}\pi_{+j}$ as opposed to the saturated model that does not have any simplification of π_{ij}. The multinomial model for contingency table data assumes that the underlying trials are independent. That is, each trial follows the same one-trial multinomial probability distribution with parameters $\{\pi_{ij}\}$. The specification of the cell probability structure may or may not assume dependence structure between variables, but trials are always assumed to be independent.

3 Flexible Multinomial Distributions

We discuss some flexible alternatives to the multinomial (MN) distribution that allow for potential excess variation as compared to the multinomial distribution – variation that may be caused by trial association. We describe the probability mechanism for each distribution and focus on how it accounts for potential trial association, setting aside for now dependence modeled through the specification of the cell probabilities.

Dirichlet-Multinomial (DM)

The Dirichlet-Multinomial (DM) is a multivariate analogue of the beta-binomial distribution that arises from a Pólya urn scheme (Mosimann, 1962). Each trial is a draw from an urn where the probability of drawing a particular category changes each time a draw occurs depending on the category of the current draw (i.e. double replacement). A vector $\mathbf{x}$ of category counts aggregated over the n draws follows a DM distribution. The dynamic nature of the category probabilities induces dependence in the trials, whereas a static set of probabilities would yield the multinomial distribution. The DM probability mass function can be written as

$$P(\mathbf{X} = \mathbf{x}) = \frac{\Gamma(c)\,\Gamma(n+1)}{\Gamma(n+c)} \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{\Gamma(n_{ij} + \alpha_{ij})}{\Gamma(\alpha_{ij})\,\Gamma(n_{ij}+1)},$$

where $\alpha_{ij} = c\ \pi_{ij}$, $c = \rho^{-2}(1-\rho^2)$ and $0 < \rho < 1$ (Neerchal and Morel, 1998). In this parameterization, ρ controls the extent of the departure from with multinomial, which is the special case at $\rho = 0$. With respect to the Pólya urn scheme, ρ controls the effect of the replacement: $\rho \nearrow \ \Rightarrow c \searrow \ \Rightarrow$ fewer objects in the urn $\Rightarrow$ larger effect of replacement on the category probabilities, i.e. more clustering. DM shares the same first moment as the multinomial distribution, but the dispersion/association parameter allows the variance to inflate relative to the multinomial distribution: $Var(\mathbf{X}) = n\left[1 + \rho^2(n-1)\right]\left[Diag(\pi) - \pi\pi^{\mathrm{T}}\right]$.

Random-Clumped Multinomial (RCM)

The Random-Clumped Multinomial (RCM) is based on a finite mixture of multinomials. Trials in a cluster are randomly "clumped" to a common but randomly selected category, whereas the remaining trials are assigned independently (Morel and Neerchal, 1993; Nagaraj et al., 1998). The probability mechanism can be notated as $\mathbf{X} = N\mathbf{X}^* + \mathbf{X}^{**}$, where $\mathbf{X}^* \sim MN(1, \pi)$, $\mathbf{X}^{**} \sim MN(n-N, \pi)$, and $N \sim Bin(n, \rho)$. The parameter ρ – the binomial success probability for the number of clustered observations N – controls the extent of the departure from multinomial, which is the special case at $\rho = 0$. Trial association is driven by the random number of trials N that have the same category as defined through $\mathbf{X}^*$. RCM moments take the same form as DM.

4 Analysis of Simulated Data

Through a simple simulation set-up we empirically assess and compare trial and variable association in data with varying levels of both types of dependence. We use a simple two-way contingency table with categorical variables X_1 and X_2, with $I = J = 2,\ 3$, or 4 levels. We generate contingency table data in two scenarios:

Scenario A, DM/RCM independence data (trial dependence only):

$$\log \pi_{ij} = \lambda_0 + \lambda_i^{X_1} + \lambda_j^{X_2} \quad \text{with} \quad \rho \in (.01, .25),$$

Scenario B, MN saturated data (variable dependence only):

$$\log \pi_{ij} = \lambda_0 + \lambda_i^{X_1} + \lambda_j^{X_2} + \lambda_{ij}^{X_1 X_2},$$

where $\lambda_1^{X_1} = \lambda_1^{X_2} = 0$ and λ_0 is defined to ensure the sum-to-one constraint on the cell probabilities. Scenario A main effects are randomly generated from $\mathcal{N}(0, 1)$; whereas Scenario B main and interactions effects are set to zero to isolate the interaction effect $\lambda_{22}^{X_1 X_2}$ which is varied from -1.5 to 1.5. We generate $R = 200$ simulated tables each with $n = 100$ or 1000 trials and fit MN, DM and/or RCM models. Data is generated and models are fit using computation methods described in Raim et al. (2015).

4.1 Estimating ρ for DM and RCM with Correct Model Specification

Large sample properties of the DM and RCM maximum likelihood estimates (MLEs) have been studied with respect to increasing the number of observed vectors of categorical counts. With only one set of categorical counts $\mathbf{X}$ observed in a contingency table, we consider MLE properties in the finite trial setting. Fitting correctly specified models on Scenario A simulated data, we find that the level of bias in estimating ρ for the DM and

RCM model fit on one data table depends on (1) the number of trials n, (2) the level of trial association, and (3) the degrees of freedom associated with the model: see Figure 1 and Figure 2. The degrees of freedom – the difference between the number of cells in the table and the number of parameters in the model – depends on the dimension of the table controlled through I and J. In a k-way table with $k > 2$, the degrees of freedom also depends on k and the the degree of variable dependence assumed in the model (e.g. mutual vs. conditional independence).

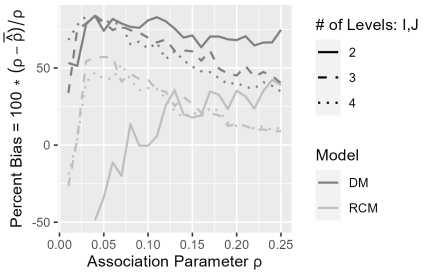


FIGURE 1. $\hat{\rho}$ % Bias, $n = 100$.

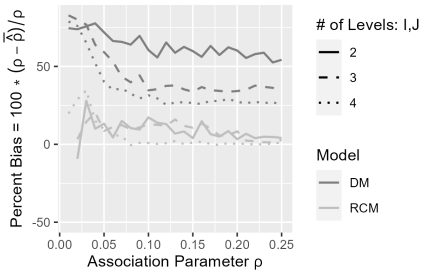


FIGURE 2. $\hat{\rho}$ % Bias, $n = 1000$.

4.2 Model Fit Comparisons with Incorrect Model Specification

Trial dependence only data (Scenario A) fit with variable dependence only model (MN saturated model).

We are interested in assessing the probability of rejecting the MN independence model ($H_0 : \lambda_{ij}^{X_1X_2} = 0$) in favor of the MN saturated model when data is simulated with only trial dependence (Scenario A). Figure 3 displays the empirical rejection rate from a likelihood ratio test with varying levels of trial dependence assuming $n = 1000$ trials. We find the MN model suggests significant interaction effects – indicating variable dependence – as trial association increases, even though variable independence was not assumed in generating the data. This suggests that the MN model accounts for the trial dependence by attributing it to variable dependence. This result is observed for a low degree of trial association when the dimension of the table is larger (i.e. I, J increases).

Variable dependence only data (Scenario B) fit with trial dependence only models (DM and RCM independence model).

We are also interested in assessing the DM or RCM model for varying levels of variable dependence in Scenario B simulated tables of $n = 1000$ trials. Figure 4 shows that, on average, the trial association parameter ρ is estimated to be greater than zero as the MN interaction effect gets further from zero. This suggests that DM/RCM accounts for variable dependence

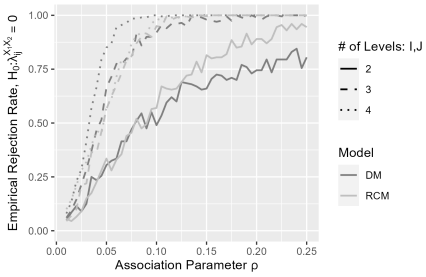


FIGURE 3. Empirical Rejection Rate, MN Fit on Scenario A Data.

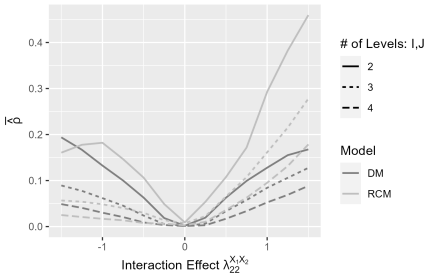


FIGURE 4. Mean $\hat{\rho}$, DM/RCM Fit on Scenario B Data.

by attributing it to trial dependence. This result is more pronounced for (1) the RCM model – recall from Section 4.1 that DM more often underestimates ρ; (2) a positive interaction effect with the RCM model – where an increase in $\lambda_{22}^{X_1X_2}$ directly translates to an increased probability (i.e. cluster) in one table cell; and (3) smaller tables – where the probability shift to/away from $(X_1 = 2, X_2 = 2)$ is spread over fewer table cells.

Figure 5 indicates that the empirical rejection rate of DM/RCM from a Pearson χ^2 test goes to one as the MN interaction effect goes away from zero. This suggests that the DM and RCM models account for some variable dependence through the trial dependence parameter ρ, but only in a significant way for smaller levels of the interaction effect. DM/RCM model evaluation though AIC similarly shows that DM/RCM performs just as well (or better) as the correctly specified MN saturated model at small levels of the interaction effect: see Figure 6. Interestingly, the RCM model performs just as well as the MN saturated model for large positive interaction effects, even though it is misspecified and relies on only one association parameter.

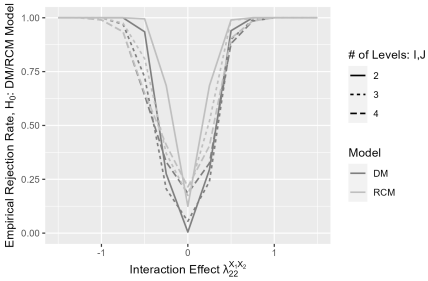


FIGURE 5. Empirical Rejection Rate, DM/RCM on Scenario B.

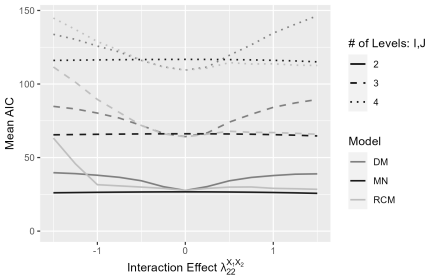


FIGURE 6. Mean AIC, Models Fit on Scenario B Data.

5 Discussion

Association in categorical variables may occur through dependence in trials and/or variables. Loglinear models for contingency tables are designed to assess relationships of variables through interaction effects capturing variable association assuming independent trials. Extended multinomial distributions such as DM and RCM that allow for dispersion – possibly caused by trial association – offer a potential alternative, particularly when trial clustering may be due to unobserved factors. However, DM and RCM encounter challenges in application to contingency table data due to the observation of only one count data vector.

In a two-way table simulation study we see that DM and RCM MLEs are sensitive to the number of trials and number of parameters in the model. Assuming $n = 1000$, we find that MN attributes trial dependence to variable dependence and DM/RCM attributes low-level variable dependence to trial dependence. The latter misspecification may be useful as the DM and RCM models depend on only one dispersion parameter rather than a set of interaction effects. Goodness-of-fit tests indicate that the one DM/RCM association parameter substitutes for the set of interaction effects in limited cases depending on the size of the table and the level of variable dependence. Further study of multi-way tables with associated varying levels of variable dependence assumptions (e.g. mutual vs. conditional) may provide better understanding of the utility of flexible multinomial models for contingency table data.

References

Agresti, A. (2012). *Categorical Data Analysis*. New York: John Wiley.

Morel, J.G. and Neerchal, N.K. (1993). A Finite Mixture Distribution for Modelling Multinomial Extra Variation. *Biometrika*, **80**, 363 – 371.

Mosimann, J.E. (1962). On the Compound Multinomial Distribution, the Multivariate β-distribution, and Correlations among Proportions. *Biometrika*, **49**, 65 – 82.

Nagaraj, K., Neerchal, N.K. and Morel, J.G. (1998). Large Cluster Results for Two Parametric Multinomial Extra Variation Models. *Journal of the American Statistical Association*, **93**, 1078 – 1087.

Raim, A.M., Nagaraj, K., Neerchal, N.K. and Morel, J.G. (2015). Modeling Overdispersion in R. *Technical Report HPCF-2015-1, UMBC High Performance Computing Facility, University of Maryland, Baltimore County*, https://hpcf.umbc.edu/publications.

Covariate-adjusted association of sensor outputs using a nonparametric estimate of the conditional covariance

Lizzie Neumann[1]

[1] Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany

E-mail for correspondence: neumannl@hsu-hh.de

Abstract: Sensor data for structural health monitoring typically depends on environmental influences such as temperature. This paper presents an approach for adjusting the association of sensor outputs using a nonparametric estimate of the conditional covariance matrix.

Keywords: Conditional Covariance; Confounding; Kernel Method; Sensor Data; Temperature Effect.

1 Introduction

In structural health monitoring, sensor data from structures such as bridges are used to monitor the condition of structures. As these measurements are typically not made under laboratory conditions, the data depend on environmental influences such as temperature.

FIGURE 1. Test Bridge UniBw M (Francesca Marsili, 2022)

FIGURE 2. Test Bridge UniBw M (Alexander Mendler, 2022)

Therefore, a model to adjust these covariates is required before the association between the sensor outputs can be analyzed.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

Viefhues et al. (2021) did a laboratory test in which they measured the acceleration on a concrete beam for five different temperatures in a climate chamber. Then they estimated the conditional covariance for these five temperature values and used interpolation for estimating the covariances for temperature values in between. In practice/reality, however, temperature is a continuous quantity. Section 2 hence presents an approach where the conditional covariance is estimated by use of a nonparametric, kernel-based technique. Section 3 illustrates the application to the Test Bridge UniBw M data set, the bridge with traffic simulation can be seen in Figures 1 and 2.

2 Conditional Covariance

Let $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ be a p-dimensional random vector describing p different sensor outputs and let z denote a potentially confounding covariate, such as temperature. First, let us assume that $\mathbf{x}$ and z are jointly normal, i.e.

$$\begin{pmatrix} \mathbf{x} \\ z \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{\mathbf{x}} \\ \mu_z \end{pmatrix}, \begin{pmatrix} \Sigma_{\mathbf{x}} & \Psi \\ \Psi^{\mathrm{T}} & \sigma_{zz} \end{pmatrix} \right) \tag{1}$$

with

$$\mu_{\mathbf{x}} = \begin{pmatrix} \mu_{x_1} \\ \vdots \\ \mu_{x_p} \end{pmatrix}, \quad \Sigma_{\mathbf{x}} = \begin{pmatrix} \sigma_{x_1 x_1} & \cdots & \sigma_{x_1 x_p} \\ \vdots & \ddots & \vdots \\ \sigma_{x_p x_1} & \cdots & \sigma_{x_p x_p} \end{pmatrix}, \quad \Psi = \begin{pmatrix} \sigma_{x_1 z} \\ \vdots \\ \sigma_{x_p z} \end{pmatrix}.$$

Then for the conditional distribution of $\mathbf{x}$ given z we have

$$\mathbf{x}|z \sim N\left(\mu_{\mathbf{x}} + \frac{1}{\sigma_{zz}} \Psi (z - \mu_z), \Sigma_{\mathbf{x}} - \frac{1}{\sigma_{zz}} \Psi \Psi^{\mathrm{T}} \right).$$

For estimating the *conditional* covariance of x_i and x_j given z

$$\sigma_{x_i x_j | z} = \sigma_{x_i x_j} - \frac{\sigma_{x_i z} \sigma_{x_j z}}{\sigma_{zz}},$$

we can use the empirical versions of $\sigma_{x_i x_j}$, $\sigma_{x_i z}$, $\sigma_{x_j z}$ and σ_{zz}.

However, the assumption (1) of $(\mathbf{x}, z)$ being jointly normal may be too restrictive. Therefore we relax the assumption by requesting that only the conditional distribution of $\mathbf{x}$ given z is normal. Then we have

$$\mathbf{x}|z \sim N\left(\mu_{\mathbf{x}}(z), \Sigma_{\mathbf{x}}(z)\right).$$

As a further generalization, we may even drop the distributional assumption of normality and focus on the conditional variances/covariance. Then, for

estimating the conditional covariance matrix $\Sigma(z)$ of $\mathbf{x}$ given z we can use a nonparametric, Nadaraya-Watson kernel estimator (Yin et al., 2010)

$$\hat{\Sigma}(z) = \left\{ \sum_{i=1}^{n} K_h(z_i - z) \left[\mathbf{x}_i - \hat{m}(z_i)\right] \left[\mathbf{x}_i - \hat{m}(z_i)\right]^{\mathrm{T}} \right\} \left\{ \sum_{i=1}^{n} K_h(z_i - z) \right\}^{-1}, \tag{2}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$, for $i = 1, \ldots, n$, are observations of $\mathbf{x}$ available with associated (e.g., temperature) measurements z_i. $K_h(\cdot)$ is a kernel function with bandwidth h, and $\hat{m}(z_i)$ is an estimate of the mean of $\mathbf{x}$ at z_i. For the latter, we can also use a kernel estimate in terms of

$$\hat{m}(z) = \left\{ \sum_{i=1}^{n} K_h(z_i - z)\mathbf{x}_i \right\} \left\{ \sum_{i=1}^{n} K_h(z_i - z) \right\}^{-1}. \tag{3}$$

At (2) and (3), we may use different bandwidths or even different bandwidths for different components of the conditional mean and conditional covariance for being adaptive in terms of smoothing (Yin et al., 2010). Also, we may use a completely different method for estimating the mean $m(z_i)$, for example, penalized regression splines (Neumann and Gertheiss, 2022).

3 Application to Data

The Test Bridge UniBw M, shown in Figures 1 and 2, is a 30-meter-long steel composite bridge on the grounds of the University of the Bundeswehr in Munich (UniBw M) (Jaelani et al., 2022). A joint group from Helmut Schmidt University, UniBw M, and the Technical University Munich collected the data. Among other things, from 11 March 2022 to 1 April 2022, the acceleration was measured with eight accelerometers in 1000 hertz and the air temperature in 1 hertz. The acceleration and temperature data were resampled to 100 hertz with the `resample` function of `signal` R-package which uses bandlimited interpolation (signal developers, 2013).

The conditional covariance of the acceleration data from the Test Bridge UniBw M is estimated as in Equation (2) with $p = 8$ acceleration sensor and bandwidth $h = 2.5$. To estimate the local mean of $\mathbf{x}$ at z_i, we use the Nadaraya-Watson kernel estimator as in Equation (3) with bandwidth $h = 1.5$.

Figure 3 shows the conditional covariance for six different temperatures, $z \in \{-5, 0, 5, 10, 15, 20\}$ in Celsius. There are small structural differences between the conditional covariances. These differences can be seen better if we have a look at the conditional covariance (or conditional correlation) as a function in z. Therefore, Figure 4 shows the conditional correlation as a function in z for $z \in [-5, 20]$. We can see that the correlation peaks at $-5°$C and then approaches zero for increasing temperatures. So for negative temperatures the sensors have a large (to small) correlation and nearly none for positive temperatures.

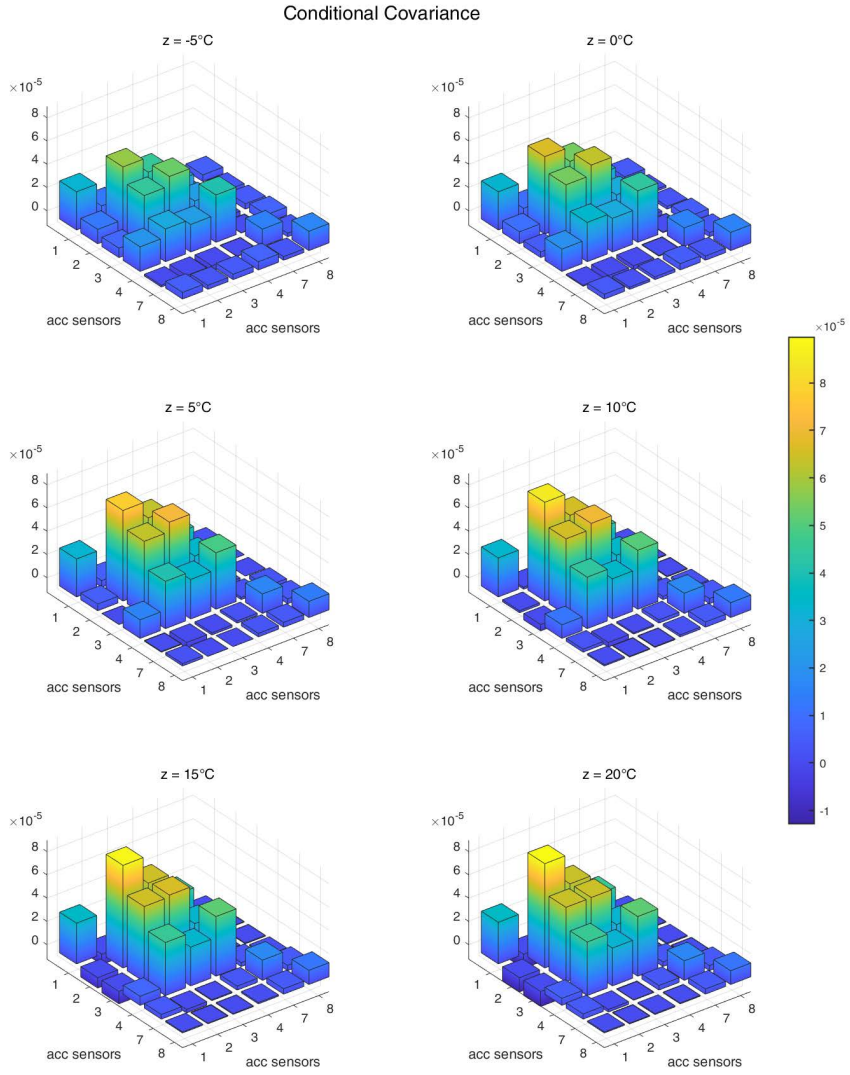


FIGURE 3. Estimates of conditional covariances for $z \in \{-5, 0, 5, 10, 15, 20\}$

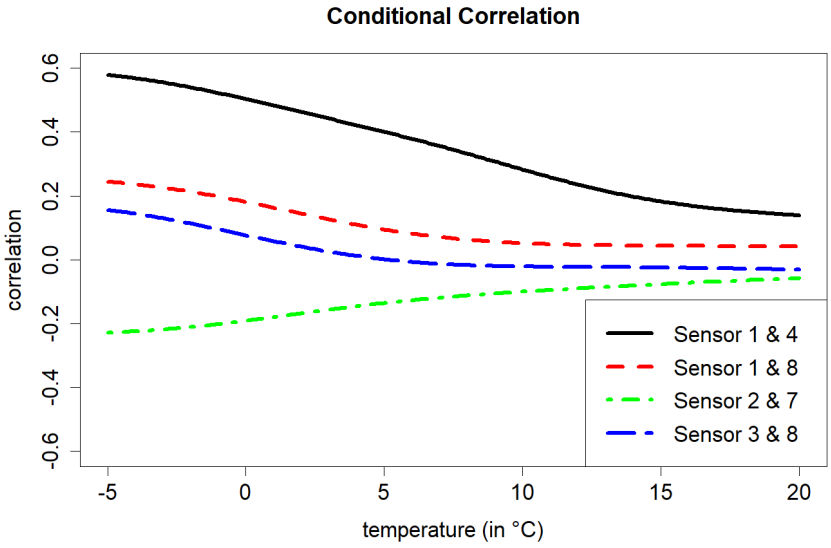


FIGURE 4. Estimated conditional correlations as function in temperature

Acknowledgments: This research paper out of the project SHM – Digitalisierung und Überwachung von Infrastrukturbauwerken is funded by dtec.bw - Digitalization and Technology Research Center of the Bundeswehr which we gratefully acknowledge. dtec.bw is funded by the European Union – NextGenerationEU. We thank Francesca Marsili and Alexander Mendler for providing the photos of the Test Bridge UniBw M.

References

Jaelani, Y., Klemm, A., Wimmer, J., Seitz, F., Köhncke, M., Marsili, F., Mendler, A., von Danwitz, M., Keßler, S., Henke, S., Gündel, M., Braml, T. and Popp, A. (2023). Developing a Benchmark Study for Bridge Monitoring. Accepted for Publication in: *Steel Construction*.

Neumann, L. and Gertheiss, J. (2022). Covariate-adjusted Association of Sensor Outputs for Structural Health Monitoring. In: dtec.bw-Beiträge der Helmut-Schmidt-Universität / Universität der Bundeswehr Hamburg: Forschungsaktivitäten im Zentrum für Digitalisierungs- und Technologieforschung der Bundeswehr dtec.bw – Band 1, 287-291, available from https://doi.org/10.24405/14566

signal developers (2013). signal: Signal processing. URL: http://r-forge.r-project.org/projects/signal/.

Viefhues, E., Döhler, M., Simon, P. Hermann, R., Hille, F., and Mevel, L. (2021). Stochastic subspace-based damage detection of a temperature affected beam structure. SHMII-10 2021-10th International Conference on Structural Health Monitoring of Intelligent Infrastructure, June 2021, Porto, Portugal. 1 – 6. hal-03276865.

Yin, J., Geng, Z., Li, R., and Wang, H. (2010). *Nonparametric Covariance Model*. In: *Statistica Sinica* 20, 469 – 479.

Bayesian probit models for preference classification: an analysis of chess players' propensity for risk-taking

Lennart Oelschläger[1], Dietmar Bauer[1]

[1] Bielefeld University, Germany

E-mail for correspondence: lennart.oelschlaeger@uni-bielefeld.de

Abstract: Probit models are widely used to analyze discrete choice behavior in fields such as transportation, marketing, and psychology. We propose a latent class model extension that allows for the classification of decider preferences. The model is estimated in a Bayesian framework, and the number of classes is determined by a Dirichlet process. In a simulation study, we verified that the dependence of the concentration prior diminishes as the number of deciders increases, resulting in stable inference. We further applied the proposed methodology in the context of chess, where chess players are classified according to three types of risk-taking propensity.

Keywords: Probit model; Dirichlet process; Preference classification; Chess.

1 Bayesian probit models

In this paper, we propose latent class probit models for the classification of decider preferences without requiring the explicit specification of the number of classes included in the model. Commonly rooted in the random utility framework, probit models assume that deciders assign utility values to discrete choice alternatives and seek to maximize them (Train, 2009). The utilities are modeled as a linear function of observable and unobservable factors, where the latter are assumed to follow a multivariate normal distribution. Specifically, decider n's choice $y_{nt} \in \{1, \dots, J\}$ at occasion t is explained through a matrix X_{nt} of choice characteristics as

$$y_{nt} = \arg\max U_{nt}, \quad U_{nt} = X_{nt}\beta + \varepsilon_{nt}, \quad \varepsilon_{nt} \sim N(0, \Sigma). \tag{1}$$

In the following, we assume that (1) has been normalized for level and scale (e.g., by taking utility differences w.r.t. a base alternative and fixing one

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

error-term variance, cf. Train, 2009, Section 5.2). Bayesian inference of the model parameters requires the computation of the posterior density

$$\Pr(\beta, \Sigma \mid y, X) \propto \Pr(\beta, \Sigma) \times L(\beta, \Sigma \mid y, X). \tag{2}$$

For the prior $\Pr(\beta, \Sigma)$, it is convenient to employ independent conjugate distributions, i.e. the normal distribution for β and the inverse Wishart distribution for Σ. The probit likelihood is formed as the product of independent multinomial distributions

$$L(\beta, \Sigma \mid y, X) = \prod_{n,t} \Pr(y_{nt} = \arg\max U_{nt}). \tag{3}$$

Evaluating (3) requires numerically expensive computations of the normal cumulative distribution function due to the error specification in (1). Instead, we augment the latent utilities $(U_{nt})_{n,t}$ as additional parameters (Imai and van Dyk, 2005), following truncated normals with truncation points determined by the observed choices $(y_{nt})_{n,t}$. This yields a straightforward and numerically fast Gibbs sampling scheme to approximate (2). The {`RprobitB`} package (Oelschläger *et al.*, 2022) implements the Gibbs sampler in `R`, including the latent class extension presented in the following.

2 Preference classification

To incorporate preference heterogeneity, we model random variation in the coefficient vector β across deciders using a Gaussian mixture with C classes:

$$\beta_n \sim \sum_{c=1}^{C} s_c N(b_c, \Omega_c), \tag{4}$$

where the weights $(s_c)_c$ a priori are assumed to follow a Dirichlet distribution with concentration parameter $\delta > 0$. This approach has two interpretations. First, it provides an arbitrarily good approximation of the true underlying mixing distribution (Oelschläger and Bauer, 2021). Second, it enables the classification of deciders with common expected preferences b_c and preference covariances Ω_c, which is our focus below.

To avoid the need to a priori select the number C of classes included, we impose a Dirichlet process prior $DP(G, \delta)$ on the distribution (4), where (assuming conjugate priors for b and Ω) the base distribution G is formed as the product of a normal and an inverse Wishart distribution. The Dirichlet process directly integrates into the Gibbs sampling scheme by iteratively updating $(b_c)_c$ and $(\Omega_c)_c$ using their posterior predictive distributions (Neal, 2000). The decider-specific assignments $z = (z_n)_n$ to either one of the existing classes $c = 1, \dots, C$ or a newly formed class $c = C + 1$

are updated based on the conditional probabilities

$$\Pr(z_n = c \mid z_{-n}, \delta) = (N - 1 + \delta)^{-1} \cdot \begin{cases} |\{z_{-n} = c\}| & c = 1, \ldots, C, \\ \delta & c = C + 1, \end{cases} \tag{5}$$

where z_{-n} denotes the vector z excluding the n-th element, and N is the total number of deciders.
Although explicit specification of C is no longer required, there is still an implicit specification through the selection of a value for the concentration prior δ. However, in our simulation study (Table 1), we found that the impact of δ on (5) diminishes as N increases, resulting in stable inference of the underlying class number.

TABLE 1. Median C for varying N and δ with standard deviations in brackets. Choice data were simulated based on the estimates reported in Table 2.

	$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$	$\delta = 2$	$\delta = 10$
$N = 100$	1 (0.33)	2 (0.62)	2 (0.68)	3 (0.79)	4 (1.28)
$N = 1000$	3 (0.15)	3 (0.54)	3 (0.50)	4 (0.78)	5 (1.25)
$N = 6174$	3 (0.22)	3 (0.40)	3 (0.55)	3 (0.77)	4 (1.10)

3 Chess players' propensity for risk-taking

The latent class probit model is well-suited for analyzing discrete choice behavior in settings that feature different groups of deciders with heterogeneous preferences. For a demonstration, we apply the model to classify chess players according to their risk-taking propensity, given that we can expect the presence of both risk-affine and risk-averse players. We further compare the model results to those obtained from the basic probit model from Section 1.
Our application is based on data from an online tournament hosted on the platform www.lichess.org (Lichess API, 2023), where $N = 6174$ participants played multiple chess games with a time limit of one minute per game. The time limit is consumend when it is the player's turn to make a move. A player whos time runs out looses the game automatically. Before the start of each round, players were presented with a risky decision: they could trade half of their clock time for the chance to earn one additional tournament point on top of the base score of two points if they won the game. Factors that influenced this decision include the player's rating, whether they had the first-move advantage, remaining tournament time, winning streak (which yielded extra points), whether the player opted for the risky option in the previous round, whether they had lost in the previous round, and the rating difference between the player and their opponent.

TABLE 2. Change in utility for taking the risky option (ceteris paribus). Reported are the means of the marginal posteriors with standard deviations in brackets.

Factor	Latent class probit			Basic probit
Intercept		-2.05 (0.03)		-1.94 (0.01)
Rating		-0.11 (0.01)		-0.08 (0.01)
Having first move		-0.04 (0.02)		-0.02 (0.01)
Minutes remaining		0.04 (0.01)		0.04 (0.01)
On a winning streak		-0.27 (0.03)		-0.21 (0.02)
Took risk last round		1.21 (0.02)		1.82 (0.02)
	Class 1	Class 2	Class 3	
Proportion	54% (0.03)	36% (0.04)	10% (0.03)	
Lost last round	-0.98 (0.09)	0.03 (0.08)	1.10 (0.18)	0.18 (0.01)
Rating difference	0.10 (0.02)	0.98 (0.06)	1.65 (0.22)	0.52 (0.01)

Both models were fitted using 5000 Gibbs iterations with a "burn-in" of 50%, the results are summarized in Table 2. The latent class model with concentration $\delta = 1$ converged to three classes that characterize different types of players:

- Type 1 players are risk-averse, rarely choosing the risky option against lower-rated opponents or after losing in the previous round.
- Type 2 players decide independently of the previous game's outcome.
- Type 3 players take more risks, with a higher likelihood of choosing the risky option after a loss and favoring it against weaker opponents.

Using the relative frequencies of the class allocation z, we can classify each player. For example, the tournament winner is of type 2 with a probability of 78%, while the runner-up is of type 1 with a probability of 94%.

4 Discussion

The proposed latent class probit model reveals heterogeneity in preferences that is not accessible via the basic probit model. We use a Dirichlet process to conduct a latent class analysis without explicit specification of the class number. For illustration, we applied the model in the context of chess and identified three types of players (cf. Figure 1). Considering the frequency of taking the risky option alone is not a sufficient indicator for risk-taking propensity, as indicated by the strong overlap of the class-wise kernel density estimates depicted in Figure 2. Further analysis is required how the results from Table 2 generalize to other chess tournaments, and how the model can be applied in other areas of preference heterogeneity.

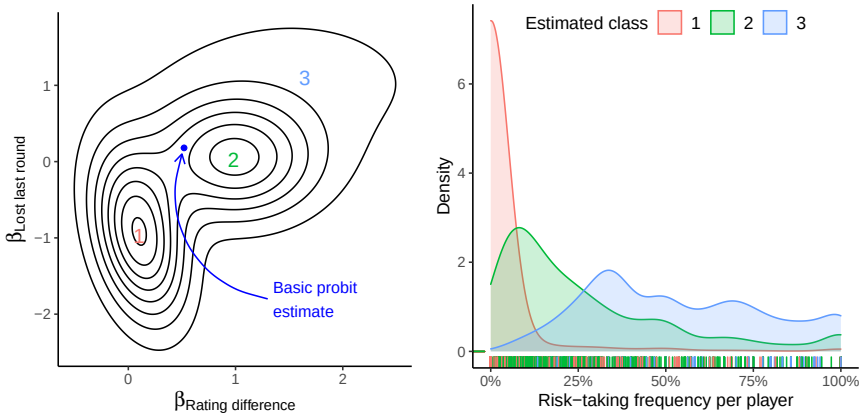


FIGURE 1. Contour plot of estimated mixing distribution with class centers.

FIGURE 2. Kernel density estimates of class-wise risk-taking frequencies.

References

Imai, K. and van Dyk, D. A. (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, **124.2**, 311 – 334.

Lichess API (2023). https://lichess.org/api/tournament/RibHfoX6. Data accessed on 17 February 2023.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, **9(2)**, 249 – 265.

Oelschläger, L. and Bauer, D. (2021). Bayes Estimation of Latent Class Mixed Multinomial Probit Models. *TRB 100th Annual Meeting*.

Oelschläger, L., Bauer, D., Büscher, S., and Batram, M. (2022). *RprobitB: Bayesian Probit Choice Modeling*. R package version 1.1.2, accessible via https://CRAN.R-project.org/package=RprobitB.

Train, K. (2009). *Discrete choice methods with simulation*. Cambridge University Press, **2. edition**.

Kriging wind on pressure levels to enrich the statistical modelling of aircraft trajectories

Rémi Perrichon[1], Xavier Gendre[2], Thierry Klein[1,2]

[1] ENAC - École Nationale de l'Aviation Civile, Université de Toulouse, France.
[2] IMT - Institut de Mathématiques de Toulouse, UMR5219, Université de Toulouse, France.

E-mail for correspondence: remi.perrichon@enac.fr

Abstract: Additional to the usual dimensions of an aircraft trajectory (longitude, latitude, altitude), it is often valuable to consider weather dimensions when studying a flight. Geostatistics provides powerful methods to associate a weather value to a given point of a trajectory. Using kriging equations allows to predict weather values for any point of the flight and to take uncertainties into account. We present the steps to perform kriging of wind speed values on pressure levels with drift and anisotropy. Focus is made on the spatial dimension.

Keywords: Geostatistics; Kriging; Drift; Anisotropy; Trajectory.

1 Motivation and problem statement

For a given flight, the position of an aircraft is recorded for a finite set of observation times. This indexed set of positions is interesting but may be an incomplete summary of the flight. Indeed, knowing the experienced weather at each observation time may help to better understand the dynamics of fuel consumption or noise emission. The goal of this work is to associate each point of a trajectory with a weather value, so that experienced weather during the flight is a piece of information that can be used in further statistical analyses.
Past weather data are not available at any instant in time (if only for storage reasons). Rather, weather data are processed so that a three-dimensional weather grid is available every hour. Because most flights in Europe last more than an hour, the task of matching weather values typically involves several weather grids as schematized in Figure 1.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

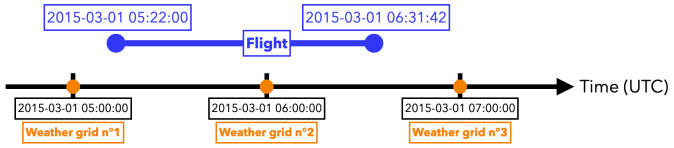


FIGURE 1. Adding weather data to a flight departing from Toulouse-Blagnac (LFBO) and landing at Paris-Orly (LFPO) in March 2015 may involve at least three weather grids.

Formally, this problem is often tackled as an *interpolation* or *spatio-temporal prediction* task. This task is common in environmental sciences as testifies the review of spatial interpolation methods written by Li and Heap (2014).
In this work, a focus is made on the spatial aspect of the problem. In other words, a simple rule is adopted for the time dimension: for each point of a trajectory, the closest weather grid in time is used to perform the spatial interpolation. The interpolation problem boils down to a three-dimensional kriging problem involving an unknown drift and anisotropy. The solution is detailed in the sequel.

2 Raw data, scope of the study

Two data sources are used in the paper.
Trajectory data are taken from the R&D data archive that contains more than 18 million flights as of January 2023. The data are collected by Eurocontrol from all commercial flights operating in and over Europe. Data are available for 4 months each year: March, June, September and December.
Weather data are taken from ERA5 hourly data on pressure levels. ERA5 is the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis for the global climate and weather.
We focus on the interpolation of the three weather grids presented in Figure 1. The weather variable of interest is the horizontal wind speed (expressed in $m.s^{-1}$) for the flight departing from Toulouse-Blagnac (LFBO) and landing at Paris-Orly (LFPO) in March 2015. For 23 pressure levels, horizontal wind speed values are given on a $0.25° \times 0.25°$ longitude-latitude grid. The weather grid on which kriging is done is three-dimensional. For a single weather grid, there are 57 (longitude values) $\times$ 41 (latitude values) $\times$ 23 (pressure values) $= 53,751$ wind values.

3 A geostatistical framework

3.1 Dealing with projection and pressure levels

Raw weather data are given on a three-dimensional grid, often called a *region of interest*, commonly denoted D in geostatistics. Projecting is a safe

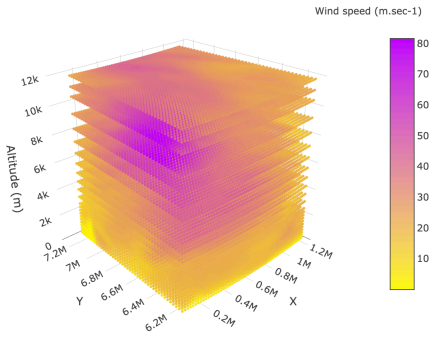


FIGURE 2. Weather grid n°1 giving wind speed values on 2015-03-01 05:00:00.

option when working with spatial data coming in longitude and latitude coordinates. It ensures that all statistical quantities based on the Euclidean distance are accurate. To safely use the Euclidean distance, pressure levels (in hectopascals) must be converted to altitude values in meters.
To go from a pressure level p to an altitude h in meters (m), the following formula is provided by the National Oceanic and Atmospheric Administration (NOAA):

$$h = \frac{145366.45\left[1 - \left(\frac{p}{1013.25}\right)^{0.190284}\right]}{3.281}.$$

It is based on the International Standard Atmosphere (ISA). The resulting grid once the two steps are performed (projection, conversion) is given in Figure 2. The Lambert 93 conformal conic projection is used as it is a very popular option for flights over Metropolitan France.

3.2 Mathematical framework

Every hour, raw data come as a collection of n *regionalized values* denoted $\{z(s_i), i = 1, ..., n\}$. Each location s on D is viewed as the realisation $z(s)$ of a random variable $Z(s)$. Values are said to be regionalized because they exhibit some spatial correlation. The family of real-valued random variables $\{Z(s), s \in D\}$ is traditionally called a *spatial random field*. In the sequel, we assume that the first moment as well as the usual second-order moments of the random field are well-defined. Contrary to usual multivariate statistics, there is only one realization of the random field making inference impossible without some assumptions. Geostatistics often relies on the *second-order stationary* hypothesis. The hypothesis is as follows:

1. The expectation exists and is constant, and therefore does not depend on the location s: $\mu(s) = \mu$.

2. The covariance exists for every pair of random variables, $Z(s)$ and $Z(s+h)$, and only depends on the vector h that joins the locations s and $(s+h)$, but not specifically on them: $C(Z(s), Z(s+h)) = C(h),\ \forall s \in D,\ \forall\ h \in \mathbb{R}^d$ such that $s+h \in D$.

3.3 A drift violates the second-order stationarity assumption

The second-order stationarity assumption doesn't hold for wind data as the mean of the random field depends on location. It is a *drift* problem. This smooth systematic non-random variation should be taken into account. To do so, the random field is broken down into the sum of two components, $Z(s) = \mu(s) + \varepsilon(s)$, where $\mu(s)$ denotes the unknown drift and $\varepsilon(s)$ the stochastic part that can be treated as second-order stationary.
Parametric models to the drift are often fit to detrend the data before attempting the analysis of the spatial correlation structure existing in the residuals. This approach is called *residual kriging* by Montero et al. (2015). This approach has been historically studied by Volpi and Gambolati (1978) through numerical simulations and applied to the mapping of an hydraulic head field of three major aquifers underlying the Venetian lagoon by Gambolati and Volpi (1979). Regarding our application, a quadratic trend has been found to be satisfactory to model the horizontal wind speed drift.
Characterization of the spatial dependence in the residuals relies on the empirical (or experimental) semivariogram. Note that the *variogram* of the random field is defined as the variance of the first differences of the random field:

$$2\gamma(s_i - s_j) = \mathbb{V}(Z(s_i) - Z(s_j)),\ \forall s_i, s_j \in D.$$

The function γ is called the *semivariogram*. In the case of second-order stationarity, the covariance function and the semivariogram are equivalent when it comes to defining the structure of spatial dependence displayed by the phenomenon. One reason for which the semivariogram is preferred to the covariogram is that it does not require the knowledge of the mean of the random field.

3.4 A key aspect: anisotropy

A given empirical semivariogram may not meet the theoretical properties of a valid semivariogram. These theoretical properties are given in most textbooks in geostatistics. The so-called *structural analysis* step is then concerned with the fitting a valid model to the empirical semivariogram. This step is necessary to make valid spatial predictions. Valid models are often *isotropic*. Isotropic covariance functions only depend on the distance between the locations s and $s+h$ as opposed to *anisotropic* ones. Regarding wind data, the dependence between $Z(s)$ and $Z(s+h)$ is obviously a function of both the magnitude and the direction of h. General anisotropy

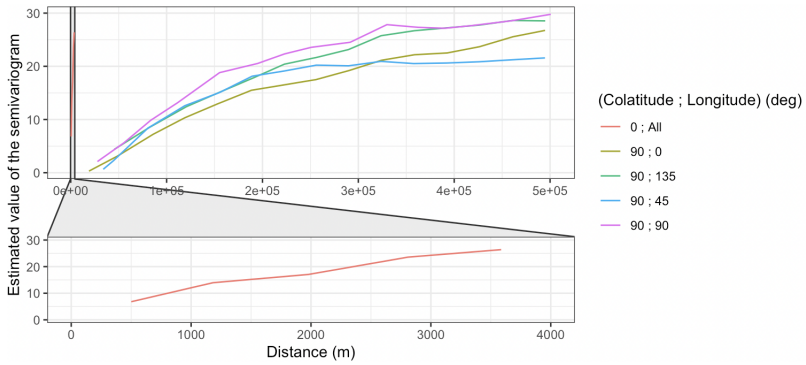


FIGURE 3. Estimated horizontal and vertical semivariograms on the residuals on 2015-03-01 05:00:00. A sample of 10,000 points (drawn at random out of 53,751 locations) is used in the estimation of the semivariograms to improve the computation time. The spatial dependence decreases rapidly in the vertical direction.

models have recently been studied by Allard et al. (2015). In practice, *geometric anisotropy* is the only one that can be corrected using a linear change of coordinates. Indeed, geometric anisotropy is obtained by some stretching of an isotropic model. Speaking in terms of semivariogram, geometric anisotropy is characterized by:

$$\gamma(h) = \gamma_{\text{iso}}(\|Ah\|_2)$$

where the matrix A defines the transformation from the initial space to the isotropic space. A linear transformation of the coordinates in enough to use an isotropic model. As put by Chilès and Delfiner (2012), the matrix A is usually written

$$A = TR_{\theta_3}R_{\theta_2}R_{\theta_1}$$

where $T = \begin{pmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{pmatrix}$ is matrix of scaling factors and R_{θ_3}, R_{θ_2}, R_{θ_2} are rotation matrices (see Chilès and Delfiner (2012), p. 99). Estimating anisotropy parameters is usually done with a *directional semivariogram*. In $\mathbb{R}^3$, taking anisotropy into account is key for good predictions because the vertical spatial dependence usually evolves very differently as compared to the horizontal one.

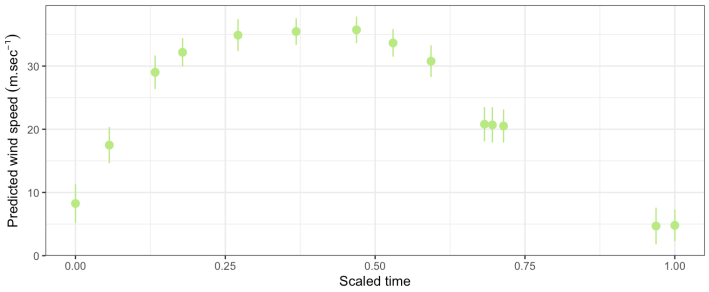


FIGURE 4. Geostatistical predictions of the wind speed values along the raw trajectory.

4 Results

For each grid, the trend is taken into account using Ordinary Least Squares (OLS). The horizontal and vertical semivariograms are then estimated on the residuals. As can be seen in Figure 3, a strong anisotropy should be taken into account, specifically in the vertical direction for which the spatial dependence is rapidly decreasing. Once corrected, predicted values are computed. Predicted wind values for the flight are shown in Figure 4. Note that the 95% confidence intervals only make sense if a Gaussian assumption holds for each weather grid. Confidence intervals are pointwise.

References

Allard, D., Senoussi R., Porcu E. (2015) Anisotropy models for spatial data. *Mathematical Geosciences*, **48** (3), 24 p.

Chilès, J.P. and Delfiner, P. (2012) *Geostatistics, modeling spatial uncertainty*. Wiley Series in Probability and Statistics.

Gambolati, G. and Volpi, G. (1979). Groundwater contour mapping in Venice by stochastic interpolators: 1. Theory. *Water Resources Research*, **15**, 281 – 290.

Li, J. and Heap, A.D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, **53**, 173 – 189.

Montero, J., Fernández-Avilés, G., Mateu, J. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. Wiley Series in Probability and Statistics.

Volpi, G. and Gambolati, G. (1978). On the use of a main trend for the kriging technique in hydrology. *Advances in Water Resources*, **1**, 345 – 349.

Wind speed/direction in complex alpine terrain and snow avalanche accidents in the western part of Austria

Christian Pfeifer[1]

[1] Department of Statistics and Economics, University of Innsbruck, Austria

E-mail for correspondence: christian.pfeifer@uibk.ac.at

Abstract: In this paper we give a proposal how to take the effect of wind on avalanche accidents into account.

Keywords: Wind speed direction; Avalanche accidents.

1 Introduction

About 10 years ago we investigated the effects of weather parameters (precipitation, temperature) on avalanche accidents, see (Pfeifer, Höller 2014). We also tried to take into account the effect of wind (speed/direction). In a recent paper we used spatial reanalysis data of ERA 5 in order to model the effects on basis of municipalities. This turned out to be not feasible because of the complex terrain in the alps. As a substitute we used the data of 3 weather stations which lead to some conclusions (Pfeifer, Höller 2021).

2 Methods

The aim of this paper is to build a spatial wind model; one possibility is to use INCA wind fields (Integrated Nowcasting through Comprehensive Analysis) such as previously proposed (Geosphere Austria, 2022). But taking into account that historical data are only available on basis of weather station data, we try to calculate daily wind field data for our own according to:

- a 3 dimensional wind model assuming mass conserving and incompressible flow

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

- with boundary conditions:
  - digital elevation data model (DEM) on a 100m resolution, 3D mesh model
  - wind velocity/speed equal to zero normal to the bounding surface
  - wind of weather station (determined by latitude and longitude) data in 2 dimensions

resulting in an elliptic partial differential equation. A numerical solution is calculated within the 'finite element model' framework.
As a consequence we are able to calculate the wind load potential (cm/day; see Conlan and Jamieson, 2016) dependent on the velocity v, the direction of wind θ_{wind} and aspect θ_{aspect} of the surface on each point of the terrain:

$$\text{wind load} = \frac{v^3}{125} \cdot \cos|\theta_{\text{wind}} - \theta_{\text{aspect}}|$$

Please note that the wind load potential can be positive/negative depending on the lee/luv side of the terrain. For objective reasons, we restrict the calculation to alpine terrain (sea level $\geq 1800m$).

3 Results and Discussion

In this paper we calculate a wind field based on the weather station 'Galzig' around the municipality 'St. Anton am Arlberg'. Looking at daily backcountry avalanche accidents we observe 54 accidents within the winter periods 1993/94 – 2011/2012 in this region . For illustration purposes Figure 1 shows the case of west wind ($\sim$ 10 m/s), which turned out to be the predominant wind direction of the weather station 'Galzig' (red point). We, however, had a focus on wind load larger then zero – ranged from black (low) to white (high) in Figure 1.
If we look at the daily number of avalanche accidents of the municipality 'St. Anton am Arlberg' the boxplots of Figure 2 show a positive effect of wind load on the number of accidents in this region. Calculating a corresponding Poisson model we observe a significant result (effect=0.05856, p=0.001).
Finally, maps just as Figure 1 are useful in the snow avalanche science community in order to recognize areas of danger in alpine terrain. However, it is necessary to compute maps of this kind in a more comprehensive way.

References

Conlan M. and Jamieson B. (2016). Naturally triggered persistent deep slab avalanches in western Canada Part I: avalanche characteristics and weather trends from weather stations. *Journal of Glaciology* 62:243 – 255.

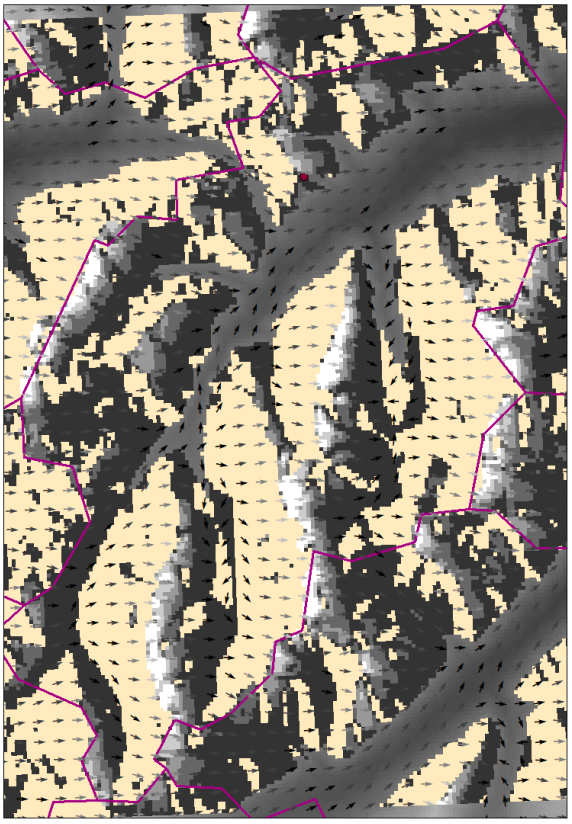


FIGURE 1. Wind and wind load Galzig/St. Anton a. Arlberg ranged from black (low) to white (high)

Geosphere Austria (2022). Integrated Nowcasting through Comprehensive Analysis (INCA). https://www.zamg.ac.at/cms/de/forschung/wetter/inca

Pfeifer C. and Höller P. (2014). Effects of precipitation and temperature in alpine areas on backcountry avalanche accidents reported in the western part of Austria within 1987 – 2009. In: *Proceedings IWSM 2014 Göttingen.*

Pfeifer C. and Höller P. (2021). Effects of Precipitation, Temperature and Wind in Alpine Areas on Backcountry Avalanche Accidents Reported in the Western Part of Austria within 1987 – 2009. Submitted to: *Natural Hazards* 2021.

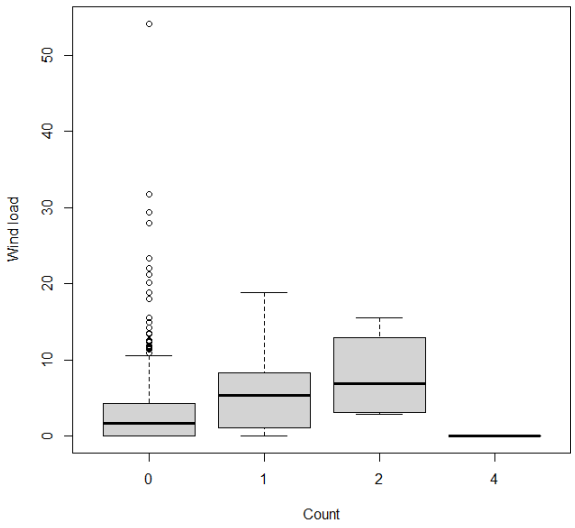


FIGURE 2. Number of avalanche accidents ('St. Anton am Arlberg') dependent on wind load > 0

Wastewater analysis in the light of Covid-19: A GAMLSS approach

Roman Pfeiler[1], Helga Wagner[1], Hans Peter Stüger[2], Karin Weyermair[2], Sabrina Kuchling[2], Patrick Hyden[2]

[1] Department of Applied Statistics, Johannes Kepler University Linz, Austria
[2] Austrian Agency for Health & Food Safety (AGES), Austria

E-mail for correspondence: roman.pfeiler@jku.at

Abstract: GAMLSS models are used to analyse longitudinal data obtained from wastewater treatment plants in Austria. The variable of interest is the amount of Covid-19 related excretion in the wastewater. The goal of the project is to model this wastewater signal as a function of covariates, specifically the vaccination rate and virus variants. Simple Gamma mixed-effects regression models are compared to more flexible GAMLSS models using the Box-Cox t distribution.

Keywords: GAMLSS; Box-Cox t distribution; Wastewater Analysis; Covid-19

1 Introduction

The outbreak of Covid-19, which was declared a pandemic by the WHO in March 2020, greatly impacted and changed the lives of many people in various ways. From a political perspective it is important to monitor and understand the virus and its development to ensure that the current pandemic management remains effective. In Austria, the number of positive Covid-19 tests (i.e. the human signal) served as the most important disease monitoring indicator up until recently. However due to the decreasing number of conducted tests, an alternative monitoring of wastewater with respect to virus copies related to Covid-19 (i.e. the wastewater signal) is considered now. The goal of this paper is to analyse data obtained from wastewater measurements to identify time trends and factors that influence the amount of virus copies found in the wastewater. To achieve flexible modelling of the wastewater signal Generalized Additive Models for Location, Scale and Shape (GAMLSS) proposed by Rigby and Stasinopoulos (2005) are employed.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Data

The data set consists of $N = 5180$ observations taken at $n = 32$ Austrian wastewater treatment plants from 2020-09-28 to 2022-10-10. The variable of interest is the wastewater signal with values on $\mathbb{R}^+$ and higher values corresponding to a higher virus concentration in the wastewater. Figure 1 shows the wastewater signal in four exemplary plants over time. There is large heterogeneity across plants with respect to time points at which measurements are taken, e.g. in the first plant measurements start much later and in the second plant there are large gaps between measurements.

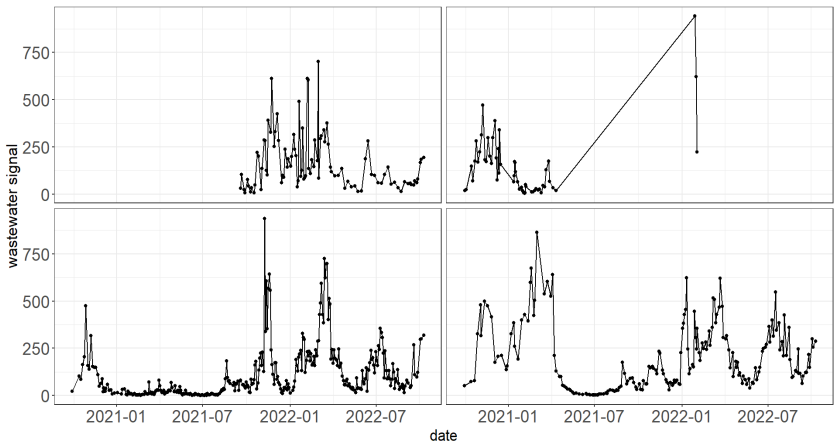


FIGURE 1. Wastewater signal for selected plants (points are measurements)

Of particular interest in analysing the development of the wastewater signal over time are the effects of vaccination as well as the dominant virus variant, since virus load is assumed to differ across virus variants. Data are available on the percentage of persons who got one, two or three vaccinations. To achieve dimension reduction we computed a vaccination score as the first principle component of the three vaccination rates, which resulted as

$$\text{vacc.score} = 0.6490 \cdot \text{vacc}_1 + 0.6369 \cdot \text{vacc}_2 + 0.4162 \cdot \text{vacc}_3 \qquad (1)$$

The principle component score of the third vaccination rate is smaller than those of the first and second which are almost identical and the explained variance is over 90 %. Figure 2 shows time series for the three vaccination rates as well as the composite score.

The dominant virus variant was categorised with categories *Alpha-Beta*, *Delta*, *Omicron-1-2*, *Omicron-4-5* and *rest/unknown*. The categorisation of *Omicron* in subcategories *Omicron-1-2* and *Omicron-4-5* is based on their genetic distance which was considered large enough to justify separate effects.

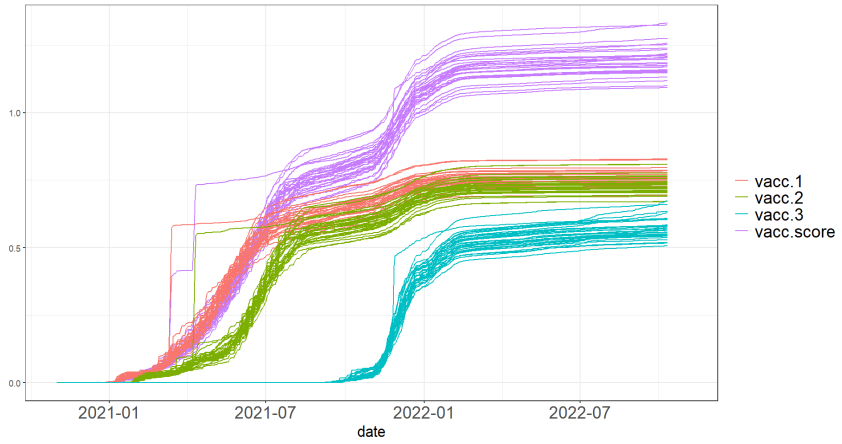


FIGURE 2. Time series of vaccination rates and the vaccination score for 32 wastewater plants.

Additional information on the laboratory that analysed the wastewater as well as the federal state are available and used as control variables.

3 GAMLSS Models for the Wastewater Signal

In a GAMLSS model the response variable y_i, $i = 1, \ldots, n$ is modelled conditional on a vector of covariates $\mathbf{x}_i$ as a realization of a distribution $\mathcal{D}$ that is not restricted to the exponential family,

$$\mathrm{y}_i|\mathbf{x}_i \sim \mathcal{D}(\boldsymbol{\theta}_i). \tag{2}$$

For the distributional parameters $\boldsymbol{\theta}_i = (\theta_{1,i}, \ldots, \theta_{k,i})'$ an additive model is specified as

$$g_k(\theta_{k,i}) = \eta_{\theta_{k,i}} = f_{\theta_k,1}(\mathbf{x}_i, \beta_{\theta_k,1}) + \cdots + f_{\theta_k,J}(\mathbf{x}_i, \beta_{\theta_k,J}), \tag{3}$$

where $g_k(\cdot)$ is an invertible, twice-differentiable link-function and the functions $f_{\theta_k,j}(\cdot)$, $j = 1, \ldots, J$ can be either parametric or nonparametric smooth functions, e.g. modelled with splines.
To model the wastewater signal y_{it} in plant $i = 1, \ldots, n$ at time point $t = 1, \ldots, T_i$ we considered several GAMLSS mixed models.
The first two models are Gamma regression models with distributional parameters $\boldsymbol{\theta}_{it} = (\mu_{it}, \sigma_{it})'$, where μ_{it} is the mean and σ_{it} is related to the standard deviation. In the first model homoscedasticity is assumed, whereas in the second also σ_{it} is modelled in terms of covariates:

$$\begin{aligned} y_{it} &\sim \mathcal{G}(\mu_{it}, \sigma_{it}), \\ \log(\mu_{it}) = \eta_{\mu_{it}}, &\quad \log(\sigma_{it}) = \text{constant}, \qquad (\texttt{Gamma-1}) \\ \log(\mu_{it}) = \eta_{\mu_{it}}, &\quad \log(\sigma_{it}) = \eta_{\sigma_{it}}. \qquad (\texttt{Gamma-2}) \end{aligned}$$

The remaining two models use the Box-Cox t distribution, which has four distributional parameters $\boldsymbol{\theta}_{it} = (\mu_{it}, \sigma_{it}, \nu_{it}, \tau_{it})'$, where μ_{it} is the median, σ_{it} is a centile-based coefficient of variation, ν_{it} relates to the skewness and τ_{it} to the kurtosis of the distribution (see Rigby & Stasinopoulos (2006) for further details on the BCT distribution):

$$y_{it} \sim \text{BCT}(\mu_{it}, \sigma_{it}, \nu_{it}, \tau_{it}),$$
$$\log(\mu_{it}) = \eta_{\mu_{it}}, \quad \log(\sigma_{it}) = \eta_{\sigma_{it}}, \quad \nu_{it} = \eta_{\nu_{it}}, \quad \log(\tau_{it}) = \eta_{\tau_{it}}.$$

The first BCT model contains a random intercept only for the location parameter, whereas the second also models the scale parameter with a random intercept. We tried also models with random intercepts in all distributional parameters but encountered convergence problems and hence did not consider these further.

Starting from a model where all covariates were included in all predictors, model selection based on AIC was performed. Table 1 gives an overview of the selected covariates in each of the linear predictors.

TABLE 1. Final models with selection based on AIC

Model	vacc.score	dom.var.	controls	γ_i	AIC	$R^2_{\text{Cox-S.}}$
`Gamma-1`	μ	μ	μ	μ	55621.98	0.7109
`Gamma-2`	μ	μ, σ	μ, σ	μ	54615.26	0.7653
`BCT-1`	μ	μ, σ, ν	μ, σ, ν, τ	μ	54387.79	0.7800
`BCT-2`	μ	μ, σ	μ, σ, ν, τ	μ, σ	54330.10	0.7833

AIC is considerably larger for the simpler Gamma than the BCT models. Moreover, for both Gamma models the wormplots in Figure 3 indicate the need for kurtosis modelling, whereas the BCT models provide an adequate fit. Based on the AIC the best model is `BCT-2`, implying that there is substantial variability concerning the scale parameter with respect to the wastewater treatment plants.

Figure 4 shows effect plots for the `BCT-2` model. The effect of the composite vaccination score on the wastewater signal is nonlinear and decreases over most of its range. Based on the estimated functional relationship, it can be stated that a higher vaccination rate corresponds to a lower signal in the wastewater.

Also the dominant virus variant has a significant effect on the median of the wastewater signal. It is lower when *Alpha-Beta* or *Delta* are the dominant virus variant than for both the *Omicron-1-2* as well as the *Omicron-4-5* subvariant. The dominant virus variant also has a significant effect on the scale parameter, which is higher for the variant group *rest/unknown* than all the others. This is likely due to the fact that this category contains several virus variants.

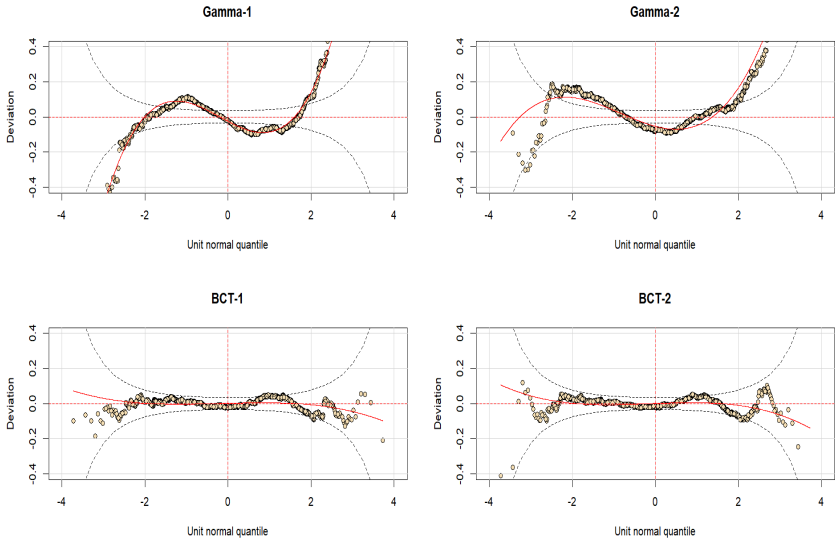


FIGURE 3. Wormplots of different GAMLSS models

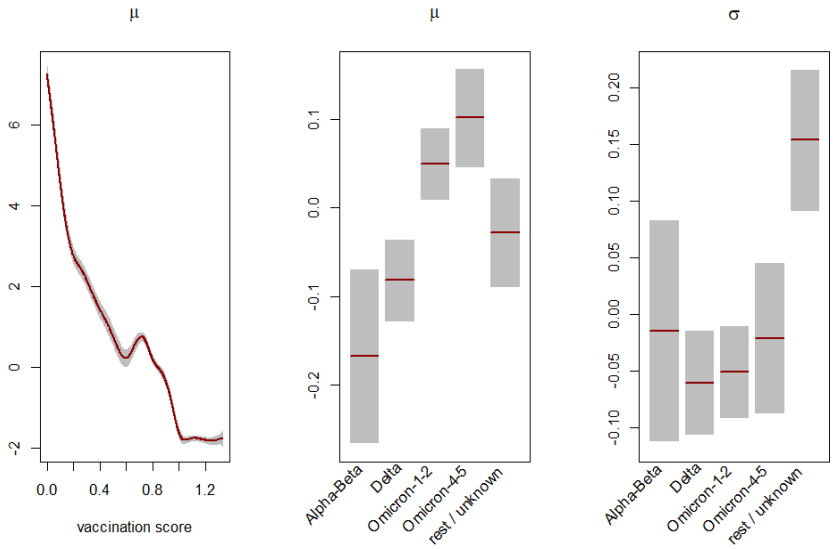


FIGURE 4. `BCT-2`: Effect plots of vaccination rate and dominant variant

Figure 5 displays the distribution of the random intercept for both distributional parameters of the best model (`BCT-2`) indicating that there are pronounced differences across wastewater treatment plants with respect to the median as well as the coefficient of variation.

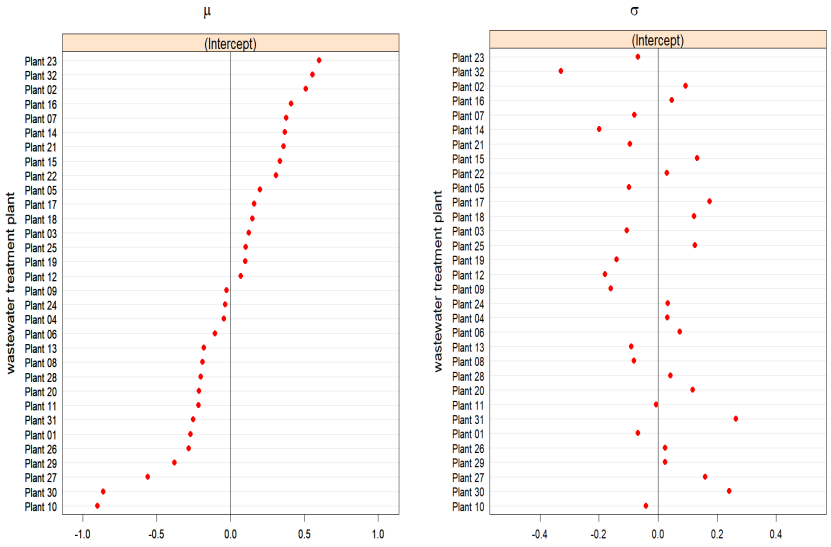


FIGURE 5. `BCT-2`: Random Intercepts for μ and σ. Wastewater treatment plants are ordered with respect to the random intercept for the median.

4 Conclusion

The more complex BCT models clearly outperformed the simpler Gamma mixed-effects regression models. The vaccination rate and the virus variant are both relevant factors for modelling the wastewater signal.

References

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

Rigby, R.A. and Stasinopoulos, D.M. (2006). Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. In: *Statistical Modelling*, **6**, 209 – 229.

Evaluating academic performance using nonparametric regression

Hildete P. Pinheiro[1], Fernando H.S. Barreto[1]

[1] State University of Campinas, Brazil

E-mail for correspondence: hildete@unicamp.br

Abstract: In this work we will use nonparametric regression estimators, such as modified Theil-Sen estimators (with high number of ties), to evaluate academic performance of students verifying the relationship between the Entrance Exam Score (EES) and the grades in courses taken at the university. The dataset is from the State University of Campinas, in Brazil, and we will focus on engineering major students and the perfomance in Calculus I according to the EES in Math and Physics and type of High School (Private or Public).

Keywords: modified Theil-Sen estimator; robust regression; U-statistics.

1 Introduction

Nonparametric methods to evaluate students' performance have been proposed by some authors. Maia et al. (2016) used test statistics based on quasi U-statistics looking at the entrance exam and the grade point average (GPA) to verify differences in academic performance according to sex and type of High School; Pinheiro et al. (2020) treated the problem using multivariate analysis and considered the grades of each of the courses taken by the students using tests based on quasi U-statistics.
Here, we will use a modified version of Theil-Sen regression estimators (Wilcox, 2021) to evaluate the relationship between the Entrance Exam Score (EES) and the grades in each course. We will focus on data from engineering major students and the performance in Calculus I (MA111) according to the EES in Math (EES_{Math}) and Physics (EES_{Phys}) and type of HS (Pr-Private x Pu-Public). We used as covariate the average of the EES in Math and Physics subtracted by its median, i.e., $X_i = (ESS_{Math} + ESS_{Phys})/2$ and $(X_i - median(X))$.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

According to Wilcox (2021), a conventional linear regression analysis with this type of data may result in some problems. For instance, low power of the tests and the assumption of homocedasticy may not hold, since we have many outliers. Therefore, more robust estimators, such as L1 regression would be a more suitable method. However, we have another problem on this dataset, which is a high number of ties. Wilcox et al.(2013) point out that in the presence of too many ties, some robust estimators may have a poor performance in terms of the Type I error, even with big sample sizes.

2 Methods

Let $\mathbf{Z}_{i(ak)} = (Z_{i(ak1)}, \ldots, Z_{i(akL_i)})^\top$ be the vector of grades of the i-th student of class/major k, who entered at year a. Let $j = 1, \ldots, L_i$ be the index indicating the course taken by student i, with L_i being the total number of courses taken by the student. Note that even though the components of $Z_{i(ak)}$ are theoreticaly continuous random variables, but in practice, we observe discrete random variables. Let $Y_{i(akj)}$ be the discrete grades of student i. For instance, $Y_{i(akj)} \in \{0.0, 0.1, 0.2 \ldots, 10.0\}$, that is,

$$\begin{cases} Y_{i(akj)} = 0.0 \text{ if } Z_{i(akj)} \in [0.0, 0.05); \\ Y_{i(akj)} = 0.1 \text{ if } Z_{i(akj)} \in [0.05, 0.15); \\ \vdots \\ Y_{i(akj)} = 9.9 \text{ if } Z_{i(akj)} \in [9.85, 9.95); \\ Y_{i(akj)} = 10.0 \text{ if } Z_{i(akj)} \in [9.95, 10.0]. \end{cases}$$

Now, define the following models:

(I) $Y_{i(akj)} = \beta_{(akj)0} + \beta_{(akj)1} X_{i(akj)} + \epsilon_{i(akj)},\ i = 1, 2, \ldots, n_{(akj)}$;

(II) $Y_{i(askj)} = \beta_{(askj)0} + \beta_{(askj)1} X_{i(askj)} + \epsilon_{i(askj)},\ i = 1, 2, \ldots, n_{(askj)}$;

(III) $Y_{i(aj)} = \beta_{(aj)0} + \beta_{(aj)1} X_{i(aj)} + \epsilon_{i(aj)},\ i = 1, 2, \ldots, n_{(aj)}$;

(IV) $Y_{i(asj)} = \beta_{(asj)0} + \beta_{(asj)1} X_{i(asj)} + \epsilon_{i(asj)},\ i = 1, 2, \ldots, n_{(asj)}$;

For student i, $a \in \{2009, 2010, \ldots, 2015\}$ is the year of entrance; $k \in \{1, 2, \ldots, 6\}$ the majors: Agriculture Engineering (1), Civil Engineering (2), Food Engineering (3), Computing Engineering (4), Electric Engineering (5) e Mechanical Engineering (6); $j \in \{1, 2\}$ the courses Calculus I (1) and Linear Algebra (2); $s \in \{1, 2\}$ is the type High School (HS): Private (1-Pr) or Public (2-Pu).
For models II and IV, the predicted variable $X_{i(\cdot)}$ may be separated in two groups according to type of HS (1-Pr and 2-Pu). The parameters $\beta_{(\cdot)0}$ and $\beta_{(\cdot)1}$, are the intercept and slope for group $(\cdot)$, respectively. $\epsilon_{i(\cdot)}$ is stochastic with distribution free.

If there is no difference between EES effects among majors, we can ignore the majors and use model III. Model IV is used to test whether or not there is difference between type of HS, ignoring the majors, i.e, $H_0 : \beta_{(a1j)\ell} = \beta_{(a2j)\ell}$, $\ell = 0, 1$.

Table 1 shows the number of ties, n_{ties}, and the number of outliers, n_{out}, for the grades in Calculus I (MA111) in years 2009 to 2012 to illustrate some of the problems of the dataset. For years 2013 to 2015, the pattern remains the same. The average of relative frequency of ties ($f_{ties} = \frac{n_{ties}}{n}$) for Calculus I is 54.96%. To detect outliers we used the method described in Carling (2000).

TABLE 1. Number of ties and outliers on the grades of Calculus I

	2009			2010			2011			2012		
	n_{out}	n_{ties}	n	n_{out}	n_{ties}	n	n_{out}	n_{ties}	n	n_{out}	n_{ties}	n
Agriculture Eng.	0	28	62	0	28	60	0	27	62	0	24	55
Civil Eng.	2	43	76	1	46	77	6	39	78	6	41	78
Food Eng.	4	66	114	9	54	106	5	54	99	0	52	98
Computer Eng.	5	86	132	3	88	137	2	90	133	7	93	134
Eletric Eng.	5	57	91	7	45	93	1	53	92	3	52	101
Mechanical Eng.	7	42	87	5	55	85	3	43	84	3	41	81

Since the distribution of the grades ($Y_{i(\cdot)}$) has severe number of ties, we will use a more robust estimator (Theil, 1950 and Sen, 1968) with its modified version (Harrell and Davis, 1982). Frequently known by **Theil-Sen estimator** of the slope (β_{TS}), it is the median based on all slopes associated with any two distinct ponts in the sample (Theil, 1950; Sen, 1968; Wilcox, 2021). Comparing it with the least square estimator, we could say that the estimated line obtained by β_{LS} is such that $\rho(X_i, \epsilon_i) = 0$, where ρ is the Pearson correlation coefficient, while the line obtained by Theil-Sen is such that $\tau(X_i, \epsilon_i) = 0$, where τ is the Kendall's tau coefficient (Kendall, 1938). A modified version of the Theil-Sen estimator is the **Harrell-Davis estimator** (Harrell and Davis, 1982).

In fact, the estimate of the slope is $\beta_{1(TS)} = \hat{\theta}_{0.5}$. The intercept is estimated by the Harrell–Davis estimate of the median based on $Y_1 - \hat{\beta}_1 X_1, \ldots, Y_n - \hat{\beta}_n X_n$. This modification will be called the Thiel-Sen-Harrell-Davis (TSHD) estimator.

Note that the choice of the intercept does not affect the Kendall's tau coefficient (τ), but it makes the median of the residuals to be approximately zero because τ is a U-statistic symmetric around zero.

To avoid problems caused by severe number of ties in the response variable, the modified version in the estimation procedure of Theil-Sen estimator, TSHD estimator, will be used here. The median of the slopes of any two points in the sample will be replaced by their respective order statistics (Wilcox et al., 2013).

To test whether the effects of EES are different across the different majors, we tested $H_0 : \beta_{(ak1)\ell} = \beta_{(ak'1)\ell}$, $\forall k \neq k'$ e $\ell = 0, 1$ using a Bootstrap procedure with $B = 1000$ as the number of bootstrap samples. For details see Wilcox (2021). Since there was no difference on the effect of EES among the majors, we adjusted models like Model III.

Figures below shows the point estimates and Confidence Intervals (CI) for model I for students who took Calculus I from 2009 to 2015. Looking at Figure 1, one can see that students with major in Agriculture Engineering have median performance in Calculus I worst than all the other Engineering major students and the results were all significant by a 0.05 level.

Adjusting model IV, we found that the EES were significant for Pr HS and Pu HS. Using model IV, we also tested the hypothesis of differences in the intercepts and slopes according to type of HS (1-Pr x 2-Pu), i.e., $H_0 : \beta_{(as1)\ell} = \beta_{(as'1)\ell}$, with $\ell = 0, 1$. There are significant differences on the intercepts (for years 2009, 2010, 2013 and 2014) in the performance of Calculus I according to type of HS (Pr x Pu), with the median of the predicted grade in Calculus I greater for Pr HS than for Pu HS.

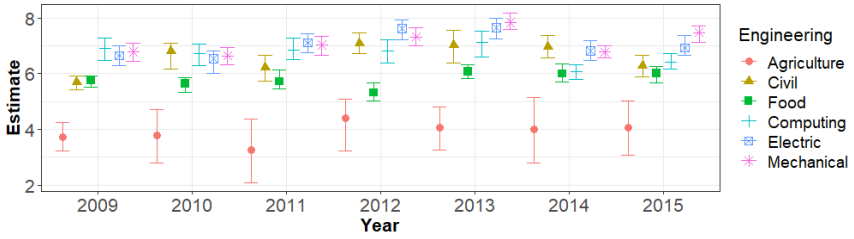


FIGURE 1. Point estimates and CI for model III: Intercepts for Calculus I

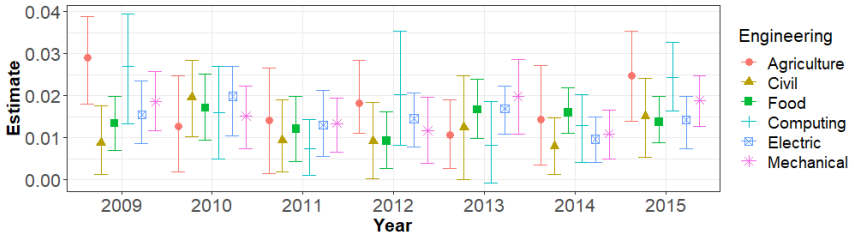


FIGURE 2. Point estimates and CI for model III: Slopes for Calculus I

Acknowledgments: Barreto, F. thanks CAPES-Coordenation for the Improvement of Higher Education Personnel and Pinheiro, H. thanks CNPq-National Scientific and Technological Development (310874/2018-1).

References

Carling, K. (2000). Resistant outlier rules and the non-Gaussian case. *Computational Statistics & Data Analysis* **33**, 249-258.

Harrell, F. E. and Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, **69**, 635-640.

Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika*, **30 (1-2)**, 81-93.

Maia, R. P.; Pinheiro, H. P.; Pinheiro, A. (2016) Academic performance of students from entrance to graduation via quasi U-statistics: a study at a Brazilian research university. *Journal of Applied Statistics*, **43**, 72-86.

Pinheiro, H. P.; Sen, P.K.; Pinheiro, A.; Kiihl, S. F. (2020) A nonparametric approach to assess undergraduate performance. *Statistica Neerlandica*, **74**(4), 538-558.

Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, **63**, 1379-1389.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae*, **12**, 85-91.

Wilcox, R. R. (2021). *Introduction to Robust Estimation and Hypothesis Testing*, 5th Edition. San Diego, CA: Academic Press.

Wilcox, R. R., Erceg-Hurn, D., Clark, and F. Carlson, M. (2013). Comparing two independent groups via the lower and upper quantiles. *Journal of Statistical Computation and Simulation.*

Bayesian effect selection in structured piecewise additive joint models using the NBPSS prior

Anja Rappl[1], Elisabeth Bergherr[2]

[1] Friedrich-Alexander Universität Erlangen-Nürnberg, Germany
[2] Georg-August-Universität Göttingen, Germany

E-mail for correspondence: anja.rappl@fau.de

Abstract: Joint Models for longitudinal and time-to-event data are an established modelling tool, yet variable selection tools for this model type are still scarce. Therefore, we apply the Normal Beta Prime Spike and Slab prior for effect selection to a Bayesian Structured Piecewise Additive Joint Model in a simulation study. The resulting effect selection is satisfactory and might prove a versatile tool for analysts.

Keywords: Bayesian statistics, effect selection, joint models, NBPSS

1 Introduction

Joint models for longitudinal and time-to-event data are an established tool to analyse data simultaneously capturing information on a longitudinal outcome and an event outcome., At the same time choosing the correct variables in their correct effect is crucial to avoid bias.The statistical toolbox knows various methods from frequentist and Bayesian statistics as well as statistical machine learning. Yet their extension to such a complex model type as the joint model is not straightforward., Thus the few methods existing employ versions of a penalized likelihood approach such as the LASSO or broken adaptive ridge and are tailored to specific data problems. A notable exception is variable selection through gradient boosting (Griesbach et al., 2023)., What these methods share is the relative simplicity of employing only linear models. On top, coefficients from prediction and allocation optimized statistical boosting techniques lack variances. Therefore, we suggest the use of a Bayesian approach able to perform not just variable

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

but also effect selection and the capacity to also be applied to structured additive models: The Normal Beta Prime Spike and Slab (NBPSS) prior (Klein et al., 2021)., The next section details the methodology behind joint models and the NBPSS prior. Section 3 presents results of a first simulation study to identify the prior's performance and the last section gives a short summary and outlook.

2 Methodological Background

A joint model traditionally consists of a potentially structured additive mixed model for the longitudinal outcome and a proportional hazards (PH) model for the time-to-event outcome. Since the equivalent counting-process notation of the proportional hazards model as a Poisson model has proven superior in performance in Bayesian joint models, we will consider the time-to-event part as being piecewise additive (Rappl et al., 2023). The formulation then takes the form

$$\begin{aligned} \boldsymbol{y}(t) &= \boldsymbol{\eta}_{\text{l}}(t) + \boldsymbol{\eta}_{\text{ls}}(t) + \varepsilon, \quad \varepsilon \sim \text{N}(0, \sigma_\varepsilon^2 \boldsymbol{I}) \\ \boldsymbol{\lambda}(t) &= \exp\left\{ f_0(t_j) + \boldsymbol{\eta}_{\text{s}} + \alpha \boldsymbol{\eta}_{\text{ls}(t)} \right\}, \quad \forall\, t \in (\kappa_{j-1}, \kappa_j], \end{aligned} \tag{1}$$

where $\boldsymbol{\eta}_{\cdot}$ are longitudinal (l), survival (s) or shared (ls) predictors respectively. The latter also connect the two model parts via the association parameter α and must include random effects. Otherwise the predictors may be specified to include linear, smooth and/or spatial effects represented in matrix notation $\boldsymbol{Z}_k \boldsymbol{\gamma}_k$ with $\boldsymbol{Z}$ being the data design matrix and $\boldsymbol{\gamma}_k$ the vector of corresponding coefficients. $f_0(t_j)$ represents the baseline hazard modelled over $j = 1, \ldots, J$ intervals of time t with boundaries $(\kappa_{j-1}, \kappa_j]$.
Effect selection then is performed via the NBPSS prior. This relies on the reformulation of an effect $\boldsymbol{Z}\boldsymbol{\gamma} = \varpi \boldsymbol{Z} \tilde{\boldsymbol{\gamma}}$, where ϖ is an importance parameter and $\tilde{\boldsymbol{\gamma}}$ are standardized coefficients.
The standardized coefficients follow the generic prior

$$p\left(\tilde{\boldsymbol{\gamma}}_k\right) \propto \exp\left\{ -\frac{1}{2} \tilde{\boldsymbol{\gamma}}_k' \boldsymbol{K}_k \tilde{\boldsymbol{\gamma}}_k \right\} \mathbb{I}_{[\boldsymbol{A}_k \tilde{\boldsymbol{\gamma}}_k = \boldsymbol{0}]}, \tag{2}$$

which through the definition of $\boldsymbol{K}_k$ may represent any effect. The constraint $\boldsymbol{A}_k$ is chosen such that it represents the basis of the null space of $\boldsymbol{K}_k$, i.e. $\boldsymbol{A}_k = \text{span}(\text{ker}(\boldsymbol{K}_k))$, in order for the prior to be identifiable and proper. This enables the NBPSS prior to discriminate between penalized (smooth) and un-penalized (linear) effect components. Selection itself happens through the importance parameter, which is subject to the NBPSS

prior of the form of the following Gamma-distribution:

$$\varpi_k^2 \mid \delta_k, \psi_k^2 \sim \text{Ga}\left(\frac{1}{2}, \frac{1}{2r_k(\delta_k)\psi_k^2}\right), \quad \psi_k^2 \sim \text{InvGa}(a_k, b_k),$$

$$\delta_k \mid \pi_k \sim \text{Bern}(\pi_k), \quad \pi_k \sim \text{Beta}(a_{0,k}, b_{0,k}), \quad r_k(\delta_k) = \begin{cases} r_k > 0 & \delta_k = 0 \\ 1 & \delta_k = 1. \end{cases}$$

The model variance σ_ε^2 is a priori inverse-Gamma-distributed with $\text{InvGa}(a_\sigma, b_\sigma)$. The likelihoods follow a Normal and a Poisson distribution respectively:

$$\boldsymbol{y} \sim \text{N}(\boldsymbol{\eta}_\text{l}(t) + \boldsymbol{\eta}_\text{ls}(t), \sigma_\varepsilon^2 \boldsymbol{I}), \quad \boldsymbol{\delta}_j \sim \text{Poi}(\log \boldsymbol{\lambda}(t)) \quad \forall\, t \in (\kappa_{j-1}, \kappa_j].$$

Both methods are implemented in BayesX.

3 Simulation study and results

From the model in (1) longitudinal measurements $\boldsymbol{y}(t)$ for $n = 200$ individuals are generated over originally $n_i = 6$ time points in the range of $t \in (0, 1)$ and an association of $\alpha = -0.3$ with the predictors

$$\begin{aligned} \boldsymbol{\eta}_\text{l} &= 0.5\ \boldsymbol{x}_{\text{l}1} + f_1(\boldsymbol{x}_{\text{l}2}), \\ \boldsymbol{\eta}_\text{ls} &= 0.9\ \boldsymbol{x}_{\text{ls}1} - 0.5\ f_2(\boldsymbol{x}_{\text{ls}2}) - 0.5\ \boldsymbol{x}_{\text{ls}3}(t) + 0.4\ \boldsymbol{t} + \boldsymbol{b}_0 + \boldsymbol{b}_1\ \boldsymbol{t} \quad \text{and} \\ \boldsymbol{\eta}_\text{s} &= 0.1\ \boldsymbol{x}_{\text{s}1} + 0.5\ f_2(\boldsymbol{x}_{\text{s}2}). \end{aligned}$$

The non-linear functions used are $f_1(x) = 0.5\ x{+}15\ \phi(2(x{-}0.2)){-}\phi(x{+}0.4)$ and $f_2(x) = \sin(x)$. All covariates are simulated from a Uniform-$U(-1, 1)$ distribution with the covariates $\boldsymbol{x}_{\text{s}\cdot}$ of the survival predictor and $\boldsymbol{x}_{\text{ls}\cdot}$ of the shared predictor being time-constant and covariates $\boldsymbol{x}_{\text{l}\cdot}$ of the longitudinal predictor and $\boldsymbol{x}_{\text{ls}3}$ of the shared predictor being time-dependent. Further the model variance is set to $\sigma_\varepsilon^2 = 0.5$ and the variances of the random intercepts and slopes are set to $\tau_1^2 = \tau_2^2 = 2$. Survival times are generated with a Weibull baseline hazard of form $\lambda_0(t) = pqt^{q-1}$ with scale $p = 0.4$ and shape $q = 1.5$ and censoring applied at $T_i = \min(T_i^*, 1)$ with an additional artificial random sampling censoring for 50% of the censored individuals in range $T_i \in (0, 1)$., In addition three non-informative covariates per predictor are sampled from $U(-1, 1)$ for variable selection. The study is carried out for $R = 100$ replications and the results are evaluated as posterior selection probability for each effect component.

Figure 1 illustrates the posterior selection probability of each effect component. For the effects in the longitudinal predictor η_l selection is perfect with correct identification of informative and non-informative components. The selection of the survival predictor η_l and the shared predictor η_ls is not yet ideal. In the former case all non-informative components were correctly de-selected, but so were all others except for the linear component of

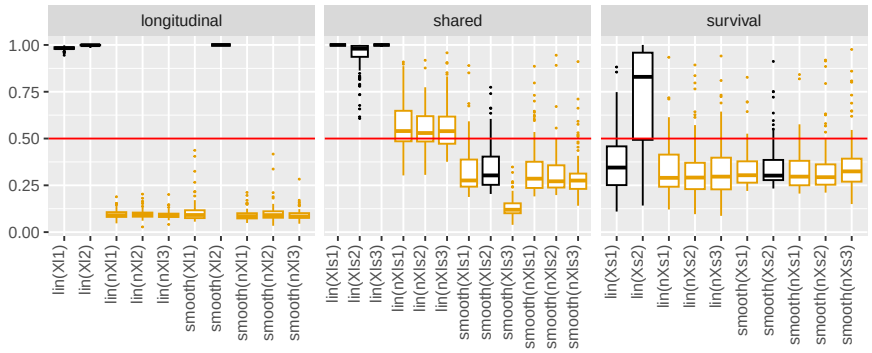


FIGURE 1. Posterior selection probabilities of effect components when using the NBPSS prior in a joint model. Black boxes belong to informative effects, yellow boxes to non-informative effects.

the informative smooth effect. In the shared predictor linear components of the informative variables were accurately selected, but the informative smooth component was missed and instead the non-informative variables were selected.

4 Results and outlook

Due to limited availability of variable selection mechanisms in joint models, we have used the NBPSS prior as an effect selection mechanism in Bayesian Structured Piecewise Additive Joint Models. A first naive application has already yielded imperfect, but satisfactory results.
What remains to be tested now is what fine tuning is needed for the prior to perform better in this setting. This constitutes a low-threshold option. Another option would be adapting the functionality of the NBPSS prior to better suit joint models.

References

Griesbach, C., Mayr, A., & Bergherr, E. (2023). Variable selection and allocation in joint models via gradient boosting techniques. *Mathematics*, *11*(2). https://doi.org/10.3390/math11020411

Klein, N., Carlan, M., Kneib, T., Lang, S., & Wagner, H. (2021). Bayesian Effect Selection in Structured Additive Distributional Regression Models. *Bayesian Analysis*, *16*(2), 545–573. https://doi.org/10.1214/20-BA1214

Rappl, A., Kneib, T., Lang, S., & Bergherr, E. (2023). Spatial joint models through bayesian structured piece-wise additive joint modelling for longitudinal and time-to-event data. https://arxiv.org/abs/2302.07020

Multivariate survival trees for prediction of lower limb injuries in professional male and female football players

Jone Renteria[1], Lore Zumeta-Olaskoaga[1,2], Eder Bikandi[3], Jon Larruskain[3], Dae-Jin Lee[4]

[1] Applied Statistics Research Line, Basque Center for Applied Mathematics, Bilbao, Bizkaia, Spain,
[2] Departamento de Matemáticas, Universidad del País Vasco UPV/EHU, Leioa, Bizkaia, Spain,
[3] Athletic Club, Medical Services, Lezama, Bizkaia, Spain,
[4] School of Science and Technology, IE University, Madrid, Spain

E-mail for correspondence: jrenteria@bcamath.org

Abstract: The aim of this study is to identify the primary risk factors for lower limb injuries in professional football players and compare the differences between male and female athletes using Multivariate Survival Trees (MST). Firstly, longitudinal data is collected for each player from several teams and seasons, including their exposure time on the field, medical injury history, and periodic screening tests. Then, a multivariate survival tree is applied to handle multiple covariates. The MST approach requires minimal statistical assumptions and provides easily interpretable prognostic rules. Finally, the analysis aims to produce trees that can help determine relevant injury factors and compare the differences between gender.

Keywords: Sports injury; Injuries risk factors; Multivariate survival trees.

1 Introduction

Sports injuries are a common occurrence in professional sports, often resulting in severe consequences for athletes and their teams. In recent years, there has been a growing interest in the prevention of sports injuries through the use of statistical models. These models aim to identify the key risk factors for injuries and develop personalized strategies to mitigate these risks.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

To fully understand the risk factors associated with injuries, it is essential to have a clear understanding of the explanatory variables collected. Several types of variables are commonly collected for injury prevention, including exposure time (the amount of time that a player spends on the field during practice and competition), movement limitations and asymmetries (any imbalances in strength or flexibility that may increase the risk of injury), functional and strength parameters, Rating of Perceived Exertion (RPE), or anthropometric data. These screening tests are used to monitor players' functional and strength parameters, assess their physical fitness, and identify movement limitations and asymmetries of the lower limbs, which are suspected to influence the risk of sports-related injury (Duke, S.R. *et al.*, 2017).

Our goal is to analyze the timing of injury occurrence and the factors that may influence it, considering variables such as exposure time, training internal load, previous injuries, and functional strength parameters. By focusing on the temporal aspect of injury risks, we aim to inform injury prevention strategies (Nielsen, R. O. et al., 2019). Incorporating functional strength parameters as covariates enables us to assess their impact on injury timing and any gender differences. This approach provides a more comprehensive understanding of injury risk factors and can inform targeted prevention strategies.

2 Data and methods

We analyzed data from five professional football teams (three male and two female) over four consecutive seasons (2017-2018 to 2020-2021). The sample includes 141 male and 61 female football players who played at least one season on their respective teams.

The R package `injurytools` (Zumeta-Olaskoaga, L., *et al.* 2022) was utilized to preprocess the data, incorporating various functions for data structuring and visualization of plots. Figure 1 shows a subsample of six players from each cohort. The figure represents the exposure time (horizontally), where the red cross indicates the injury date and the blue circle the recovery date. Moreover, the data used to generate the survival analysis dataset includes screening tests captured from periodic health examinations, players' accumulated load over time, and health records containing information about exertion perception tests and injuries across seasons. In a second step, imputation of missing values is required. This is because not all players undergo all screening tests each season. Therefore, the Multiple Imputation by Chained Equation (MICE) method is applied, after preselecting 16 of the original 397 physical test variables based on medical expert judgment. Using the Permuted Mean Matching (PMM) method, and after completing five iterations, the MICE method imputes the best value for each missing observation, resulting in a complete screening dataset.

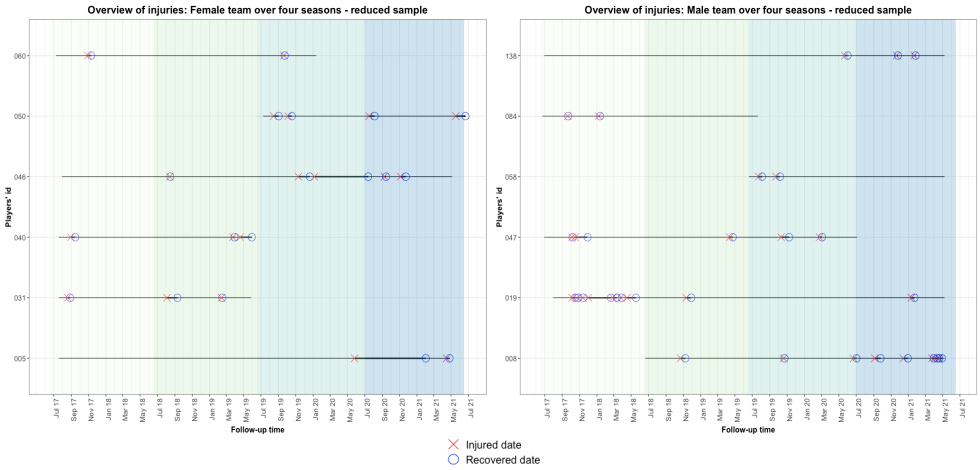


FIGURE 1. Overview of the injuries of 12 players over time. The follow-up time is split into each of the four seasons analyzed, represented with a different vertical color in the graphs.

The method selected to model the data is known as Multivariate Survival Trees (MST), which allows to define models using highly correlated data. The procedure of fitting such trees is divided into three steps: i) growing the initial tree by splitting nodes, partitioning the data into heterogeneous groups; ii) pruning the tree generated in the first step, generating subtrees; and iii) selecting the optimal size tree. For step i), three different methods are applied to handle the correlation within the variables: exponential frailty, gamma frailty, and marginal approaches. Moreover, for the ii) and iii) steps, control parameters are used: *minsplit* and *minbucket*, respectively. The former avoids overfitting while controlling the minimum number of observations at each node, and the latter controls the minimum number of observations required at the leaf node. The analysis is completed using the package `MST` in R Statistical software (Calhoun, P. *et al.*, 2018).

3 Results and concluding remarks

A selection of the optimal tree size was completed both for the female and male cohorts for each of the three different methods applied to handle the correlation within the variables. Figure 2 represents the best model generated using the marginal approach for each of the cohorts. The distinction in the sample size in both genders makes the *minsplit* and *minbucket* control parameters to be different for each of the groups. The time variable (x-axis)

within the final nodes is represented by the accumulated workload, which is the value of the time spent on the field (playing or training) multiplied by the value of the perceived exertion scale (also known as Borg).

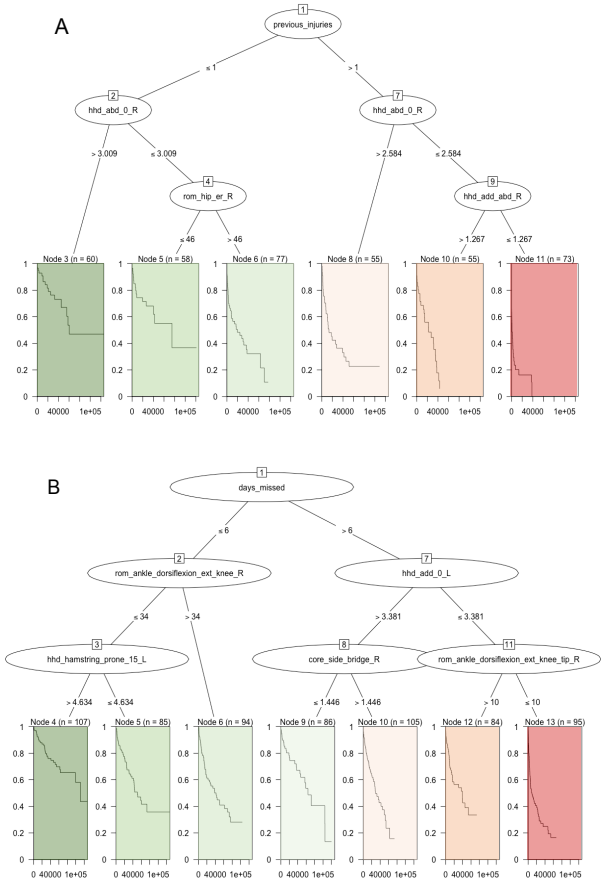


FIGURE 2. MST model results with the marginal approach. Figure (A) shows the best-fitted tree for the female cohort and Figure (B) shows the male cohort tree.

Both genders differ on the starting point node. Females start with the quantification of the preceding injuries (*previous injuries*) and males, start with the burden caused by those injuries (*days missed*). These starting point variables only consider the timeframe of the season prior to the completion of the screening test. For the female's tree, the value of the hand-held dynamometer abductor exercise is the second splitting point, followed by the abductor-adductor test and the measurement of the range of motion of the hip. However, on the men's tree, different variables have a higher effect on

the results, combining hand-held dynamometer tests on other body parts (e.g., hamstring) and core or range of motion values for the second and third node levels. Both trees on Figure 2 are sorted such that each split to the left has a lower risk of failure.

In conclusion, the use of survival trees to estimate the risk of lower limb injuries has shown promising results. By incorporating functional strength parameters as covariates, we were able to identify their impact on the timing of injuries and compare the effects between genders. This approach provides a more comprehensive understanding of the factors that contribute to injury risk and can inform targeted injury prevention strategies. With further research and validation, survival trees have the potential to become a valuable tool for sports medicine professionals in assessing and mitigating the risk of lower limb injuries in football and other sports.

Acknowledgments: We thank Medical Services of Athletic Club for data support. We acknowledge the support of the Basque Government through the BERC 2022-2025 program; of the Ministry of Science, Innovation and Universities through BCAM Severo Ochoa accreditation and through SEV-2017-0718 PRE2018-084007 funding; of AEI/FEDER, UE through the PID2020-115882RB-I00 and acronym "S3M1P4R"; and of Provincial Council of Bizkaia through the 6/12/TT/2022/00006 and acronym "MATH4-SPORTS".

References

Calhoun, P. et al. (2018). Constructing Multivariate Survival Trees: The MST Package for R. *Journal of Statistical Software*, **83**, 1 — 21.

Duke, S. R. et al. (2017). Preseason functional movement screen predicts risk of time-loss injury in experienced male rugby union athletes. *The Journal of Strength and Conditioning Research.*, **31**, 2740 — 2747.

Nielsen, R. O. et al. (2019). Time-to-event analysis for sports injury research part 2: time-varying outcomes. *British journal of sports medicine*, **53**, 70 — 78.

Zumeta-Olaskoaga, L., and Lee, D.-J. (2022). injurytools: A toolkit for Sports Injury Data Analysis. R package version 1.0.1, https://CRAN.R-project.org/package=injurytools

Focussed information criteria for model selection - A Bayesian perspective

Bijit Roy[1], Emmanuel Lesaffre[1]

[1] Katholieke Universiteit Leuven, Belgium

E-mail for correspondence: bijit.roy@kuleuven.be

Abstract: Most model selection methods are based on a global measure of model adequacy. Often one is more interested in estimating some other derived quantities or parameters from the model. The Focused Information Criteria (FIC) is a model selection tool which uses accuracy measures on the parameter of interest. In this paper we propose a Bayesian analogue to the FIC for model selection.

Keywords: Model Selection; FIC; Bayesian analysis.

1 Introduction

Model selection is an important step in statistical modelling because there are usually many different possible models that can be used to represent the data, and selecting the wrong model can lead to inaccurate predictions or incorrect inferences. Most model selection techniques involve evaluating different models based on some form of global goodness-of-fit criterion which is then penalized for model complexity. Traditional model selection criteria such as AIC (Akaike, (1974)) and BIC (Schwarz, (1978)) focus on goodness of fit for the response variable. Same is true for popular model selection measures in a Bayesian paradigm, such as DIC (Spiegelhalter *et al.* (2002)) and WAIC (Vehtari *et al.* (2017)).
On the other hand often one may be more interested in estimating or predicting some specific parameters in the model or some quantities derived from the response. Focused Information Criteria (Claeskens and Hjort (2003)), select models based on precision of the estimates of the parameter of interest (focus parameter). In this article we propose a simple Bayesian analogue of the FIC.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

2 Motivation: Modelling BMI using growth curves

Our motivating data-set comes from a data warehouse project by Danone Nutricia Research. We have data on human infant growth (weight and height), as well as background information such as country of birth, gender, and birth weight. The data contains weight and height measurements for around 1000 infants from 3 countries, recorded at several time points till 1 year of age. Our main research question is to see if there are differences in the development trajectory of BMI among the birth-weight categories (large, normal, small) of the infant, while accounting for country and gender effects.

For modelling height and weight of infants several commonly used growth curve models exist, and each of these models have a particular age range where they perform best (Chirwa *et al*, 2014). Multilevel modelling is used to account for the individual level effects. BMI has a deterministic relationship with weight and height, thus a joint bi-variate growth model for weight and height will allow us to construct a model for BMI (Roy and Lesaffre (2020)). In this context we want model selection techniques that focus on BMI, a derived quantity of the responses.

3 The FIC paradigm

Focused Information criteria (FIC) are based on the mean square errors (MSE) of a focus parameter of interest, instead of goodness of fit of the responses. The FIC for a model M is the estimate for the MSE of the focus parameter $\widehat{\mu_M}$, $FIC(M) = \widehat{mse(\mu_M)}$. Under the original version of FIC as introduced in Claeskens and Hjort (2003, 2008) the FIC was calculated under a locally asymptotic framework and an assumed true unknown data generating model. This was extended to a "fixed wide model" framework in Claeskens *et al.* (2019), which allowed for FIC to be applied to complicated data structures including longitudinal models. A sufficiently large and comprehensive wide model $\mathbf{Y} \sim F_{wide}(Y; \theta_{wide})$ is assumed to replace the true model. All the variances and biases for calculating the FIC are now calculated under this estimated wide model. Thus $FIC(M) = Var_{\widehat{F_{wide}}}(\widehat{\mu_M}) + (E_{\widehat{F_{wide}}}(\widehat{\mu_M}) - \widehat{\mu_{wide}})^2$, and $FIC(wide) = Var_{\widehat{F_{wide}}}(\widehat{\mu_{wide}})$.

4 Proposed Bayesian FIC

In a Bayesian approach we simply replace the expected bias and variance with their posterior counterparts. In analogous to the original FIC approach, we still use the parameter estimates from the wide model in place of the true values of the parameter while calculating the bias. This the Bayesian version of the FIC can be written as

$$BFIC(M) = Var_M(\widehat{\mu_M}|X) + (E_M(\widehat{\mu_M}|X) - E_{wide}(\widehat{\mu_{wide}}|X))^2,$$

with $E_M(.|X)$ denoting the posterior expectation under the model M.

5 Results

5.1 Simulation Results

We replicate the simulation framework for longitudinal models in Section 6 of Cunen *et al.* (2020). The 100 data-sets each has 20 groups with 15 observations in each, and 4 potential fixed and random effects (the wide model M_0. The four experiments each has a different focus parameters. We refer the readers to Cunen *et al.* (2020) for the details. Figure 1 shows the FIC for different focus parameters, models and simulated data-sets. (The red crosses denote the models with least FIC for each data set.) We see that in most cases the models M_0, M_1, and M_2 are preferred by the FIC. In contrast the frequentist version of FIC prefers mostly M_1 and M_2.

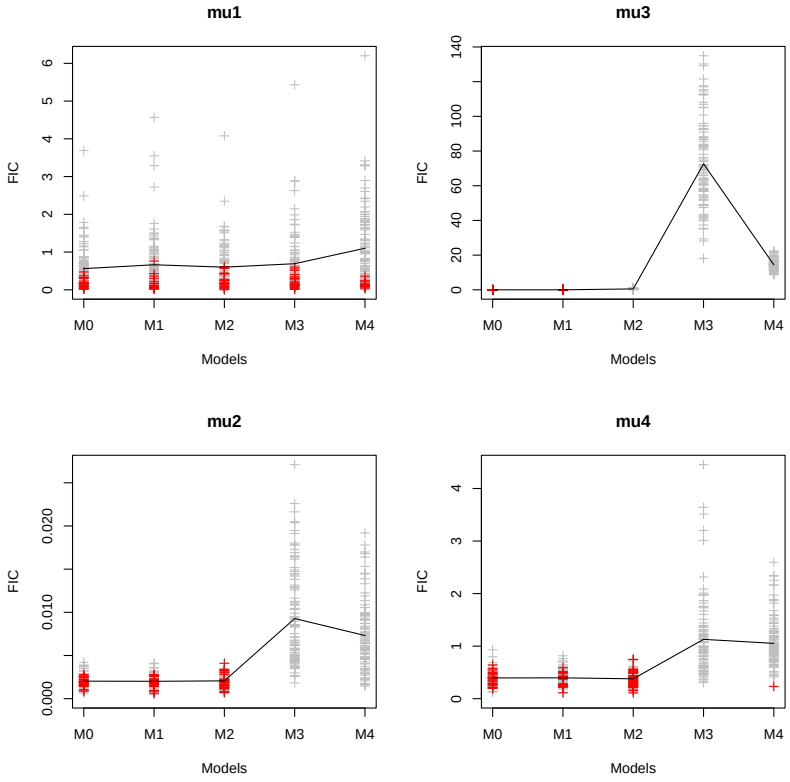


FIGURE 1. FIC for 100 simulated data-sets.

5.2 Application to BMI modelling

We have longitudinal growth data (weight and height) for around 1000 infants form 3 countries, Australia, Spain and Thailand. Together with the

3 birth-weight categories, and gender this makes 18 subgroups. We are interested in BMI trajectories vary among the 3 birth-weight categories. To show the easy applicability of the proposed BFIC approach, we use two approaches for modelling the BMI trajectory. First, we will use a joint bivariate growth curve model (the first order Berkey-Reed model (Berkey and Reed (1987))) for weight and height, as in Roy and Lesaffre (2020). Second, we will use generalized additive models (GAM) (Wood (2017)), with smooths for both individual and group effects to model the BMI directly. Since our main research question wants to know if there are significant differences in BMI trajectory for the three weight categories, we use mean BMI for each of the subgroups at 1 year mark as our focus parameter.

TABLE 1. Models used for BMI

Model 0	$(wt, ht) \sim sex : wt\_cat : country+$ $sex : wt\_cat : country : (t + ln(t) + 1/t)+$ $(1 + t + ln(t) + 1/tID)$
Model 1	$(wt, ht) \sim sex + wt\_cat + country+$ $sex : (t + ln(t) + 1/t) + wt\_cat(t + ln(t) + 1/t)+$ $country : (t + ln(t) + 1/t) + (1 + t + ln(t) + 1/t\|ID)$
Model 2	$(wt, ht) \sim sex + wt\_cat + country+$ $(t + ln(t) + 1/t) + ((1 + t + ln(t) + 1/t\|ID)$
Model 3	$(wt, ht) \sim sex + wt\_cat + country+$ $country : (t + ln(t) + 1/t) + (1 + t + ln(t) + 1/t\|ID)$
Model 4	$(wt, ht) \sim sex + wt\_cat + country+$ $sex : (t + ln(t) + 1/t) + (1 + t + ln(t) + 1/t\|ID)$
Model 5	$(wt, ht) \sim sex + wt\_cat + country+$ $wt\_cat : (t + ln(t) + 1/t) + (1 + t + ln(t) + 1/t\|ID)$
Model 6	$(wt, ht) \sim sex + wt\_cat + country+$ $sex : wt\_cat : (t + ln(t) + 1/t) + (1 + t + ln(t) + 1/t\|ID)$
Model 7	$(wt, ht) \sim sex + wt\_cat + country+$ $wt\_cat : country(t + ln(t) + 1/t) + (1 + t + ln(t) + 1/t\|ID)$
Model 8	$(wt, ht) \sim sex + wt\_cat + country+$ $sex : country : (t + ln(t) + 1/t) + (1 + t + ln(t) + 1/t\|ID)$
Model 9	$BMI \sim s(t) + s(t\|sex : wt\_cat : country) + s(t\|ID)$
Model 10	$BMI \sim sex + wt\_cat + country + s(t) + s(t\|ID)$
Model 11	$BMI \sim wt\_cat + country + s(t) + s(t\|sex) + s(t\|ID)$
Model 12	$BMI \sim sex + country + s(t) + s(t\|wt\_cat) + s(t\|ID)$
Model 13	$BMI \sim sex + wt\_cat + s(t) + s(t\|country) + s(t\|ID)$

The wide model M_0 is the fully saturated model with the intercept and slopes all varying for the 18 subgroups. Models $M_1 - M_8$ used for comparison were sub-models of M_0. These sub-models involves the intercepts and slopes varying with different levels of grouping. Models $M_9 - M_{13}$ were GAMs used to directly model the BMI. Model M_9 had group level

smooths for all the 18 subgroups, as well as individual level smooths. Models $M_{10} - M_{13}$ had group level smooths with different levels of grouping. Refer to Table 1 for details of the models.

In this example we treat the expected BMI at 1 year for each subgroup as the focus parameter of interest. Figure 2 shows the variance, bias square and FIC for each of the 18 subgroups, as well as the average over the 18 subgroups. Model M_8 has the best average FIC (1.026) among the chosen models (in contrast the wide model has an average FIC of 1.331).

Focusing on the actual quantity of interest (BMI) allows us to choose a simpler model than the fully saturated model. This provides other incidental benefits like faster convergence in a Bayesian context. Moreover our selected model shows that in the context of predicting average BMI at 1 year the birth weight categories do not influence the BMI development trajectories (The slopes do not vary with birth weight category.) However, Model M_6, which included weight category as a factor was in a very close second place (average FIC 1.054). In contrast the GAMs using smoothing splines had worse FIC, and were thus not selected. This also shows the validity of our approach of using two simple models for height and weight, instead of going for a more complicated spline type model for BMI directly. However, it should be kept in mind that the FIC procedure should not be used for deciding whether certain factors actually effect the responses, as a larger data-set or different focus parameters may have different results.

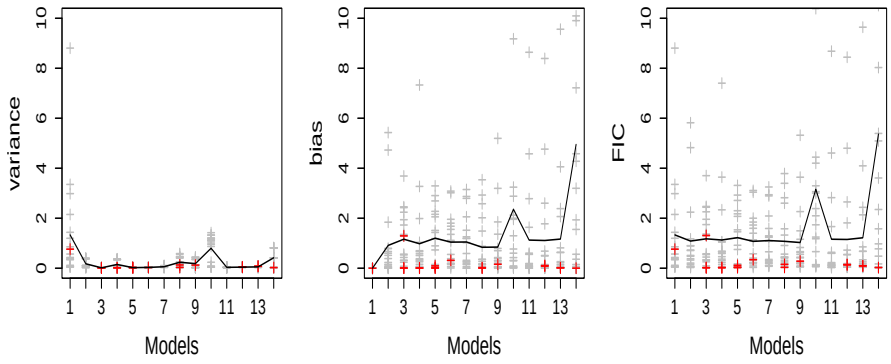


FIGURE 2. Variance, bias, and FIC for average BMI at 1 year for 18 subgroups.

6 Conclusion

We have proposed an analogue to the FIC approach for model selection in a Bayesian paradigm. One benefit of our suggested approach is that this is applicable to a much wider class of models and more general classes of focus parameters of interest than the original FIC approach. Any quantity

for which we can get a posterior estimate can now be a focus parameter. Also our proposed FIC avoids complicated Jacobian calculations specific to each model and focus parameter, and can be easily calculated from the posterior samples, and can be used to compare among widely different classes of models. This is illustrated as the two categories of models we used are, bi-variate growth models for weight and height, and GAMS with group and individual level smooths for BMI.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19(6)**, 716-723.

Berkey, C.S. and Reed, R.B. (1987). A model for describing normal and abnormal growth in early childhood. *Human Biology*, **59**, 973 – 987.

Chirwa, E.D. *et al.* (2014). Multi-level modelling of longitudinal child growth data from the Birth-to-Twenty Cohort: a comparison of growth models. *Annals of Human Biology*, **41 (2)**, 168 – 179.

Claeskens, G., and Hjort, N. L. (2003). The focused information criterion [with discussion contributions and a rejoinder]. *Journal of the American Statistical Association*, **98(464)**, 900-916.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Claeskens, G., Cunen, C., and Hjort, N. L. (2019). Model selection via focused information criteria for complex data in ecology and evolution. *Frontiers in Ecology and Evolution*, **7**, 80.

Cunen, C., Walloe, L., and Hjort, N. L. (2020). Focused model selection for linear mixed models with an application to whale ecology. *Journal of Agricultural, Biological, and Environmental Statistics*, **25**, 404-420.

Roy, B., and Lesaffre, E. (2020). Bayesian modelling of complex functional forms. In: *Proceedings of the 35th International Workshop on Statistical Modelling*, 414-418.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6(2)**, 461-464.

Spiegelhalter, D. J. *et al.* (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64(4)**, 583-639.

Vehtari, A., Gelman, A., and Gabry, J. (2017) . Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27(5)**, 1413-1432.

Wood, S. N. (2017) . *Generalized additive models: an introduction with R*. CRC press.

Spatio-temporal modelling using an opportunistically sampled open-survey data: a simulation study based on the Belgian Great Corona Study.

Alejandro Rozo[1,2], Christel Faes[2], Thomas Neyens[1,2]

[1] KU Leuven, L-BioStat, Belgium
[2] Hasselt University, CenStat, Belgium

E-mail for correspondence: josealejandro.rozoposada@kuleuven.be

Abstract: Online open surveys are a popular source of information to complement epidemiological surveillance programs. Due to the absence of a sampling protocol, open surveys typically yield to opportunistically sampled data. This requires care when interpreting statistical modelling results, as these methods commonly assume a randomised study design to underlie the data collection process. It remains unclear to which extent such data, which usually lead to spatio-temporal imbalanced samples, is useful to detect spatio-temporal trends in epidemiological phenomena such as disease risk. We propose a simulation study based on Flemish and Brussels COVID-19 symptoms data obtained via the Great Corona Study, a Belgian large-scale, online, open survey operational throughout the COVID-19 pandemic. We show the impact of opportunistic sampling on detecting epidemiological trends when spatio-temporal modelling approaches are considered. We find that traditional spatio-temporal disease mapping methods often work well when sample sizes are large but tend to perform poorly when substantial spatial and temporal data imbalance simultaneously occur.

Keywords: Opportunistic sampling; Sample size; Spatio-temporal modelling; Spatial modelling; Epidemiological surveys.

1 Introduction

Large-scale open, and often online, surveys have become a popular and cost-effective means to gather epidemiological information. However, many of these surveys lack a sampling protocol, as anyone interested in participating can provide data. Consequently, resulting samples are often termed

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

'opportunistic', in contrast to 'randomised'. This distinction can be problematic because traditional models to analyse areal data as spatial and spatio-temporal models typically assume a randomised sample obtained through adherence to a study protocol.

Opportunistic samples can exhibit multiple data characteristics that are suboptimal for data analysis, many of which are associated with strong variability in individuals' participation efforts. These characteristics include survey incompleteness, underrepresentation of individuals with lower digital literacy, and preferential sampling, which occurs when the epidemiological trend of interest and trends in sampling are stochastically related (Diggle, P.J. et al. (2010)). Although model corrections for opportunistic samples exist (e.g. Diggle, P.J. et al. (2010)), they are often difficult to implement. Therefore, the standard methods are often still applied in such circumstances, assuming that the large sample sizes will allow the detection of the important spatial or spatio-temporal trends sufficiently well. However, the validity of this assumption remains uncertain, especially since opportunistic samples typically result in geographical and temporal imbalances in survey participation rates. The impact of such imbalances on statistical inference and predictions is not yet well understood.

The current research is motivated by a case study that compared the spatio-temporal incidence estimates based on two data sources: (i) confirmed COVID-19 cases data collected by Belgian governmental agencies; and (ii) COVID-19 symptoms based on self-reporting, collected via a weekly Belgian online open survey, titled the Great Corona Study (GCS)(University of Antwerp et al. (2020)). The case study, which is not reported here, showed that despite the very large amount of self-reported symptomatic information, the opportunistically sampled COVID-19 symptoms data provided only, to a limited extent, valuable insights into the spatio-temporal trends in COVID-19 incidences. There were strong suggestions that the opportunistic sampling nature of the data lay at the root of this problem.

A simulation study was therefore undertaken aiming at (i) assessing the effect of imbalanced opportunistic samples and (ii) samples sizes when detecting epidemiological insights through large-scale open surveys.

2 Methods

This study used two data sources collected during six weeks from the early period of the Belgian COVID-19 pandemic (March 17- May 11, 2020): (i) the daily test-confirmed cases reported by the Belgian Health Institute, Sciensano, in all municipalities of two regions of Belgium, i.e., Flanders and the Brussels-region; and (ii) the weekly counts of people experiencing COVID-19 symptoms obtained from the GCS. The confirmed COVID-19 cases were temporally aggregated on a weekly level. We carried out this study in three stages: (i) simulation; (ii) modelling; and (iii) comparison.

In the **simulation stage**, we first defined distinct sampling strategies that differed in the spatio-temporal balance of the sample and the total sample size.

Regarding the spatio-temporal balance, we considered four scenarios: (i) balance in time and space; (ii) balance in time but not in space; (iii) balance in space but not in time; and (iv) unbalance in time and space. Note that scenarios (ii), (iii) and (iv) correspond to opportunistic samples. Concerning the sample size, we started from the total sample sizes obtained in each of the scenarios, which we multiplied with multiple constants (1/6, 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5 and 6). Combining the four scenarios and the different sample sizes led to 44 different sampling strategies to be compared.

Secondly, we specified the sample size for each municipality i at each week of analysis t according to each sampling strategy. Thirdly, using these sample sizes, we simulated, within each sample, the number of people experiencing a symptom of COVID-19, assuming that the data-generating mechanism is a binomial distribution with two parameters. (i) The sample size, N_{it}, associated with one of the 44 sampling strategies, and (ii) the proportion of occurrence, p_{it}, corresponding to the proportion of COVID-19 cases observed at each municipality and week, information obtained from the official governmental reports. Finally, each sampling strategy was simulated 100 times, producing 4400 different datasets.

During the **modelling stage**, we analysed two types of data; (a) the weekly observed COVID-19 cases from the governmental agencies and (b) each of the simulated samples of people experiencing COVID-19 symptoms. We, therefore, undertook a spatio-temporal modelling approach that use conditional autoregressive smoothing across the spatial and temporal dimensions, and a type I spatio-temporal interaction proposed by Knorr-Held (2000).

Finally, in the **comparison stage**, we obtained incidence estimates from the observed COVID-19 cases, as well as from each simulated sample of people experiencing COVID-19 symptoms. Subsequently, we compared the incidence estimates from the observed COVID-19 cases with the incidence estimate from each simulated sample. Two metrics were used: (i) the correlation coefficient and (ii) the proportion of agreement in the top 20 ranking of highest incidences each week of analysis. Note that the comparison was based on estimates from both modelling approaches. The correlation coefficient intends to measure to which extent the incidence estimates of people experiencing COVID-19 symptoms, under different sampling strategies, can capture the actual overall trends of the disease spread that the incidence estimates of COVID-19 cases show. In comparison, the proportion of agreement seeks to assess how similar the detection of hotspots is.

3 Results and discussion

Figure 1 shows that larger sample sizes are associated with higher correlation coefficients between incidence estimates. Moreover, any type of balance, either in time or space, produces scenarios with similar results as in the full balance scenario. However, the scenario with sampling unbalances in both space and time underperforms substantially.

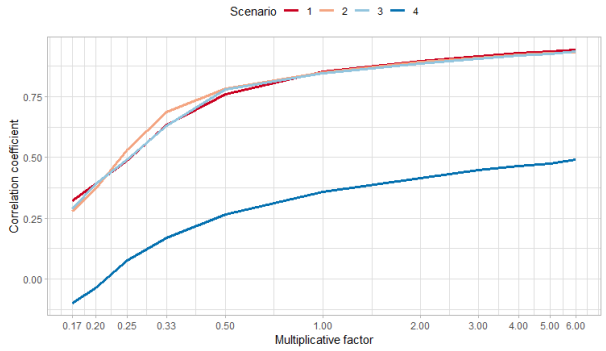


FIGURE 1. Average correlation coefficients between incidence estimates of COVID-19 cases and incidence estimates of simulated COVID-19 symptoms, for each of the different sampling strategies.

Figure 2 shows the proportion of agreement in the top 20 ranking of municipalities with the highest incidence estimates; we observe that as the sample size increases, the proportion of agreement increases linearly.

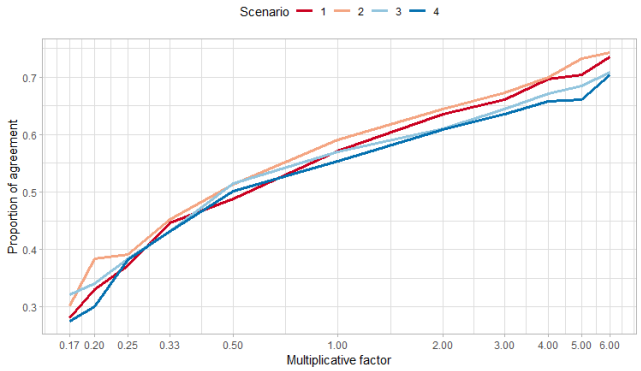


FIGURE 2. Average proportion of agreement in the top 20 ranking of the municipalities with the highest incidence estimates of COVID-19 cases and simulated COVID-19 symptoms, for each of the different sampling strategies.

This simulation study takes out of the picture some of the limitations that the case study presents, we assume that the data-generating mechanism of the symptoms is the same as the COVID-19 cases, meaning that all COVID-19 cases are symptomatic and that there is no time lag between the onset of symptoms and confirmation of a COVID-19 case. This assumption allows us to isolate the effect of the sample size and spatio-temporal balance and explore its implication providing epidemiological insights.
Finally, we observed that the sample size is essential and helps to obtain better results. However, models based on opportunistic samples can underperform, even when the opportunistic samples have large sample sizes. Furthermore, the sample's representativeness is also vital, partly achieved by spatio-temporal balance. This research finds that the methods remain useful when there is some unbalance (in either time or space). However, the models seem to break down when unbalance occurs in time and space simultaneously. This is an important message to researchers setting up such studies: they should ensure that some form of balance is achieved.

References

Diggle, P.J., Menezes, R., & Su, T.L. (2010). Geostatistical inference under preferential sampling. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 59(2), 191-232.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. In *Statistics in medicine* 19(17-18):2555–2567.

Murray J, and Cohen AL. (2017). Infectious Disease Surveillance. In: *International Encyclopedia of Public Health.* Alabama, USA, 222–9.

University of Antwerp, KU Leuven, Hasselt University (2020). The Great Corona Study. https://www.uantwerpen.be/en/projects/great-corona-study/

Meta-analysis of variability in survival outcomes in precision oncology trials

Maximilian Schuessler[1,2], Elizaveta Skarga[2], Pascal Geldsetzer[3], Ying Lu[1] & Maike Hohberg[4]

[1] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[2] Institute for Global Health, University Hospital Heidelberg, Ruprecht-Karls-University Heidelberg, Heidelberg, Germany
[3] Division of Primary Care and Population Health, Department of Medicine, Stanford University, Stanford, CA, USA
[4] Department of Medical Statistics, University Medical Center Goettingen, Goettingen, Germany

E-mail for correspondence: maike.hohberg@med.uni-goettingen.de

Abstract: Heterogeneity in survival outcomes for biomarker-enriched alone versus enriched and non-enriched cohorts combined is not yet fully understood. Coefficient of variation ratios have been proposed as a meta-analytic measure of variability. However, their application to survival data from clinical trials remains a methodological gap. We develop a methodological procedure that leverages coefficient of variation ratios (CVRs) to meta-analyse survival outcomes from clinical trials. We apply this procedure to biomarker-enriched subgroup versus trial population (enriched and non-enriched cohorts combined) in cancer trials. Our preliminary results suggest that CVRs, derived from treatment and control groups, are smaller in biomarker-enriched subgroup than that in the trial population, and that this reduction in variability is driven by the treatment group in the biomarker-enriched cohort.

Keywords: coefficient of variation; meta-analysis; precision medicine

1 Introduction

Molecular profiling and novel trial designs are powerful to match patients' tumor profile with more targeted therapies in oncology. Despite the benefits and promises of increasingly refined therapies, the heterogeneity in survival outcomes for biomarker-enriched subgroup alone versus total trial

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

population (enriched and non-enriched cohorts combined) is not fully understood. Approaches for meta-analytic measures of variability have been successfully developed (senior et al. 2020, Nakagawa et al. 2015) and applied (Winkelbeiner et al. 2019) for normally distributed outcomes, but have not yet been developed for survival outcomes.
The objective of this study is to develop a methodological procedure for using coefficients of variation ratios (CVRs) for meta-analysing survival outcomes from cancer trials. The second objective is to assess whether biomarker-based treatments result not only in mean improved survival outcomes, but also increased precision in treatment response of biomarker-enriched subgroups in cancer trials.

2 Methodology

To construct a trial cohort for our meta-analysis, we conducted a literature search of biomarker-stratified phase II and phase III cancer trials. Clinical trials were included if i) they reported survival outcomes, including overall survival (OS) and/or progression free survival (PFS), for both the total trial population and biomarker-enriched subgroups, and ii) the treatment under investigation had a an established mechanistic link that would drive differential outcomes in the biomarker-enriched subgroup.
To estimate variability in the control and treatment groups, we reconstructed pseudo individual participant data (IPD) from published Kaplan-Meier survival curves making use of the `R` package `kmdata` (Redd et al. 2022). Using this pseudo-IPD data, we estimate parametric models, starting with a lognormal model, and rely on restricted mean survival time (RMST) to calculate:

$$E\left(Y_{t,k}(\tau)\right) = \int_0^\tau S_{t,k}(y;\theta)dy \quad \text{and}$$
$$sd^2\left(Y_{t,k}(\tau)\right) = 2\int_0^\tau yS_{t,k}(y;\theta)dy - E^2\left(Y_{t,k}(\tau)\right)$$

where $Y_{t,k}$ is the survival time for treatment ($t = 1$) or control ($t = 0$) group in the k'th trial, S denotes the survival function of the parametric model with parameters θ, and τ is the time limit of the RMST. We rely on RMST here, since calculating the mean and variance of the distribution extrapolates to unobserved survival times.
To compare the variation in treatment and control group, we calculate coefficient of variation ratios (CVR) for each trial, as described by Senior (2020):

$$\ln(CVR) = \ln\left(\frac{sd_T}{\bar{x}_T} \,/\, \frac{sd_C}{\bar{x}_C}\right) + \frac{1}{2}\left(\frac{1}{n_T - 1} - \frac{1}{n_C - 1}\right) + \frac{1}{2}\left(\frac{sd_T^2}{n_T \bar{x}_T^2} - \frac{sd_C^2}{n_C \bar{x}_C^2}\right), \quad (1)$$

where sd denotes the standard deviation, n the sample size, and $\bar{x}$ the mean survival time of treatment group T and control group C. The first term of the formula is a "naïve" estimator of the CVR while the latter terms serve as a bias correction for small samples. The associated sampling variance of $ln(CVR)$ is given by:

$$Var(\ln CVR) = \frac{sd_C^2}{n_C \bar{x}_C^2} + \frac{sd_C^4}{2n_C^2 \bar{x}_C^4} + \frac{n_c}{(n_c - 1)^2} + \frac{sd_T^2}{n_T \bar{x}_T^2} + \frac{sd_T^4}{2n_T^2 \bar{x}_T^4} + \frac{n_T}{(n_T - 1)^2}. \quad (2)$$

After calculating coefficients of variation ratios (CVRs) for all trials as in equation (1), we perform meta-analysis using random effects models to estimate the relative variability between treatment and control groups, where the inverse of the variance is used as weight. Note that in the future, we will replace the variance formula in equation (2) to also account for censored data. Finally, we compare CVRs (with bootstrap confidence intervals) from biomarker-enriched subgroups with those from intention-to-treat (ITT) populations.

3 Results

For our preliminary analysis, we identified seven breast cancer trials with a total of around 3,500 participants and successfully reconstruct IDP for total and biomarker-enriched cohorts and their respective outcomes overall survival and progression-free survival. Figure 2 forest shows the CVRs from each individual trial for the ITT and the biomarker enriched cohorts, respectively. The preliminary results show lower variability ratios between treatment and control for biomarker-enriched subgroups compared to the ITT cohort, suggesting more homogeneous response in biomarker-enriched subgroups compared to the ITT trial population. For all but one trial, we see a trend towards smaller CVRs for the biomarker-enriched subgroups, suggesting a more homogenous response in groups that are biomarker-enriched. However, the associated confidence intervals overlap for the two cohorts.

When aggregating the individual CVRs in a random effects meta-analysis, the CVR is 0.950 [0.892, 1.002] for the ITT group and 0.865 [0.790, 0.945]

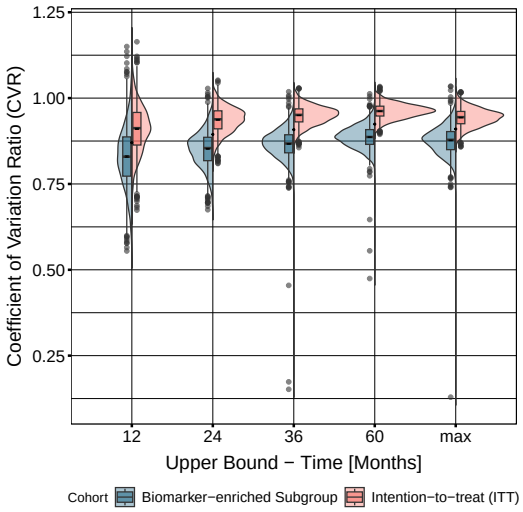


FIGURE 1. The violin plot depicts the distribution of the CVR in 2000 bootstrap samples for different time restrictions τ. Maximum upper bounds (max) are the maximum of observed survival times. Boxplots depict the interquartile range (IQR) and one standard error of deviation.

for the biomarker-enriched. This suggests that for the biomarker-enriched subgroup, response to treatment has lower variability in the treatment group relative to the control group.

To assess if the decrease in CVR is driven by a variability decrease in the treatment or control group, we compare the assess ratio of the coefficient of variation (CV) for subgroup over the CV of the ITT group, i.e. $CV_{enriched}/CV_{ITT}$.

Since we do not expect a change in the variation of the control group between ITT cohort and biomarker-enriched subgroup, this ratio should be around 1 for most trials. For the treatment group, we expect values below 1 if there is less variation in the biomarker-enriched subgroup, i.e. treatments targeted to a biomarker lead to more precise response. Our preliminary results point towards this trend, with 6 out of 7 trials having a CV ratio in the treatment group that is below 1. Figure 1 shows violin plots of meta-analyzed CVRs in 2000 bootstrap samples in ITT and biomarker-enriched cohorts for varying time windows used to construct RMST. This corroborates our previous results that the CVR is lower for biomarker-enriched cohorts. Since both distributions are overlapping, more data are necessary to confirm this trend. Finally, our results seem to be robust across the different choices of τ, with distribution becoming narrower for larger τ.

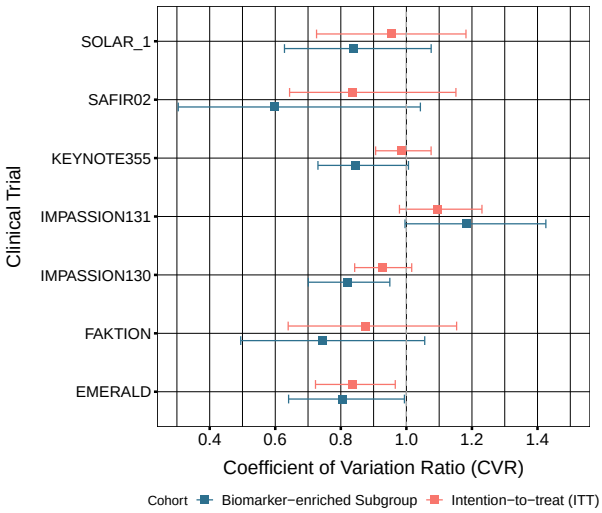


FIGURE 2. The forest plot shows the CVRs for each trial for overall survival with estimation restriction at 36 months, stratified by the intention-to-treat (ITT) cohort and the biomarker-enriched subgroup. Error bars indicate 95% bootstrap percentile intervals.

4 Conclusion

We demonstrate a feasible approach to construct coefficient of variation ratios (CVRs) for survival outcomes using Kaplan-Meier curves from precision oncology trials. We apply this approach to seven breast cancer trials that report overall survival and progression free survival for total trial populations and biomarker-enriched subgroups. Preliminary results show a trend towards lower CVRs for biomarker-enriched subgroups compared to their intention-to-treat group, suggesting more homogeneous response in biomarker-enriched subgroups. Further work in progress includes incorporating more trials from different tumor entities, assessing the quality of the reconstructed pseudo-data, and comparing different (semi-)parametric models within our procedure.

References

Nakagawa, S.. Poulin, R., Mengerson, K., Reinhold, K., Engqvist, L., Lagisz, M., Senior, A.M. (2015). Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, **6**, 143 – 152.

Redd R., Fell G., Rahman R. (2022). kmdata: A Database of Reconstructed Individual Patient Level Data from Oncology Clinical Trials. *R package version 1.0.1*

Senior, A.M., Viechtbauer, W., and Nakagawa, S. (2020). Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Research Synthesis Methods*, **11**, 553 – 567.

Winkelbeiner, S., Leucht, S., Kane, J., and Homan, P. (2019). Evaluation of differences in individual treatment response in schizophrenia spectrum disorders: a meta-analysis. *JAMA psychiatry*, **76**, 1063 – 1073.

Challenges in statistical consulting for Animal Science

Sabine K. Schnabel[1]

[1] Biometris, Wageningen University and Research, The Netherlands

E-mail for correspondence: sabine.schnabel@wur.nl

Abstract: In statistical consultation different research domains come with different data and analysis challenges. In this overview we are presenting a few of the common situations we encounter when collaborating with researchers in Animal Science. We describe different types of studies, data types and data collection as well as problems that arise during the analysis. The general description is illustrated with three examples that have been analyzed within a (generalized) linear mixed model framework.

Keywords: Experimental studies; behavioral data; agriculture; GLMM.

1 Introduction to data from Animal Science

Different research domain create and work with different types of data. The focus here is on data in the field of Life Sciences, more specifically from research projects within Animal Science. This field includes for example research on different types of species, their interactions, their influence on climate and other questions. Traditionally data from animals can be collected in observational as well as experimental studies. Often observational data is more used in ecological studies. While these data also have a lot of challenges in their analysis, this will not be the focus of the abstract. In the following we will be only looking at experimental data. As a glimpse into a few cases we will list some examples in Section 2.
In statistical consulting and analysis for experimental data from animal studies we are mainly dealing with data from farm animals that are living in stationary buildings. Most data is collected on animals such as cows, calves, pig(let)s, sows, hens and chickens. The animals are usually housed in smaller or larger units (pens) with multiple animals in one of them. The housing units can be equipped with different installations in addition to

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

the required standards. Oftentimes different housing conditions are used as experimental factors in the design, such as different types of flooring, heating, additional ventilation, lighting, number of animals per pen, extra equipment such as brushes, more water stations etc.
As the aims of the analyses are very diverse, also the type of variables and aggregation levels can be very different. In addition to numeric variables we also often deal with percentages or proportions (e.g. proportions of active animals in a pen at any given time), scores (e.g. for indicating the severeness of a (health) condition), binomial data etc. Due to the nature of the setup of the experiments as well as the different treatments and measurements data are often on different levels of aggregation: mainly animal based or e.g. small housing unit based. In addition observations are often longitudinal (on different scales).
Different types of data and study aims require different types of methods. In the examples listed below we ultimately chose for a (generalized linear) mixed model framework. While we aim to consult with the researchers before start of the experiments in order to discuss the experimental design from a statistical point of view, more often than not are we confronted with designs that can be challenging to accommodate in more commonly used models and will thus require adapted or new methods. For example this can be due to multiple layers of nested and crossed random effects due to the interaction between housing arrangements and treatments. Depending on the type of experiment and variables we also are faced with different sample sizes. For example studies on behavior of animals are often smaller as the data collection can be very expensive (e.g. in terms of time). Large amounts of longitudinal data are often generated from the use of sensors.
We are presenting the studies in a neutral description focussing on the statistical challenges. No ethical questions will be discussed. In the examples cited below all necessary ethical permissions were approved.

2 Examples from consultation practice

2.1 Health status of cows using sensor data

The aim of this project was to relate the postpartum health status of a cow to (derived) measurements from sensors that were placed on the animals pre-partum as well as some general characteristics of the cows. The experiment included 180 cows in 4 different farms (Van Dixhoorn et al., 2023) followed over a period of 8 weeks. The housing conditions and other circumstances were standardized between the different locations. The response variable in this case was a measure for the global health assessment of the cow. It is expressed as a numeric variable. The animals are monitored with sensors on the neck and leg. The sensors recorded different activity features (respectively eating, ruminating, inactive or active or step counts and a count of transitions from lying to standing or walking). These

were aggregated into hourly data. Characteristics from these time series were derived and used as explanatory variables in the analysis. Additionally we control for parity in the analysis as well the influence of farm (as a random effect). Through different stages of variable selection (univariate pre-selection, stepwise/all possible subset selection) we identified a final model including only few of the variables derived from the sensor measurements. The results of this study will form the basis for further research to use these non-invasive sensor measurements as a diagnostic tool to identify a problematic health status.

2.2 Influential factors on calf weight

A larger dataset (250000 records) was collected by companies supplying feed for calf rearing. While all calves are subject to the same nutritious content and feeding regime, they come from a large number of different farms of origin and will be housed on a large number of farms in the second phase of their development. The variable of interest is here the growth of the calves. The aim of the analysis is two-fold: trying to identify a ranking of calf-rearing farms as well as investigating the factors that influence the growth of the animals. These factors are very diverse and range from characteristics of the animals themselves, to variables relating the farm of origin and destination, grouping over time (as animals arrive at a certain age, potentially together from the same farm) and other. Some of the variables are from surveys while other are measured on animal or housing unit level.

2.3 Housing conditions affect behavior

This is an example for a study in animal behavior. This is usually studied in smaller studies due to the labor-intensive data collection. Often the variables are generated from video imaging that still involves a lot of manual annotation. Treatments in these types of studies can be very diverse and can range from size of the pen, number of animals housed to design and equipment of the unit, different access to food sources etc. Here, challenges for the analysis often involve the proper treatment of the type of data describing a behaviour: present/not present, percentages of shown behavior in a certain timeframe and the like. The level of aggregation over time or over animals versus housing units is regularly discussed with the researchers. Additional complications arise when the behaviors of interest are rare.

Acknowledgments: Special thanks to all researchers involved in the examples, especially from Wageningen Lifestock Research as well as the faculty of Animal Sciences at Wageningen University.

References

van Dixhoorn, I.D.E., de Mol, R.M., Schnabel, S.K. et al. (2023). Behavioral patterns as indicators of resilience after parturition in dairy cows. *Journal of Dairy Science*, in Press.

Neural additive quantile regression

Quentin E. Seifert[1], Elisabeth Bergherr[1], Benjamin Säfken[2]

[1] Chair of Spatial Data Science and Statistical Learning, University of Goettingen, Goettingen, Germany
[2] Chair or Applied Data Science, TU Clausthal, Clausthal, Germany

E-mail for correspondence: quentinedward.seifert@uni-goettingen.de

Abstract:
We introduce the Quantile Neural Additive Model (QNAM), an extension of the Neural Additive Model (NAM), which brings the Generalized Additive Model (GAM) into the deep learning context by replacing its smoothers with feature specific subnetworks. Due to the conceptual simplicity of the NAM, this extension of the framework deeper into the field of distributional regression is straightforward to implement and its ability to circumvent problems of established quantile regression approaches will make it a valuable tool for quantile regression in the future.

Keywords: Neural Networks; Additive Models; Quantile Regression.

1 Introduction

Quantile regression aims to estimate conditional quantiles of data dependent on explanatory variables and is used in settings in which one is interested in covariate effects on extreme values. Compared to other distributional regression models, quantile regression does not rely on any distributional assumptions in its basic form, however, Bayesian extensions of quantile regression make use of auxiliary distributions.
The additive quantile models (QGAMs) introduced in Fasiolo et al. (2021) offer the same conveniences as Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1986), hence they allow for the inclusion of many different types of effects and smoothers with automatic smoothing parameter choice. However, working with very large datasets might become problematic. Furthermore, the common smoothers have issues estimating jagged shape functions or dealing with structural breaks. Both of these problems do not play a role for neural networks. Making use of gradient descent

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

based optimizers with mini-batches, they are known to be easily scalable. Additionally, due to not relying on low-rank smoothing approaches, they have the advantage of being able to flexibly approximate more challenging shape functions.

Neural Additive Models (NAMs) as introduced in Agarwal et al. (2021) are a novel type of artificial neural network with directly interpretable covariate effects. Conceptually, NAMs take an established model class and simply translate it into the deep learning context by replacing the smoothers from GAMs with neural networks. Making use of the NAM-framework, we introduce the Quantile Neural Additive Model (QNAM). For the QNAM, we simply modify the NAM by changing the loss function from the negative log-likelihood of the assumed distribution to the pinball loss introduced by Koenker and Bassett Jr (1978).

2 Methods

In GAMs, the predictor η is the sum of smooth terms denoted as $f_j(z_j)$ and a parametric model part $\boldsymbol{X}\boldsymbol{\beta}$ such that

$$\eta = \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{J} f_j(z_j)$$

with $\eta = g(\mu)$, μ being the canonical parameter of a distribution of the exponential family which is linked to the predictor using the link function g. The smooth terms can usually be expressed as linear combinations of basis functions evaluated at z. NAMs follow the same basic structure as GAMs and model the predictor to be the sum of multiple independent covariate effects. However, instead of using the common smoothers, usually splines, these effects are learned using neural networks. Compared to conventional fully connected neural networks, every input feature has its own subnetwork, the outputs of which are summed for the model prediction. As in GAMs, this additive structure allows for interpretability of the covariate effects.

The QNAM is simply a NAM which minimizes the pinball loss defined as

$$\sum_{i=1}^{n} w_\tau(y_i, \hat{y}_{i,\tau})|y_i - \hat{y}_{i,\tau}|,$$

with $\tau \in (0, 1)$ being the quantile to be estimated, $\hat{y}_{i,\tau}$ being the fitted values and

$$w_\tau(y_i, \hat{y}_{i,\tau}) = \begin{cases} \tau & y_i \geq \hat{y}_{i,\tau} \\ (1-\tau) & y_i < \hat{y}_{i,\tau} \end{cases}.$$

Figure 1 shows an example of a QNAM consisting of J feature networks. By simply adjusting the number of output neurons and specifying the loss

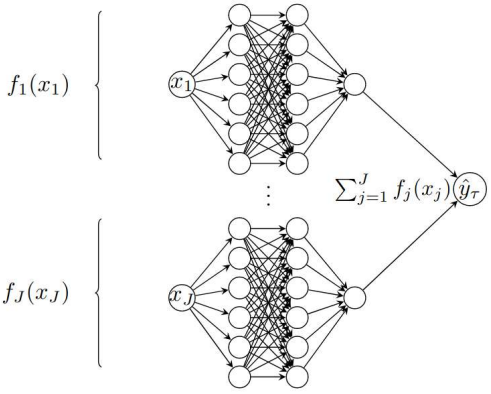


FIGURE 1. Exemplary archicture of a QNAM: The network prediction is the sum of multiple feature networks.

function to be the sum of the individual quantile losses, the QNAM can also be extended to estimate multiple quantiles at once.

3 Simulations

To demonstrate the advantage of our model over other established approaches, we simulate a large dataset with $N = 100,000$ observations and a structural break, where the response y depends on a single covariate x. The upper left panel of Figure 2 shows a QNAM, the upper right panel shows a QGAM estimated using the `R`-package `qgam` (Fasiolo et al., 2021). Both models estimate the 99%-quantile. For the QGAM, we use the default specifications for the non-linear effect, a thin-plate regression spline with rank $k = 10$. The QNAM consists of two hidden layers with 64 and 32 neurons.

Whereas the QGAM smooths over the structural break and appears to fit the data quite badly in the area around it, the QNAM captures it quite well. By increasing the rank of the used smoother, one could improve the ability of the QGAM to capture structural breaks, however, this may lead to very jagged estimated effects and quickly becomes computationally problematic for a dataset of this size.

The lower panel of Figure 2 demonstrates the easy extendability of the QNAM and shows a model which estimates multiple quantiles at once with $\boldsymbol{\tau} = (0.01, 0.2, 0.4, 0.6, 0.8, 0.99)$. While the known issue of quantile crossing does not play a role for a dataset of this size, the simultaneous estimation of quantiles could be used to further extend the model by including measures to prevent quantile crossing for smaller datasets.

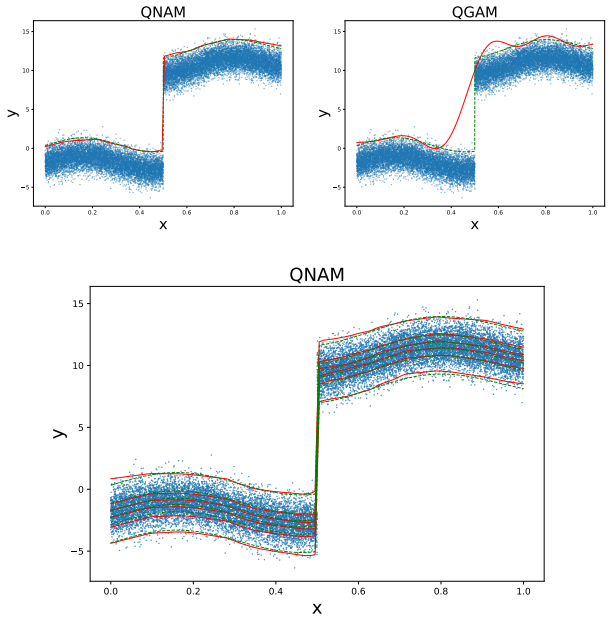


FIGURE 2. Comparison of QNAM and QGAM. The models in the upper two panels estimate the 99%-quantile of the data with the estimated quantiles plotted in red and the true values in green. The lower panel shows a QNAM estimating multiple quantiles at once.

4 Probabilistic Load Forecasting on the Household Level

The deployment of smart meters that monitor the electricity consumption of individual households and the data they collect open up new possibilities for probabilistic electric load forecasting on the household level which can provide valuable insights for power grid operators and other actors in the energy industry. The London Smart Meter dataset, collected by the UK Power Networks[1], available in a preprocessed version combined with weather data on Kaggle[2], contains half-hourly measurements from 5,567 households in the greater London area from between November 2011 and February 2014. The total amount of observations amounts to about 167 million. The size of the dataset necessitates a model that is equipped to handle large amounts of data. Furthermore, the heterogeneity of the data

[1]https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[2]https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london

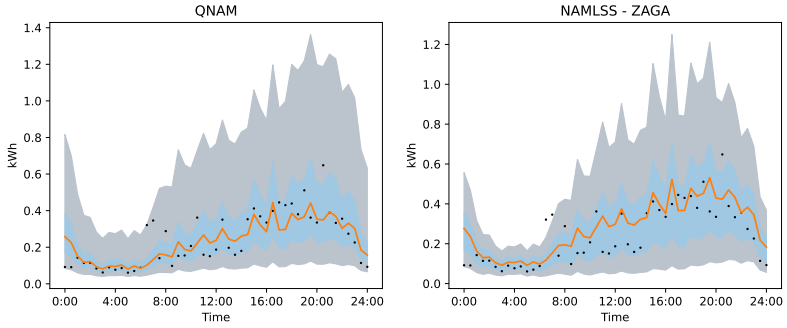


FIGURE 3. Estimated distribution of predicted electricity consumption for one household over the course of one day. Left panel shows a QNAM, right panel a NAMLSS based on the ZAGA distribution. The shaded areas show the 90% and 50% prediction intervals respectively, the true observed values are depicted as black dots. For the QNAM, the orange line is the predicted median, in the NAMLSS it is the predicted mean of the distribution.

requires a model that can estimate its distribution in a flexible manner. We use a QNAM to estimate the conditional distribution of the energy consumption y in time $t + h$ such that

$$F_{t,h}(y \mid \boldsymbol{x}_t) = P(y_{t+h} \leq y \mid \boldsymbol{x}_t)$$

where $\boldsymbol{x}_t$ contains information known at time t, such as past energy consumptions and exogenous covariates, such as time of day, position within a year, weekday and temperature at time $t+h$, as well as a factor variable for the specific household. As in Taieb (2016), we obtain the distribution by estimating multiple quantiles simultaneously. For our model, we estimate 21 quantiles with $\boldsymbol{\tau} = (0.01, 0.05, 0.1, \ldots, 0.95, 0.99)$ and a forecast horizon of $h = 48$.

We compare our model to a NAMLSS (Thielmann et al., 2023), another extension of the NAM which estimates a parametric distribution in a way similar to GAMLSS (Stasinopoulos and Rigby, 2008). For the NAMLSS, we assume a Zero Adjusted Gamma distribution (ZAGA), since it allows for skewness as well as zero values. The NAMLSS is trained by minimizing the negative log-likelihood of the ZAGA distribution. It consists of three sub-models, one for each distribution parameter that define a ZAGA distribution, that are trained simultaneously.

Figure 3 visualizes the results of a model trained on a subset of 500 households and showcases how the proposed model can be used for forecasting the energy consumption of one specific household over the course of a day. Both models appear to predict the conditional distribution similarly and seem to be able to capture household specific electricity consumption profiles adequately.

Acknowledgments: This work was supported by the Freigeist-Fellowships of Volkswagen Stiftung, project "Bayesian Boosting - A new approach to data science, unifying two statistical philosophies".

References

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. (2021). Neural Additive Models: Interpretable Machine Learning with Neural Nets. *Advances in Neural Information Processing Systems*, **34**, 4699 – 4711.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., and Goude, Y. (2021). Fast Calibrated Additive Quantile Regression. *Journal of the American Statistical Associations*, **116(535)**, 1402 – 1412.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, **1(3)**, 297 – 310.

Koenker, R. and Bassett Jr, G. (1978). Regression Quantiles. *Econometrica*, **46(1)**, 33 – 50.

Stasinopoulos, D. M. and Rigby, R. A. (2008). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1 – 46.

Taieb, S. B., Huser, R., Hyndman, R. J., and Genton, M. G. (2016). Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. *IEEE Transactions on Smart Grid*, **46(1)**, 2448 – 2455.

Thielmann, A., Kruse, R. M., Kneib T. and Säfken B. (2023). Neural Additive Models for Location Scale and Shape: A Framework for Interpretable Neural Regression Beyond the Mean. *arXiv preprint arXiv:2301.11862*.

Mixed effects neural networks for longitudinal k-inflated count responses

Nastaran Sharifian[1], Kevin Burke[1]

[1] University of Limerick, Ireland

E-mail for correspondence: nastaran.sharifian@ul.ie

Abstract: In some real problems, the count response is inflated in a particular value k and a k-inflated power series (SIPS) distribution is used as its distribution. Here, a generalized neural network mixed mixture (GNNMM) model is applied for predicting outcomes as nonlinear functions of predictors in longitudinal k-inflated power series data.

Keywords: Neural networks; k-inflated; Longitudinal data; Mixture model

1 Introduction

Finite mixture models are commonly used to handle count data with inflated zeros. However, more generally, some datasets have inflated counts at a value k, which is not necessarily zero. Such k inflation can arise in questionnaire data due to the nature of the question and/or responses, for example, Arora (2018) found inflation at $k = 6$ where women were asked about the number of smear tests taken in the past six years (where $k = 6$) corresponds to once per year; Figure 1 displays count data with inflation at $k = 6$.
Generalized Linear Mixed Models (GLMMs) provide a flexible framework for modeling longitudinal data but can have poor predictive power when covariate effects are non-linear, and, hence, Mandel et al (2021) proposed a mixed effects neural network.
The aim of this paper is to extend Mandel's work to model the k-inflated count longitudinal responses.

2 k-Inflated Power Series Distribution

The power series (PS) class of discrete distributions, which includes the Poisson, binomial, and negative binomial, can be described

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

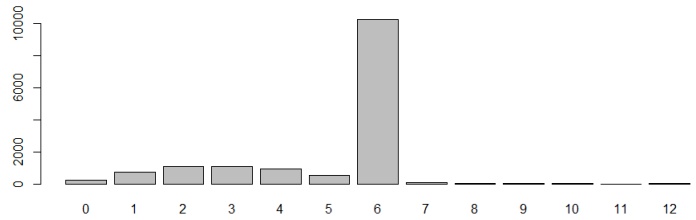


FIGURE 1. The bar charts of the simulated k-inflaed poisson data, $k = 6$.

by the probability mass function

$$f_{PS}(n|\nu) = p(N = n|\nu) = \frac{a(n)\nu^n}{g(\nu)}, \quad n = 0, 1, 2, ...,$$

where $\nu > 0$, $a(n) \geq 0$ and $g(\nu) = \sum_{n=0}^{\infty} a(n)\nu^n$ is the normalizing constant. This can be extended to a k-inflated power series (kIPS) model for a longitudinal count variable N_{it} via

$$p(N_{it} = n_{it}|\pi_{it}, \nu_{it}) = \begin{cases} \pi_{it} + (1 - \pi_{it})\dfrac{a(k)\nu_{it}^k}{g(\nu_{it})} & ; n_{it} = k, \\ (1 - \pi_{it})\dfrac{a(n_{it})\nu_{it}^{n_{it}}}{g(\nu_{it})} & ; n_{it} \neq k, \end{cases} \tag{1}$$

where π_{it}, $0 \leq \pi_{it} \leq 1$, is a mixing proportion for the value k, and $t = 1, \ldots, T_i$ is the time index for the ith individual.

3 GLMM Model

For longitudinal count response N_{it}, let $\boldsymbol{X}_{it}$ and $\boldsymbol{Z}_{it}$ be the $p \times 1$ and $r \times 1$, known and fixed vectors of covariates, for the ith subject at time t, respectively. Also, $\boldsymbol{W}_{it}$ is some $s \times 1$ sub-vector of $\boldsymbol{X}_{it}$. Then, the kIPS GLMM for longitudinal k-inflated count data is given by

$$\begin{aligned} &N_{it}|\boldsymbol{b}_i, \nu_{it}, \pi_{it} \sim kIPS(\nu_{it}, \pi_{it}), \\ &q_1(\nu_{it}^{\boldsymbol{b}}) = \boldsymbol{X}_{it}'\boldsymbol{\gamma}_1 + \boldsymbol{W}_{it}'\boldsymbol{b}_i, \\ &q_2(\pi_{it}) = \boldsymbol{Z}_{it}'\boldsymbol{\gamma}_2, \end{aligned} \tag{2}$$

where q_1 and q_2 are some known link functions for ν_{it} and π_{it}, respectively (Sharifian et al., 2021). Typically, q_1 will be taken to be the log link function for ν_{it}, and q_2 will be taken to be the logit link function for π_{it}. The vector of random effects, $\boldsymbol{b}_i$ is used to include within-subject dependence through time and is typically assumed to follow a multi-variate normal distribution.

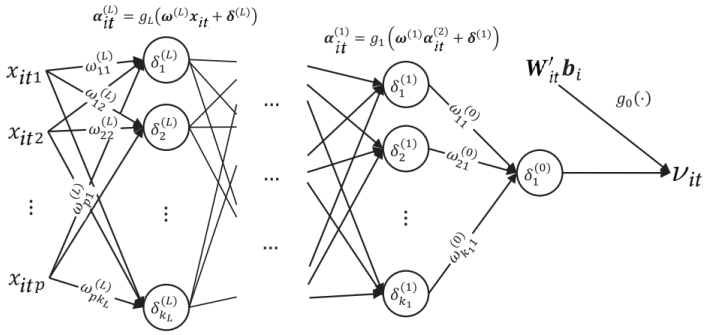


FIGURE 2. The architecture of a multi-layer generalized neural network mixed model (Mandel et al., 2021)

3.1 Parameter Estimation

To obtain the MLEs for the parameters of the mixture model (2) it is reasonable to apply the EM algorithm. Let $\boldsymbol{n}_i = (n_{i1}, ..., n_{iT_i})'$, $\boldsymbol{n} = (\boldsymbol{n}_1^{'}, ..., \boldsymbol{n}_m^{'})'$, $\boldsymbol{u}_i = (u_{i1}, ..., u_{iT_i})'$, and $\boldsymbol{u} = (\boldsymbol{u}_1^{'}, ..., \boldsymbol{u}_m^{'})'$. Then the complete data for the EM algorithm are $(\boldsymbol{n}, \boldsymbol{u}, \boldsymbol{b})$. Now, we characterize the model by the latent variable U_{it} where $Pr(U_{it} = 1) = \pi_{it}$. If it is supposed that $N_{it}|(U_{it} = 1)$ and $N_{it}|(U_{it} = 0)$ have the mass function degenerated at the value of k and the mass function of a power series distribution, the joint density of (N_{it}, U_{it}) is given by $f(n_{it}, u_{it}|\boldsymbol{b}_i) = [\pi_{it} I_{\{k\}}(n_{it})]^{u_{it}} [(1 - \pi_{it}) f_{PS}(n_{it}|\nu_{it})]^{1-u_{it}}$. Based on the complete data $(\boldsymbol{n}, \boldsymbol{u}, \boldsymbol{b})$, the kernel log-likelihood, $\log f(\boldsymbol{n}, \boldsymbol{u}|\boldsymbol{b})$, is obtained as follows:

$$\log f(\boldsymbol{n}, \boldsymbol{u}|\boldsymbol{b}) = \sum_{i=1}^{m} \sum_{t=1}^{T_i} \log f(n_{it}, u_{it}|\boldsymbol{b}) = \sum_{i=1}^{m} \sum_{t=1}^{T_i} \Big[u_{it} log \pi_{it} \Big]$$
$$+ \sum_{i=1}^{m} \sum_{t=1}^{T_i} \Big[(1 - u_{it}) \big(\log(1 - \pi_{it}) + \log f_{PS}(n_{it}|\nu_{it}) \big) \Big].$$

Because this log-likelihood is linear in u_{it}, it is straightforward to compute its expected value (E Step), which is then followed by maximisation (M Step); we iterate between these two steps until convergence.

4 Neural network extension

Consider a feed-forward ANN with L hidden layers, $\boldsymbol{X}_{it}$ as the p inputs, and an univariate output $\nu_{it}^{\boldsymbol{b}}$. Figure 2 shows the architecture of our proposed neural network extension model. The neural network's output $\nu_{it}^{\boldsymbol{b}}$ can be

written as a nonlinear function of the predictors $\boldsymbol{X}_{it}$ as well as the network weights $\boldsymbol{\omega}^{(l)}$ and biases $\boldsymbol{\delta}^{(l)}$ through a series of nested activation functions $g_l(.)$ for layers $l = 0, 1, 2, ..., L$. The $\boldsymbol{X}_{it}$ enters into the neural network through the Lth hidden layer consisting of k_L nodes, producing

$$\boldsymbol{\alpha}_{it}^{(L)} = g_L\Big[\boldsymbol{\omega}^L \boldsymbol{X}_{it} + \boldsymbol{\delta}^{(L)}\Big] \tag{3}$$

where $\boldsymbol{\omega}^{(L)}$ is a $K_L \times p$ weight matrix, $\boldsymbol{\delta}^{(L)}$ is a bias vector of length k_L, and $g_L(.)$ is the activation function applied to its input vector. For the lth hidden layer ($l = 1, ..., L-1$) with k_l nodes, the layer's output is

$$\boldsymbol{\alpha}_{it}^{(l)} = g_l[\Big[\boldsymbol{\omega}^l \boldsymbol{\alpha}_{it}^{(l+1)} + \boldsymbol{\delta}^{(l)}\Big], \tag{4}$$

where $\boldsymbol{\omega}^{(l)}$ is a $K_l \times k_{l+1}$ matrix and $\boldsymbol{\delta}^{(l)}$ is a vector of length k_l. The output from the neural network, the parameter of the power series distribution, is as follows:

$$\boldsymbol{\nu}_{it}^{\boldsymbol{b}} = g_0\Big[\boldsymbol{\omega}^0 \boldsymbol{\alpha}_{it}^{(1)} + \delta^{(0)} + \boldsymbol{W}_{it}^{'} \boldsymbol{b}_i\Big], \tag{5}$$

where $\boldsymbol{\omega}^{(0)}$ is $1 \times K_l$. The linear predictor for the random effects $W_{it}^{'}\boldsymbol{b}_i$ is included in the final layer of the network. For estimation, we apply the previously described EM algorithm but with a neural network in place of the linear predictor.

5 Discussion

The proposed GNNMM is highly flexible as it can handle non-linear, longitudinal k-inflated data. We will demonstrate the utility of the GNNMM compared to existing approaches on both real and simulated data in our presentation.

Acknowledgments: This work was supported by the Confirm Smart Manufacturing Centre (https://confirm.ie/) funded by Science Foundation Ireland (Grant Number: 16/RC/3918).

References

Arora, M (2018). *Extended Poisson Models for Count Data With Inflated Frequencies.* Doctor of Philosophy (PhD), Dissertation, Mathematics and Statistics, Old Dominion University, DOI: 10.25777/nz1ed763.

Mandel, F., Ghosh, R. P., and Barnett, I. (2021). Neural networks for clustered and longitudinal data using mixed effects models, *BIOMETRIC METHODOLOGY*, DOI: 10.1111/biom.13615

Sharifian N, Bahrami Samani E, Ganjali M (2021). Joint modeling for longitudinal set-inflated continuous and count responses *Commun Stat-Theory Methods*, **50** (5), 1134 – 1160.

A flexible non-mixture cure model for recurrent gap time data

Ivo Sousa-Ferreira[1,2,3], Ana Maria Abreu[2,4], Cristina Rocha[1,3]

[1] Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal
[3] CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal
[4] CIMA – Centro de Investigação em Matemática e Aplicações, Portugal

E-mail for correspondence: ivo.ferreira@staff.uma.pt

Abstract: A new cure model for gap times between recurrent events is proposed. The model is characterized by a fully parametric rate function derived from a non-homogeneous Poisson process (NHPP). To obtain flexible shapes of the rate function, the baseline log-cumulative rate function is modelled as a restricted cubic spline (RCS) function of log time. Furthermore, a shared frailty is included in order to develop a survival model for heterogeneity that accounts for zero-recurrence subjects. With this purpose, we assume that the frailty has a non-central chi-squared distribution with zero degrees of freedom (d.f.), which gives rise to a non-mixture cure model. The model also includes covariates, acting multiplicatively on the rate function. The usefulness of the model is emphasized through its application to hospital readmission data.

Keywords: Gap times; Non-mixture cure model, Non-homogeneous Poisson process; Restricted cubic splines; Shared frailty.

1 Introduction

Recurrent gap time data often arise in biomedical research when it is intended to study the time between consecutive events (Cook and Lawless, 2007). However, scientific advances have led to an increase in the number of subjects that will never experience a recurrence, who are designated as zero-recurrence subjects. Zhao and Zhou (2012) developed a semiparametric mixture cure model considering that the recurrence process is derived

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

from a NHPP. Under this approach, Sousa-Ferreira et al. (2020) specified a flexible parametric form for the baseline rate function based on a RCS. Nevertheless, the within-subject correlation problem triggered by the unobserved heterogeneity, which can lead to biased estimators, has not yet been addressed.

2 The flexible non-mixture cure model

Suppose there are n independent subjects in study and each one can experience a maximum of K_i $(i = 1, \ldots, n)$ recurrences of an event. For the ith subject, let T_{ik} be the time of the kth event $(k = 1, \ldots, K_i)$, $Y_{ik} = T_{ik} - T_{i,k-1}$ the gap time and W_i a non-negative random variable. Following the approach of Zhao and Zhou (2012), the recurrence process is assumed to be a NHPP. Then, we consider a multiplicative model in which, conditional on the shared frailty $W_i = w_i$, the cumulative rate function (crf) of the kth gap time is given by

$$H(y|t_{i,k-1}, w_i, \boldsymbol{z}_{ik}) = w_i H(y|t_{i,k-1}) \exp(\boldsymbol{\beta}' \boldsymbol{z}_{ik}), \tag{1}$$

where $H(y|t_{i,k-1}) = H_0(y + t_{i,k-1}) - H_0(t_{i,k-1})$ is a baseline crf, $\boldsymbol{z}_{ik}$ is the covariate vector and $\boldsymbol{\beta}$ is the regression coefficients vector. As considered in Sousa-Ferreira et al. (2020), we propose to model the $\log H_0(\cdot)$ as a RCS function of log time. The complexity of the curve is regulated by the number of d.f., given by d.f. $= m + 1$, where m is the number of internal knots of the RCS. By convention, when d.f. $= 1$ the baseline rate function has a Weibull hazard form.

In certain settings, the unobserved heterogeneity may be originated by the existence of some non-susceptible subjects, while the others have a varying degree of susceptibility. Motivated by Rocha (1996), we accommodate this situation considering that the frailty has a non-central chi-squared distribution with zero d.f. and non-centrality parameter $\gamma > 0$, denoted by $\chi_0^{'2}(\gamma)$. This distribution can be obtained as a Poisson mixture of central chi-squared distributions with even d.f. Some properties of $\chi_0^{'2}(\gamma)$ are described in Rocha (1996). In particular, its Laplace transform (LT) is $L_W(s) = \pi^{1-(1+2s)^{-1}}$, with $\pi = \exp(-\gamma/2)$. Note that $\chi_0^{'2}(\gamma)$ has a mass at zero given by $P\{W = 0\} = \pi$ $(0 < \pi < 1)$, following that $L_W(\infty) = \pi$. Then, the unconditional (population) crf of the flexible non-mixture cure model is obtained by taking the expectation of (1) over W_i, yielding

$$H_{\text{pop}}(y|t_{i,k-1}, \boldsymbol{z}_{ik}) = -\log\left\{\pi^{1-\left[1+2H(y|t_{i,k-1})\exp(\boldsymbol{\beta}'\boldsymbol{z}_{ik})\right]^{-1}}\right\}. \tag{2}$$

So, two situations can occur: the ith subject experiences at least one recurrence, being a recurrent subject with $P\{W > 0\} = 1 - \pi$; or the ith subject does not suffer any event, being either a recurrent subject with $P\{W > 0\} = 1 - \pi$ or a zero-recurrence subject with $P\{W = 0\} = \pi$.

The inferential procedure is based on the maximum likelihood method, under a non-informative right-censoring mechanism and assuming that the frailties $W_1, \ldots, W_n$ are independent and identically distributed random variables. Since the proposed model is fully specified, the parameter estimation is based on the unconditional likelihood function, which can be expressed as

$$\mathcal{L} = \prod_{i=1}^{n} (-1)^{d_i} \left\{ \prod_{k=1}^{K_i} h(y_{ik}|t_{i,k-1}, \boldsymbol{z}_{ik})^{\delta_{ik}} \right\} L_W^{(d_i)} \left[\sum_{k=1}^{K_i} H(y_{ik}|t_{i,k-1}, \boldsymbol{z}_{ik}) \right],$$

where $h(y|t_{i,k-1}, \boldsymbol{z}_{ik}) = dH(y|t_{i,k-1}, \boldsymbol{z}_{ik})/dy$, $d_i = \sum_{k=1}^{K_i} \delta_{ik}$, δ_{ik} is the usual right-censoring indicator and $L_W^{(d)}(\cdot)$ is the dth derivative of the LT of W_i. The computational implementation was developed in `R` software (R Core Team, 2023), using the Broyden–Fletcher–Goldfarb–Shanno method.

3 Application to hospital readmission data

The analysed data represent the gap times (in days) of successive hospital readmissions of 403 patients diagnosed with colorectal cancer, after surgery to remove their tumours. The maximum follow-up time was 2176 days ($\approx$ 6 years) and a total of 861 readmissions were recorded, with 199 patients (49.4%) having no recurrence at all. The data are available in the `R` library `frailtypack` and contain the following covariates: chemotherapy (0: untreated, 1: treated); gender (0: male, 1: female); Dukes' stage (1: stage A−B, 2: stage C, 3: stage D); and Charlson comorbidity index (0: index 0, 1: index $1-2$, 3: index ≥ 3).

The Akaike information criterion (AIC) was used to informally select the number of d.f. of the RCS. Thus, models with 1 to 4 d.f. were fitted, without including covariates or the proportion of zero-recurrence subjects. The AIC values (7067.6, 7046.0, 7047.7 and 7048.5) indicate that the most adequate

TABLE 1. Results obtained from fitting the flexible non-mixture cure model to hospital readmission data.

Parameters		Estimate	$\widehat{\text{SE}}$	p-value of Wald test
Spline part	ξ_0	−8.343	0.605	—
	ξ_1	1.207	0.154	—
	ξ_2	0.011	0.004	—
Chemo [ref. untreated]		−0.189	0.159	0.234
Gender [ref. male]		−0.643	0.152	2.170e-05
Dukes' stage [ref. A−B]				
C		0.294	0.180	0.102
D		1.188	0.216	3.908e-08
Charlson index [ref. 0]				
$1-2$		0.543	0.294	0.065
≥ 3		0.676	0.151	7.517e-06
π		0.140	0.035	—

number is d.f. = 2. Then, the regression model characterized by (2) was implemented and the results are shown in Table 1.

In our model, the reference group consists of untreated male patients, who are in Dukes' stage A−B, have Charlson index 0 and frailty $W = 1$. For this group, the zero-recurrence proportion estimate is $\hat{\pi} = 0.140$. The chemotherapy coefficient estimate is negative, with a non-significant effect on the gap time between readmissions, suggesting that the treatment reduces the rate of readmission but has a negligible effect. Recurrent female patients have significantly longer gap times to hospital readmissions compared to the recurrent males. The other two important prognostic factors are the Dukes' stage D and Charlson index ≥ 3, both yielding a significant increasing effect on the rate of readmission. Notice that the regression coefficients hold a cluster-level relative risk interpretation, referring to comparisons between subjects that share the same values of frailty and remaining observed covariates.

The model-based estimates of the conditional and unconditional rate functions, for subjects with null covariate vector, are depicted in Figure 1. All estimates exhibit a right-skewed unimodal shape. When $W = 1$ (reference group), the readmission rate reaches its maximum value 24 days after the previous recurrence. However, for the population, the maximum value is attained after 18 days.

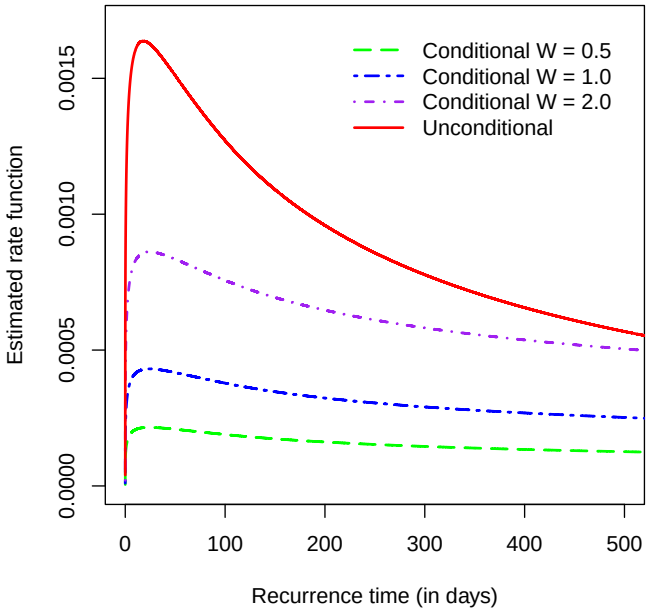


FIGURE 1. Estimated conditional and unconditional rate functions, based on the flexible non-mixture cure model with d.f. = 2 for the spline part.

4 Final remarks

In this paper, the approach of Zhao and Zhou (2012) is extended to include a shared frailty. In particular, we propose that the frailty follows a $\chi_0^{'2}(\gamma)$ distribution, which has a probability mass at zero and is continuously distributed on the positive real line. Therefore, it allows to account simultaneously for the within-subject dependence among recurrent gap times and the existence of zero-recurrence subjects in the population.
For future research, it would be interesting to conduct a simulation study to evaluate the performance of the inferential procedure in several scenarios.

Acknowledgments: This research was partially sponsored by portuguese funds through *FCT – Fundação para a Ciência e a Tecnologia*, under the projects UIDB/00006/2020 (*Centro de Estatística e Aplicações*) and UIDB/04674/2020 (CIMA – Center for Research in Mathematics and Applications, from the Statistics, Stochastic Processes and Applications group). I. Sousa-Ferreira is grateful to FCT for his PhD grant DFA/BD/6459/2020.

References

Cook, R.J., and Lawless, J. (2007). *The Statistical Analysis of Recurrent Events*. Springer Science & Business Media. ISBN: 978-1-4419-2415-5.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

Rocha, C.S. (1996). Survival models for heterogeneity using the non-central chi-squared distribution with zero degrees of freedom. In: Jewell, N.P., Kimber, A.C., Lee, M.-L.T., and Whitmore, G.A. (eds.), *Lifetime Data: Models in Reliability and Survival Analysis*, Springer, Boston, MA, 275 – 279.

Sousa-Ferreira, I., Rocha, C., and Abreu, A.M. (2020). A flexible marginal rate model for recurrent events with a zero-recurrence proportion. In: Irigoien, I., Lee, D.-J., Martínez-Minaya, J., and Rodríguez-Álvarez, M.X. (eds.), *Proceedings of the 35th International Workshop on Statistical Modelling*, Bilbao, Spain, 417 – 420.

Zhao, X., and Zhou, X. (2012). Modeling gap times between recurrent events by marginal rate function. *Computational Statistics & Data Analysis*, **56**(2), 370 – 383.

A tool to dectect nonlinearity and interactions in generalized regression models

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: In generalized regression models the effect of continuous covariates is commonly assumed to be linear. This assumption, however, may be too restrictive in applications and may lead to biased effect estimates. While a multitude of alternatives for the flexible modeling of continuous covariates have been proposed, methods that provide guidance for choosing a suitable functional form are still limited. To address this issue, we propose a detection algorithm that evaluates several approaches for modeling continuous covariates and guides practitioners to choose the most appropriate alternative. The performance of the algorithm was assessed in a simulation study. To illustrate the proposed algorithm, we analyzed data of patients suffering from chronic kidney disease.

Keywords: Nonlinearity; Interactions; Functional Forms; Tree-based Modeling

1 Introduction

Generalized linear models (GLMs) are one of the most popular tools for regression analysis. In GLMs the outcome of interest is related to a set of covariates using a linear combination of the covariate values (i.e., continuous covariates are fitted by simple linear terms). Although this linear modeling approach is often considered the default and rarely questioned in practice, assuming linearity may often be too restrictive, and misspecifying the functional form of a continuous covariate may lead to biased effect estimates. There exist a number of established alternatives for the flexible modeling of continuous covariates – among others, categorization, structural breaks, polynomial regression, generalized additve models (GAMs) and classification and regression trees (CART) – that go beyond classical GLMs. However, because each method exhibits specific benefits and draw-

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

backs, the choice of the most appropriate one remains highly challenging and its importance is frequently neglected (Sauerbrei et al., 2020).
To address this issue, we propose an algorithm that examines various modeling alternatives and is able to detect nonlinearity and interactions between covariates, if they are present. The two-step algorithm (described in Section 3) utilizes tree-based splits which makes the resulting effects easily interpretable. More specifically, it indicates whether (i) linear effects are sufficient, (ii) varying linear effects should be included in the model formula, (iii) one or several covariates exhibit non-linear effects or (iv) interaction effects occur in the data.

2 A group of generalized regression models

We consider generalized regression models, where the expectation of an outcome $y_i,\ i = 1, \ldots, n$, is linked to a vector of p covariates $\boldsymbol{x_i} = (x_{i1}, ..., x_{ip})^\top$ in the form $\mathbb{E}(y_i|\ \boldsymbol{x_i}) = g^{-1}(\eta(\boldsymbol{x_i}))$, where $g(\cdot)$ denotes a suitable link function and $\eta(\cdot)$ denotes the predictor function. Assuming that the effect of a covariate x_j on the outcome is simply *linear* yields the model with predictor function

$$\eta(\boldsymbol{x_i}) = \beta_0 + \beta_j\, x_{ij}\,, \tag{1}$$

where β_0 is the intercept and β_j is the linear regression coefficient. Alternatively, one can consider a predictor function with a *piecewise constant* effect of the form

$$\eta(\boldsymbol{x_i}) = \beta_0 + \gamma_j\, I(x_{ij} > c_j)\,, \tag{2}$$

where $I(\cdot)$ denotes the indicator function, c_j is a split point in x_j and γ_j is the corresponding regression coefficient. A more complex model using an *additive combination* of the linear and piecewise constant effect yields the predictor function

$$\eta(\boldsymbol{x_i}) = \beta_0 + \beta_j\, x_{ij} + \gamma_j\, I(x_{ij} > c_j)\,. \tag{3}$$

Note that both, the linear and the piecewise constant model, are nested in model (3). When using a *multiplicative combination* of the linear and piecewise constant effect, the predictor function is given by

$$\eta(\boldsymbol{x_i}) = \beta_0 + \beta_{j1}\, x_{ij} + \beta_{j2}\, I(x_{ij} > c_j)\, x_{ij}\,, \tag{4}$$

The linear model is nested in in model (4), as setting $\beta_{j2} = 0$ yields (1). Finally, we consider an extension of (2) allowing for an *additional split* in x_j, which has the form

$$\eta(\boldsymbol{x_i}) = \begin{cases} \beta_0 + \gamma_{jr}\, I(x_{ij} > c_j) \\ \quad + \gamma_{j\ell}\, I(x_{ij} \leq c_j \wedge x_{ij} > c_{j\ell})\,, & \text{if split in } \{x_{ij} \leq c_j\}\,, \\ \beta_0 + \gamma_{j\ell}\, I(x_{ij} \leq c_j) \\ \quad + \gamma_{jr}\, I(x_{ij} > c_j \wedge x_{ij} > c_{jr})\,, & \text{if split in } \{x_{ij} > c_j\}\,. \end{cases} \tag{5}$$

Importantly, models (1) to (5) form a group of nested models that can all be fitted using the framework of tree-structured varying coefficient (TSVC) models (Berger et al., 2019). In the presence of multiple continuous covariates $x_1, \ldots, x_p$ each part of the predictor function can take the form as given by (1) to (5). Furthermore, models (4) and (5) allow for an interaction between two covariates.

3 Algorithm

We propose a two-step algorithm that examines the modeling alternatives introduced in the previous section and automatically chooses the most appropriate one according to their predictive performance. More specifically, we compute the predicted log-likelihood of the models using leave-one-out cross validation (LOOCV). In addition, we apply the so-called "one standard error rule" (1SE rule), which is an established strategy for the selection of tuning parameters in regularized regression (Chen and Yang, 2021).
Let us again consider one continuous covariate x_j. In the first step of the algorithm, the linear model (1) and the piecewise constant model (2) are evaluated. Among these two models, the model with the larger predictive log-likelihood is selected and compared to the null model (with intercept β_0 only). If the condition of the 1SE rule is met, the selected model is confirmed and the algorithm continues with step 2. Otherwise, no effect of x_j is found and the algorithm is terminated.
In the second step of the algorithm, if a linear effect was selected in step 1, the models (3) and (4) are evaluated. Otherwise, if a piecewise constant effect was selected in step 1, the models (3) and (5) are evaluated. In the same way as in step 1, the algorithm first computes the predictive log-likelihood values using LOOCV. Afterwards, it compares the better performing model to the simpler one applying the 1SE rule. In a scenario with multiple continuous covariates the models are adjusted for all effects of the other covariates selected in step 1.

4 Empirical evaluations

To illustrate the proposed algorithm, we consider 100 simulated data sets with five covariates $x_1, \ldots, x_5 \sim N(0, 1)$ and a data-generating model of the form

$$\begin{aligned} y_i = &\beta_0 + \beta_{11} x_{i1} + \beta_{12} I(x_{i2} > 0) x_{i1} + \gamma_{3\ell} I(x_{i3} \leq 0) + \\ &\gamma_{3r} I(x_{i3} > 0 \wedge x_{i4} > 0) + \varepsilon_i \,, \quad i = 1, \ldots, 500 \,, \end{aligned} \tag{6}$$

where $\beta_0 = 1$, $\beta_{11} = 0.6$, $\beta_{12} = 1.2$, $\gamma_{3\ell} = -1$, $\gamma_{3r} = 2$. The independent error terms ε_i were drawn from a zero-mean normal distribution with standard deviation $\sigma \in \{1, 1.5, 2\}$. The results in Table 1 show that the

TABLE 1. Results of the simulation. Proportion of simulation runs in which the effects were identified correctly.

True effect	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 2$
Type (4) linear effect of x_1 modified by x_2	1.00	0.89	0.50
Type (5) interaction of x_3 and x_4	1.00	0.95	0.78
No effect of x_5	1.00	1.00	1.00

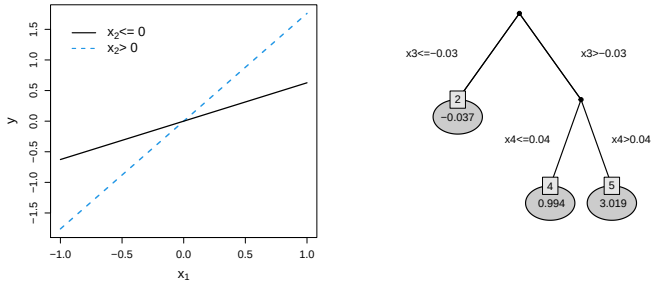


FIGURE 1. Results of the simulation. Estimated effects of x_1 and x_2 (left) and x_3 and x_4 (right) using a data set, where all effects were correctly identified.

tree-structured interaction (5) between x_3 and x_4 was more likely to be identified than the varying effect (4) of x_1 with regard to x_2 (particularly in the scenario with large noise). The absence of the effect of x_5 was perfectly detected illustrating the conservative impact of the 1SE rule. Figure 1 shows the estimated effects of one exemplary data set, where all effects were correctly identified by the proposed algorithm.

In addition, we applied the algorithm to real-word data from chronic kidney disease patients, where the objective was to to identify suitable functional forms for the effect of BMI and the biomarker HbA(1c) on the probability of suffering from diabetic nephropathy. The results indicate that BMI exhibits a linear effect, whereas a piecewise constant effect with split point 49.3 mmol/mol is most suitable for HbA(1c).

References

Berger, M., Tutz, G., and Schmid, M. (2019). Tree-structured modelling of varying coefficients. *Statistics and Computing*, **29**, 229 – 2019.

Chen, Y. and Yang, Y. (2012). The one standard error rule for model selection: does it work? *Stats*, **4**, 868 – 892.

Sauerbrei, W., Perperoglou, A., Schmid, M., et al. (2020). State of the are in selection of variable and function forms in multivariable analysis – outstanding issues. *Diagnostic and Prognostic Research*, **4**, 619 – 678.

Variable selection for statistical fine-mapping and prediction modelling of polygenic traits

Christian Staerk[1], Carlo Maj[2], Oleg Borisov[3], Hannah Klinkhammer[1,3], Peter Krawitz[3], Andreas Mayr[1]

[1] Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Bonn, Germany
[2] Center for Human Genetics, University of Marburg, Marburg, Germany
[3] Institute for Genomic Statistics and Bioinformatics, Medical Faculty, University of Bonn, Bonn, Germany

E-mail for correspondence: christian.staerk@imbie.uni-bonn.de

Abstract: Classical variable selection approaches for large-scale genotype data are typically based on marginal associations of genetic variants with the phenotype of interest. In this work we consider more advanced variable selection methods based on multivariable regression models for statistical fine-mapping of variants in relevant genomic regions and combine the selected variants in prediction models. We illustrate our approach on large UK Biobank genotype data for the prediction of height as a polygenic trait. Based on our results we discuss the interplay between model sparsity, predictive performance and generalizability.

Keywords: Genetic epidemiology; Prediction modelling; Stochastic search; UK Biobank; Variable selection.

1 Introduction

In genetic epidemiology one is often confronted with large-scale data, where both the sample size n and the number of variables p are very large. Due to computational and memory issues, classical variable selection methods for genotype data are typically based on univariate (marginal) associations of the genetic variants with the phenotype of interest, which are conveniently available as summary statistics from genome-wide association studies (GWAS). However, from a statistical modelling perspective it is desirable to select informative variants not only based on marginal associations but by using multivariable regression methods, which can be directly applied to the individual-level genotype data.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

In this work we investigate different modern variable selection methods for statistical fine-mapping of informative variants in genomic regions. In particular, we consider the Adaptive Subspace (AdaSub) method (Staerk et al., 2021) for ℓ_0-type selection criteria, as well as Probing (Thomas et al., 2017) for statistical boosting. To facilitate computations, we apply the variable selection methods on separate genomic regions and combine the selected variants from the individual regions in final prediction models. Using large-scale data from the UK Biobank we illustrate our approach for the prediction of height as a polygenic trait, which is influenced by several common genetic variants.

2 Variable selection for statistical fine-mapping

To apply state-of-the-art variable selection techniques on large-scale genetic data we consider a three-step approach (cf., Maj et al., 2022). In a first step, we divide the genome into smaller regions, where relevant regions are determined by a marginal screening approach (at least one genetic variant associated with the phenotype with $p < 5 \times 10^{-8}$) and are pre-filtered for suggestively significant variants ($p < 10^{-5}$). In a second step, we apply variable selection based on multivariable linear regression methods to fine-map the signal in the relevant genomic regions. Note that variants in the same region tend to be highly correlated due to linkage disequilibrium (LD), so that it is important to investigate their effects in a multivariable model. In a third step, we combine the selected variants from the different regions into a final multivariable polygenic prediction model using statistical boosting. We consider two different variable selection methods for fine-mapping. The Adaptive Subspace (AdaSub) method (Staerk et al., 2021) conducts a stochastic search to address the discrete optimization problem of identifying the best model according to an ℓ_0-type selection criterion. AdaSub solves multiple low-dimensional sub-problems of the original high-dimensional problem in an adaptive way, where the probability of each variant to be included in a new sub-problem is sequentially adjusted based on its selection frequency in previous sub-problems. We apply AdaSub to minimize the extended Bayesian information criterion (EBIC), defined by

$$\mathrm{EBIC}_\gamma(S) = n \log\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_S + \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_S - y_i\right)^2\right) + \left(\log(n) + 2\gamma\log(p)\right)|S|,$$

for a set of variants $S \subseteq \{1, \ldots, p\}$. Here, $\boldsymbol{x}_i \in \{0,1,2\}^p$ denotes the genotype and $y_i \in \mathbb{R}$ the phenotype for subjects $i = 1, \ldots, n$, while $\hat{\mu}_S$ is the estimated intercept and $\hat{\boldsymbol{\beta}}_S \in \mathbb{R}^p$ is the least squares estimate for the linear model with variants in S (i.e. $\hat{\boldsymbol{\beta}}_{S,j} = 0$ for $j \notin S$). The parameter $\gamma \in [0,1]$ in EBIC_γ controls the induced sparsity.

As a second variable selection approach we consider Probing (Thomas et al., 2017) for statistical boosting with linear component-wise base-learners

and squared error loss. Statistical boosting constructs an adaptive ensemble of simple linear models by sequentially fitting the current model residuals using the best performing base-learner. In Probing, for each variant X_j an additional base-learner is incorporated based on a "shadow variant" (probe) $\tilde{X}_j$, which is a randomly permuted version of the original X_j and therefore not associated with the phenotype. The boosting algorithm is stopped as soon as the first base-learner for one of the probes $\tilde{X}_1, \ldots, \tilde{X}_p$ is selected, encouraging the sparsity of the resulting model.

3 UK Biobank data: The prediction of height

We illustrate the fine-mapping methods on UK Biobank data to identify informative variants and build prediction models for height. In particular, we analyse imputed genotype counts for $p = 9{,}812{,}717$ variants (single-nucleotide polymorphisms, SNPs), considering $n_{\text{train}} = 272{,}726$ samples as training data for fine-mapping and $n_{\text{test}} = 135{,}291$ samples as test data for evaluating the prediction accuracy (all individuals of British ancestry).

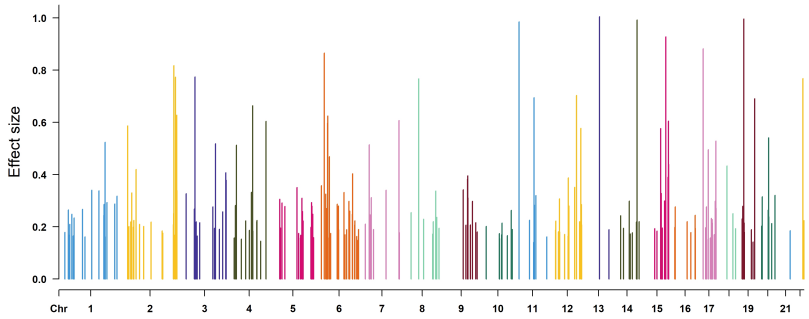


FIGURE 1. Effect sizes (absolute regression coefficients) of final boosted model for height based on fine-mapping with AdaSub for EBIC_1. The x-axis reflects the position of variants in the genome (chromosomes indicated by different colours).

Figure 1 illustrates the effect sizes of the selected variants in the final prediction model for height based on fine-mapping with AdaSub for EBIC_1. While the distribution of effect sizes across the genome reflects the general polygenicity of height, it is also apparent that the final model is relatively sparse with some pronounced estimated effects for single variants.
Table 1 shows that the final model by AdaSub for EBIC_1 indeed only includes 292 variants. As expected, the final model by AdaSub for $\text{EBIC}_{0.5}$ is less sparse including 594 variants, while Probing results in the largest model with 2,788 variants (which is still relatively sparse, e.g., compared to a BayesR model with 546,011 variants, Klinkhammer et al., 2023). Despite

TABLE 1. Results of fine-mapping for the prediction of height, in terms of number of selected variants (N variants), R^2 on training data, squared correlation r^2 between predicted and observed heights on test data, and root mean squared error (RMSE) of prediction on test data (unit: centimeter). All models include the covariates age, sex and the first ten principal components of the genetic data.

Method	N variants	Training R^2	Test r^2 (% train.)	Test RMSE
AdaSub EBIC_1	292	0.6073	0.6030 (99.3%)	5.835 (cm)
AdaSub $\text{EBIC}_{0.5}$	594	0.6292	0.6201 (98.6%)	5.708 (cm)
Probing	2,788	0.6581	0.6389 (97.1%)	5.567 (cm)

the sparsity of the models, all methods yield reasonable prediction performance with $r^2 > 0.6$ on test data (cf. Klinkhammer et al., 2023). However, there is a clear trend that sparser and hence more interpretable models tend to come at the prize of a reduced prediction accuracy, which might be explained by the high polygenicity of height. On the other hand, sparser models tend to yield a more stable predictive performance (i.e. less overoptimistic results on training set compared to test set). This may also hint at potential benefits of sparser fine-mapped models regarding the generalizability to different populations (cf. May et al., 2022). Further research is warranted to investigate the interplay between invariant predictions, sparsity and causality in the context of polygenic models (cf. Bühlmann, 2020).

Acknowledgments: This research has been conducted using the UK Biobank Resource under Application Number 81202.

References

Bühlmann, P. (2020). Invariance, causality and robustness. *Statistical Science*, **35**, 404 – 426.

Klinkhammer, H., Staerk, C., Maj, C., Krawitz, P.M., and Mayr, A. (2023). A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Frontiers in Genetics*, **13**, 1076440.

Maj, C. et al. (2022). Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol. *Genetic Epidemiology*, **46**, 589 – 603.

Staerk, C., Kateri, M., and Ntzoufras, I. (2021). High-dimensional variable selection via low-dimensional adaptive learning. *Electronic Journal of Statistics*, **15**, 830 – 879.

Thomas, J., Hepp, T., Mayr, A. and Bischl, B. (2017). Probing for sparse and fast variable selection with model-based boosting. *Computational and Mathematical Methods in Medicine*, **2017**, 1421409.

Long-term foehn reconstruction combining unsupervised and supervised learning

Reto Stauffer[1,2], Georg J Mayr[3], Achim Zeileis[1]

[1] Dept. of Statistics, Universität Innsbruck, Austria
[2] Digital Science Center, Universität Innsbruck, Austria
[3] Dept. of Atmospheric and Cryospheric Sciences, Universität Innsbruck, Austria

E-mail for correspondence: `Reto.Stauffer@uibk.ac.at`

Abstract: Foehn is a wind on the leeward side of a mountain range and is characterized by a sharp increase in wind speed and changes in temperature and relative humidity. Strong foehn winds can damage trees, overturn cars, or contribute to the spread of wildfires. Therefore, understanding changes in foehn occurrences over time due to a changing climate is of great interest. Unfortunately, foehn cannot be measured directly (such as, e.g., temperature) but meteorological observations with sufficiently high resolution are needed for classification, often based on expert judgment. Consequently, foehn classifications are limited to specific periods where the necessary data and expertise are available.
To obtain fully objective probabilistic foehn classifications over long time periods we propose a novel combination of unsupervised and supervised statistical learning. Based on 10-minute observations from automated weather stations, which are only available in the last 10–20 years, probabilities for foehn occurrence are obtained from a Gaussian mixture model with concomitant variables and two components (capturing foehn vs. no foehn). The resulting probabilities are aggregated hourly to a binary foehn indicator variable which can be linked to an atmospheric reanalysis dataset that is available only at a coarser resolution but over a much longer time period (about 70–80 years). Therefore, a probabilistic binary classifier – here we use logistic regression with lasso-based stability selection – can be learned on the most recent 10–20 years and then employed to obtain out-of-sample predictions – so-called reconstructions – for foehn occurrence on the previous decades.
The method is illustrated for long-term foehn reconstruction at several stations in the European Alps. This allows to investigate possible changes in foehn occurrence in relation to climate change or to use foehn as an input for other models for which it is relevant but lacking so far.

Keywords: mixture model, variable selection, classification, reconstruction

---

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

1 Methodology

1.1 Unsupervised learning: Two-component mixture model

Figure 1 shows the stations used in this study, spanning across Switzerland and Austria. For each site, meteorological observations for the location itself as well as for a nearby mountain station are available for the last 17–22 years with a 10-minute temporal resolution.
Since foehn shows a characteristic wind direction, the mixture model is only fitted to the data where the wind blows from a pre-defined wind sector, specific for each location. If the wind direction is outside the sector, the foehn probability is set to zero.

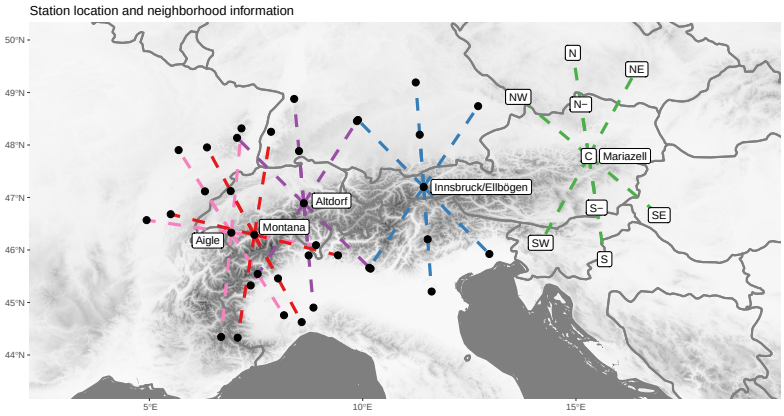


FIGURE 1. Location of the stations plus information about neighbouring locations used for calculating covariates based on the ERA5 reanalysis data set (e.g., spatial differences). Exemplarily labeled for the most eastern station (Mariazell).

To obtain the probability for foehn occurrence, a two-component Gaussian mixture model with additional concomitant variables is used (Grün and Leisch 2008). While the potential temperature difference to the nearby mountain station ($\mathbf{y}$) is used to model the parameters θ for the two Gaussian components $f()$ (foehn vs. no foehn), relative humidity and wind speed are used as concomitant variables $\mathbf{X}$ for modeling the probability π for an observation falling into the second component. The resulting two-component distribution is thus specified as follows:

$$h(\mathbf{y}, \mathbf{X}, \theta, \alpha) = \underbrace{\big(1 - \pi(\mathbf{X}, \alpha)\big) \cdot f(\mathbf{y}, \theta_1)}_{\text{first component}} + \underbrace{\pi(\mathbf{X}, \alpha) \cdot f(\mathbf{y}, \theta_2)}_{\text{second component}}$$

For the concomitant model, logistic regression is used:

$$\log\left(\frac{\pi}{1-\pi}\right) = \mathbf{X}^{\mathrm{T}}\alpha; \quad \pi = \frac{\exp(\mathbf{X}^{\mathrm{T}}\alpha)}{1+\exp(\mathbf{X}^{\mathrm{T}}\alpha)}$$

Once the required parameters θ,α are estimated, the final a-posteriori probability $\hat{p} \in [0,1]$ can be calculated.

$$\hat{p}(\mathbf{y},\mathbf{X},\theta,\alpha) = \frac{\pi(\mathbf{X},\alpha)\cdot f(\mathbf{y},\theta_2)}{\big(1-\pi(\mathbf{X},\alpha)\big)\cdot f(\mathbf{y},\theta_1)+\pi(\mathbf{X},\alpha)\cdot f(\mathbf{y},\theta_2)}$$

The probability $\hat{p}$ is the result of the objective classification and allows to obtain foehn probabilities for the same period and on the same temporal resolution for which the observations from the weather station are available.

1.2 Supervised learning: Reconstruction

For the foehn reconstruction, the result ($\hat{p}$) from the unsupervised classification is converted into a binary foehn indicator to then be combined with atmospheric reanalysis data. As the reanalysis data is only available on an hourly temporal resolution, $\hat{p}$ needs to be aggregated. The foehn indicator is set to 1 if $\hat{p} \geq 0.5$ for at least 1/3 of an hour (i.e., at least 20 minutes of foehn per hour), else to 0. This then serves as the binary response variable for the reconstruction, thus supervised learning.
As covariates, data from the fifth generation atmospheric reanalysis data set (ERA5; Hersbach et al. 2023) is used. This includes information interpolated to the target location itself and spatial and temporal differences to neighbouring locations as illustrated in Figure 1, which overall yields more than 550 possible covariates.
While any binary response model can be used here, this study makes use of logistic regression with lasso-based stability selection (Meinshausen and Bühlmann 2010) to account for the large number of covariates. Separate models are estimated for each of the six sites and for each hour of the day. Once estimated, out-of-sample prediction is used to reconstruct the foehn probability on an hourly temporal scale for the period where reanalysis data is available (1950–2023).

2 Results

By combining the unsupervised and supervised learning approaches from the previous section we first diagnose/classify foehn based on more recent 10-minute observations and to subsequently predict/reconstruct the foehn occurrence probability based on 1-hour ERA-5 data over several decades. This opens up many new possibilities for studying long-term developments of foehn occurrence as well as using foehn occurrence as an explanatory variable in other studies, e.g., in ecology or wildfire research.

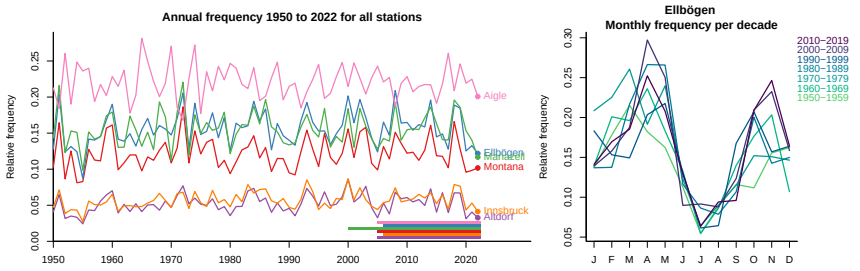


FIGURE 2. Left: Annual frequency for all six stations based on the reconstruction; the horizontal bars show the periods where observations are available. Right: Monthly frequencies for Ellbögen, separately for each decade.

Figure 2 provides some first insights into long-term trends. The left panel shows annual averages for all six stations 1950–2022. This reveals that occurrences vary substantially between stations but less so over time, without any long-term increases or decreases. The right panel shows decadal averages for each month at station Ellbögen, revealing that the periods with low foehn occurrences in summer remain very stable, while there is more variation in spring and fall. Specifically, in November there seems to be an upward trend with more foehn events in the recent two decades.
Future research will refine these insights by further extending reconstruction periods, more locations (outside Europe), and a full analysis of calibration and validation of different supervised learners (e.g., also including random forests, gradient boosting, or neural networks).

Acknowledgments: The computational results presented here have been achieved using the LEO HPC infrastructure of the Universität Innsbruck.

References

Hersbach, H., et al. (2023). ERA5 Hourly data on Pressure Levels/Single Levels from 1940 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*, (Accessed on 22-03-2023).

Grün, B. and Leisch, F. (2008). FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, **28**(4), 1 – 35.

Meinshausen, N. and Bühlmann, P. (2010). Stability Selection. *Journal of the Royal Statistical Society B*, **72**, 417 – 473.

Friedman J., Tibshirani R., and Hastie T. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1), 1 – 22.

Asymmetry model and its properties for square contingency tables

Kouji Tahata[1], Yusuke Kori[1]

[1] Tokyo University of Science, Chiba, Japan

E-mail for correspondence: kouji_tahata@rs.tus.ac.jp

Abstract: Many symmetric and asymmetric models have been proposed for square contingency tables with ordinal categories. However, fitting some symmetry and asymmetry models in statistical packages takes much work. This paper gives an equivalent expression of the asymmetry model based on the f-divergence. Additionally, the asymptotic distribution of estimators is given. We also implement the goodness-of-fit test of that model in R language. It helps compare test results using various functions f because this implementation can take arbitrary function f.

Keywords: Goodness-of-fit test; Ordinal category; Quasi symmetry; R language.

1 Introduction

Square contingency tables may arise when each subject is observed on an ordinal response at two different points in time. Such tables also occur in sample pairs of matched individuals and experiments conducted on matched pairs. For a square contingency table with the same row and column ordinal classifications, many observations concentrate on main diagonal cells or near. Thus, we are interested in considering symmetry rather than independence between row and column variables. The issues of symmetry have been treated in many studies, for example, Kateri and Papaioannou (1997), Kateri and Agresti (2007), and Tahata (2020, 2022).

For square contingency table analysis, Tahata (2020) proposed an asymmetry model based on f-divergence. We shall refer to this model as the $\mathrm{AS}_k[f]$ model. The $\mathrm{AS}_k[f]$ model includes the $\mathrm{QS}[f]$ model (Kateri and Papaioannou, 1997) and the $\mathrm{OQS}[f]$ model (Kateri and Agresti, 2007) as special cases. This study finds an equivalent representation of the $\mathrm{AS}_k[f]$ model. Additionally, we give the asymptotic distribution for the estimator

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

of parameters in that model using the result in Lang (2004). Using this representation, we implement the goodness-of-fit test of the $\mathrm{AS}_k[f]$ model in the language of R. Our implementation can take function f, score $\{u_i\}$, and k as argument. It enables users unfamiliar with contingency table analysis to analyze easily.

2 Model and its properties

Let π_{ij} denote the probability that an observation will fall in the (i,j)th cell of an $r \times r$ contingency table $(i = 1, \ldots, r; j = 1, \ldots, r)$. Assume that a set of known scores $u_1 < \cdots < u_r$ can be assigned to the rows and the columns. For a given k $(k = 1, \ldots, r-1)$, the $\mathrm{AS}_k[f]$ model is defined as

$$F(2\pi_{ij}^c) = \sum_{h=1}^{k} u_i^h \alpha_h + \gamma_{ij} \quad (i = 1, \ldots, r; j = 1, \ldots, r), \tag{1}$$

where $\gamma_{ij} = \gamma_{ji}$, $\pi_{ij}^c = \pi_{ij}/(\pi_{ij} + \pi_{ji})$, and $F(t) = f'(t)$. It should be noted that f is a twice-differential and strictly convex function on $(0, \infty)$ with $f(1) = 0$, $f(0) = \lim_{t \to 0} f(t)$, $0 \cdot f(0/0) = 0$, and $0 \cdot f(a/0) = a \lim_{t \to \infty}(f(t)/t)$. The $\mathrm{AS}_k[f]$ model has properties that can be reduced to some asymmetry models like the $\mathrm{QS}[f]$ model (when $k = r-1$) and the $\mathrm{OQS}[f]$ model (when $k = 1$) by specifying k.
When $k = r-1$, the $\mathrm{AS}_{r-1}[f]$ model can be expressed as

$$F(2\pi_{ij}^c) = \sum_{h=1}^{r-1} \prod_{s=1}^{h} (u_i - u_s)\alpha_h^* + \gamma_{ij} \quad (i = 1, \ldots, r; j = 1, \ldots, r),$$

where $\gamma_{ij} = \gamma_{ji}$. It is equivalent to (1) when $k = r-1$. When $\alpha_1^* = \cdots = \alpha_{r-1}^* = 0$, this model reduces to the symmetry model, i.e., $\pi_{ij} = \pi_{ji}$ $(i < j)$. Although the details are omitted here, we get for $j = 2, 3, \ldots, r$,

$$\alpha_{j-1}^* = \frac{F(2\pi_{1j}^c) - F(2\pi_{j1}^c) + \sum_{h=1}^{j-2} \left\{ \prod_{i=1}^{h} (u_j - u_i) \right\} \alpha_h^*}{-\prod_{i=1}^{j-1} (u_j - u_i)}. \tag{2}$$

The parameter α_h^* in the right-hand side can be replaced by (π_{ij}). Namely, equation (2) enables us to deal α_h^* as function of (π_{ij}).
Let n_{ij} denote the (i,j) cell observation in the $r \times r$ table with $n = \sum_i \sum_j n_{ij}$. Assume that a multinomial distribution with n and $\boldsymbol{\pi}$ applies to the table, where $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1r}, \ldots, \pi_{r1}, \ldots, \pi_{rr})^{\mathrm{T}}$. Also, let m_{ij} denote the expected frequency in the (i,j) cell, that is, $m_{ij} = n\pi_{ij}$.
The model class comprises MPH models expressed as $\boldsymbol{h}(\boldsymbol{m}) = \boldsymbol{0}$ using expected frequency vector $\boldsymbol{m} = (m_{11}, \ldots, m_{1r}, \ldots, m_{r1}, \ldots, m_{rr})^{\mathrm{T}}$ and constraint function $\boldsymbol{h}$ is introduced in Lang (2004). Maximum likelihood fitting and large-sample inference for MPH models are described

in the paper. The $\mathrm{AS}_k[f]$ model can be expressed as $\boldsymbol{h}(\boldsymbol{\pi}) = \mathbf{0}$ where $\boldsymbol{h}(\boldsymbol{\pi}) = (h_{1,k+2}(\boldsymbol{\pi}), \ldots, h_{1r}(\boldsymbol{\pi}), h_{23}(\boldsymbol{\pi}), \ldots, h_{2r}(\boldsymbol{\pi}), \ldots, h_{r-1,r}(\boldsymbol{\pi}))^{\mathrm{T}}$ with

$$h_{ij}(\boldsymbol{\pi}) = F(2\pi_{ij}^c) - F(2\pi_{ji}^c) - \sum_{l=1}^{k} \left\{ \prod_{s=1}^{l} (u_i - u_s) - \prod_{t=1}^{l} (u_j - u_t) \right\} \alpha_l^*.$$

It should be noted that the parameters α_l^* $(l = 1, \ldots, k)$ are the functions of $\boldsymbol{\pi}$ given by equation (2), and $\boldsymbol{h}(\boldsymbol{\pi})$ is $d_k \times 1$ vector, where $d_k = r(r-1)/2-k$. Therefore, the MPH model includes the $\mathrm{AS}_k[f]$ model in a particular case. Let $\hat{\pi}_{ij}$ and $\hat{m}_{ij}$ denote the maximum likelihood estimate (MLE) of π_{ij} and m_{ij} under the $\mathrm{AS}_k[f]$ model, respectively. That is, $\hat{m}_{ij} = n\hat{\pi}_{ij}$. The MLEs $(\hat{m}_{ij})$ may be obtained using the Newton-Raphson method to the log-likelihood equations. Let $\boldsymbol{\alpha}^*(\boldsymbol{\pi}) = (\alpha_1^*(\boldsymbol{\pi}), \ldots, \alpha_k^*(\boldsymbol{\pi}))^{\mathrm{T}}$. We consider the asymptotic distribution of $\boldsymbol{\alpha}^*(\hat{\boldsymbol{\pi}})$ with $\hat{\boldsymbol{\pi}} = (\hat{\pi}_{11}, \ldots, \hat{\pi}_{1r}, \ldots, \hat{\pi}_{r1}, \ldots, \hat{\pi}_{rr})^{\mathrm{T}}$. Note that the function $\alpha_l^*(\boldsymbol{\pi})$ $(l = 1, \ldots, k)$ is defined by equation (2). From Theorem 3 in Lang (2004), we can obtain $\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \underset{d}{\to} N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{D} - \boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}} - \boldsymbol{D}\boldsymbol{H}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{D}$. Herein, $\boldsymbol{D}$ denotes a diagonal matrix with the ith component of $\boldsymbol{\pi}$ as the ith diagonal component, and $\boldsymbol{H}$ denote the $r^2 \times d_k$ matrix of partial derivatives of $\boldsymbol{h}(\boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$. That is, $\boldsymbol{H} = \partial \boldsymbol{h}^{\mathrm{T}}(\boldsymbol{\pi})/\partial \boldsymbol{\pi}$. Using the delta method, the asymptotic distribution of $\boldsymbol{\alpha}^*(\hat{\boldsymbol{\pi}})$ can be obtained as follows. $\sqrt{n}(\boldsymbol{\alpha}^*(\hat{\boldsymbol{\pi}}) - \boldsymbol{\alpha}^*(\boldsymbol{\pi})) \underset{d}{\to} N(\mathbf{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}})$, where $\boldsymbol{A} = \partial \boldsymbol{\alpha}^*(\boldsymbol{\pi})/\partial \boldsymbol{\pi}^{\mathrm{T}}$. Let $\widehat{\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}}$ denote $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}$ with π_{ij} replaced by $\hat{\pi}_{ij}$. Additionally, the lth diagonal component of $\widehat{\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}}$ is denoted by $\widehat{v(\alpha_l^*)}$ for $l = 1, \ldots, k$. Therefore, the approximate confidence interval of α_l^* is obtained.

3 An example

We implemented `DisplayASkfResult` function that displays the result of the goodness-of-fit test in R language. Table 1 is the cross-classification of the influence of scientists (row) and officials (column) on the global warming policy (Smith et al., 2006). Note that category (1) is a great deal of influence, (2) is a fair amount, (3) is a little influence, and (4) is no influence. As an example, we fit the model with $f(t) = t\log(t)$, $u_i = i$, and $k = 1$, i.e., the linear diagonals-parameter symmetry model (Kateri and Agresti, 2007). The result of the function `DisplayASkfResult` for the data in Table 1 is shown in Figure 1. The details of the data analysis will be given in the poster presentation. Source code exists in *https://github.com/icy-mountain/MasterResearch*.

Acknowledgments: We would like to thank Professor Joseph B. Lang for providing us with the code `mph.fit`. Also, we would like to thank Professor Maria Kateri for her helpful comments.

TABLE 1. Cross-classification based on a survey about the appropriate influence level of scientists and officials on global warming policy.

Scientists's	Official's policy level				
policy level	(1)	(2)	(3)	(4)	Total
(1)	98	150	135	53	436
(2)	37	131	133	43	344
(3)	9	16	33	15	73
(4)	4	1	4	21	30
Total	148	298	305	132	883

Source: GSS (2006).

```
> freq <-c(98,150,135,53,37,131,133,43,9,16,33,15,4,1,4,21)
> ASkf_result <- DisplayASkfResult(freq=freq, f="t * log(t)", name="t", score=1:4, k=1)
**********result**********
k: 1
f: t * log(t)
df: 5
G2: 16.26
pValue: 0.006144
alpha_stars:
             Estimate  Std.Error  Confidential.Interval
alpha_star1    -1.4524    0.1010   [   -1.6504,   -1.2544] *
(*) means interval excluding 0.
```

FIGURE 1. Capture image of R Console.

References

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

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

Lang, J.B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *The Annals of Statistics*, **32**, 340 – 383.

Smith, T.W., Marsden, P., Hout, M. and Kim, J. (2006). General Social Surveys, 1972-2014[machine-readable data file] /Principal Investigator, Tom W. Smith; Co-Principal Investigator, Peter V. Marsden; Co-Principal Investigator, Michael Hout; Sponsored by National Science Foundation. -NORC ed.- Chicago: NORC at the University of Chicago [producer and distributor].

Tahata, K. (2020). Separation of symmetry for square tables with ordinal categorical data. *Japanese Journal of Statistics and Data Science*, **3**, 469 – 484.

Tahata, K. (2022). Advances in quasi-symmetry for square contingency tables. *Symmetry*, **14**, 1051.

A superiority test for comparing sensitivity, specificity, and predictive values of two diagnostic tests

Kanae Takahashi[1], Kouji Yamamoto[2]

[1] Hyogo Medical University, Japan
[2] Yokohama City University, Japan

E-mail for correspondence: kan-takahashi@hyo-med.ac.jp

Abstract: Sensitivity, specificity, positive and negative predictive values are widely used as performance measures of the diagnostic test, and these are often used simultaneously. Although there are several methods to test the superiority of these performance measures, these approaches separately compare sensitivity, specificity, positive and negative predictive values. Therefore, we propose a superiority test that can confirm that compared to an existing diagnostic test, a new diagnostic test is superior regarding at least one of the performance measures. This allows for a comprehensive determination of the superiority of the new test based on four measures. A simulation study showed that the performance of the proposed testing procedure is appropriate when the sample size is small.

Keywords: negative predictive value; positive predictive value; sensitivity; specificity.

1 Introduction

In medicine, diagnostic tests are important for early detection and treatment of disease. The sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) are widely used as performance measures of the diagnostic test. The SE is the probability that the diagnostic test result is positive among diseased subjects, and the SP is the probability of a negative that the diagnostic test result is negative among undiseased subjects. On the other hand, the PPV is the probability of having the disease when the diagnostic test result is positive, and the NPV is the probability of not having the disease when the diagnostic test result is negative.

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

If the study design is paired, the comparison of the two SEs and SPs can be performed by the McNemar's test (Zhou et al, 2011). Also, there are several methods to compare the two PPVs and NPVs (Leisenring et al, 2000; Wang et.al, 2006; Kosinski, 2013). These methods compare each performance measure separately. However, the performance of medical tests is often evaluated using SE, SP, PPV, and NPV simultaneously.
Therefore, in this study, we propose an superiority test that investigates the superiority of at least one of performance measures for a new diagnostic test compared to an existing test. Furthermore, we execute simulation studies to evaluate the performance of the proposed superiority test.

2 Proposed test

The superiority test proposed in this study is an appropriation of the idea of the approximate likelihood ratio test (Tang et al, 1989). The performance measures of the new diagnostic test are denoted as SE_1, SP_1, PPV_1, NPV_1, and those of the existing diagnostic test are denoted as SE_2, SP_2, PPV_2, NPV_2. The null hypothesis of the proposed superiority test is H_0: $SE_1 = SE_2 \cap SP_1 = SP_2 \cap PPV_1 = PPV_2 \cap NPV_1 = NPV_2$, and the alternative hypothesis is H_1: $SE_1 > SE_2 \cup SP_1 > SP_2 \cup PPV_1 > PPV_2 \cup NPV_1 > NPV_2$.
Let $\boldsymbol{m}$ be a vector whose components are $SE_1 - SE_2$, $SP_1 - SP_2$, $PPV_1 - PPV_2$, $NPV_1 - NPV_2$, and let $\boldsymbol{\Sigma}$ be a variance–covariance matrix of $\boldsymbol{m}$. Applying the delta–method and the multivariate central limit theorem, we can calculate $\boldsymbol{\Sigma}$. Let $\boldsymbol{A}$ be a positive definite matrix such that $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$. The statistic $\boldsymbol{u} = \sqrt{N}\boldsymbol{A}\boldsymbol{m}$ is approximately distributed as a four-variate normal distribution with mean $\sqrt{N}\boldsymbol{A}\boldsymbol{m}$ and covariance matrix $\boldsymbol{I}$, where N is the total sample size of observed data.
H_0 is rejected at significance level α when

$$\sum_{i=1}^{2} max(\hat{u}_i, 0)^2 > c,$$

where $\hat{\boldsymbol{u}} = (\hat{u}_1, \hat{u}_2, \hat{u}_3, \hat{u}_4)^T$ is composed of $\widehat{\boldsymbol{\Sigma}}^{-1} = \widehat{\boldsymbol{A}}^T\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{m}}$, which are calculated from observed data, and c is the critical value for the proposed test. c is determined by

$$\sum_{i=0}^{4} \frac{{}_4C_i}{2^4} Pr(\chi_i^2 \geq c) = \alpha,$$

where χ_i^2 denotes the χ^2 distribution with i degrees of freedom, and χ_0^2 is defined as the constant zero.

3 Simulation

We performed simulation studies to investigate the performance of the superiority test (Tsup). We also compared the performance of the proposed test with that of a method that combines the McNemar test and tests of positive and negative predictive values (Kosinski's test (Tkosi), Leisenring's test (Tleis), and Wang's test (Twang)) with the Holm-Bonferroni method for adjusting multiplicity. The Monte Carlo simulations were conducted and repeated 100,000 times for each method. The seven true parameters, SE_1, SE_2, SP_1, SP_2, conditional correlations between test outcomes for disease present ($\rho_{D+} = 0.40$) and for disease absent ($\rho_{D-} = 0.40$), and prevalence of disease ($\pi = 0.30$), are set for each simulation scenario to generate simulation data. The total sample size N was assumed to be 25, 50, 100, 200, and 500.

Table 1 shows the actual type 1 error rate for comparing performance measures of new and existing diagnostic tests. It shows that the actual type 1 error rates for all testing methods do not exceed the nominal type 1 error rate 0.05.

Table 2 and 3 show the empirical powers of tests. It shows that the empirical power of proposed superiority test is the highest among four methods when the sample size is small. However, that is lower than other testing methods when the sample size is large.

4 Discussion

By the result of the simulation studies, we consider that the proposed superiority test may be useful for the performance measures of comparing two diagnostic tests simultaneously when the sample size is small because the proposed test have higher power than other methods.

TABLE 1. Actual type 1 error rate
($SE_1 = SE_2 = 0.80, SP_1 = SP_2 = 0.85$,
$PPV_1 = PPV_2 = 0.70, NPV_1 = NPV_2 = 0.91$)

N	Tsup	Tkosi	Tleis	Twang
25	0.046	0.004	0.010	0.001
50	0.029	0.013	0.018	0.009
100	0.022	0.023	0.026	0.021
200	0.019	0.027	0.028	0.026
500	0.017	0.029	0.029	0.029

TABLE 2. Empirical power
($SE_1 = 0.90, SE_2 = 0.80, SP_1 = SP_2 = 0.85$,
$PPV_1 = 0.72, PPV_2 = 0.70, NPV_1 = 0.95, NPV_2 = 0.91$)

N	Tsup	Tkosi	Tleis	Twang
25	0.088	0.009	0.020	0.002
50	0.107	0.039	0.056	0.025
100	0.187	0.146	0.174	0.133
200	0.364	0.396	0.401	0.389
500	0.780	0.830	0.832	0.829

TABLE 3. Empirical power
($SE_1 = 0.85, SE_2 = 0.80, SP_1 = 0.90, SP_2 = 0.85$,
$PPV_1 = 0.78, PPV_2 = 0.70, NPV_1 = 0.93, NPV_2 = 0.91$)

N	Tsup	Tkosi	Tleis	Twang
25	0.087	0.016	0.032	0.004
50	0.121	0.073	0.099	0.051
100	0.190	0.198	0.213	0.179
200	0.342	0.387	0.396	0.372
500	0.735	0.783	0.786	0.778

References

Kosinski, A.S. (2013). A weighted generalized score statistic for comparison of predictive values of diagnostic tests. *Statistics in medicine*, **32**, 964 – 977.

Leisenring, W. et al. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics*, **25**, 2215 – 2229.

Tang, D-I., Gnecco, C., and Geller, N.L. (1989). An Approximate Likelihood Ratio Test for a Normal Mean Vector with Nonnegative Components with Application to Clinical Trials. *Biometrika*, **76**, 577 – 583.

Wang, W. et al. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. *Statistics in Medicine*, **25**, 2215 – 2229.

Zhou, X-H., Obuchowski, N.A., and McClish, D.K. (2011). *Statistical Methods in Diagnostic Medicine, Second Edition.* Hoboken: John Wiley & Sons.

Individual participant data meta-analysis: pooled effect of EEF funded educational trials on low baseline attaining group

G. Uwimpuhwe[1], A. Singh[1], N. Akhter[1], B. Ashraf[1], T. Coolen-Maturi[1], T. Robinson[1,2], S. Higgins[1], J. Einbeck[1]

[1] Durham University, Durham, United Kingdom
[2] London School of Economics and Political Science, London, United Kingdom

E-mail for correspondence: germaine.uwimpuhwe@durham.ac.uk

Abstract: The Education Endowment Foundation (EEF), a charity aiming to break the link between socioeconomic disadvantage and pupil attainment, has commissioned over 200 randomised controlled trials. The collection of data from these trials, the 'EEF Data Archive', forms a rich repository. It is vital to understand the overall impact of EEF-funded interventions on disadvantaged pupils' attainment, as well as the 'attainment gap' to their peers. The EEF Data Archive allows gaining such understanding. This study utilized individual-level pupils' data from 100 trials available in this archive, with disadvantaged pupils being indicated by students belonging to the lowest tertile of the baseline scores for each trial. A Bayesian multilevel IPD meta-analysis was applied to the standardised outcome measures (mathematics and literacy) to estimate the pooled effect size and the attainment gap. The preliminary analysis revealed that EEF-funded interventions improved low-attaining pupils' attainment for literacy (effect size: 0.033, 95% CI: 0.011, 0.055) and mathematics outcomes (effect size: 0.019, 95% CI: -0.001, 0.038). Overall, the results align with the EEF mission of increasing the attainment of disadvantaged pupils and closing the attainment gap.

Keywords: IPD meta analysis; multilevel models; effect size; educational interventions

1 Introduction

Raising the attainment of pupils from disadvantaged backgrounds, commonly defined in terms of family socioeconomic status, can help all pupils in achieving their potential (Macleod et al., 2015). This is the main mission

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

of the Education Endowment Foundation (EEF), an independent charity who sponsor education trials. For example, Higgins et al. (2016) found that an EEF-funded mastery learning intervention is a promising strategy for narrowing attainment gaps. Similarly, Gorard et al. (2014) demonstrated that pupils with low attainment prior to an EEF-funded reading intervention showed greater positive results. These individual results are promising, but only tell a partial story of the global impact EEF interventions have had on reducing attainment gaps.

In this study, we investigated the pooled effect of EEF-funded interventions on disadvantaged pupils' attainment. We estimated the attainment gaps between middle/high and low attainers using a new three-category indicator called the Baseline Attainment Group (BAG), created from the tertiles of standardised prior attainments. For each project, pre-test scores were z-standardised and then categorised into three equal groups. Tertiles of the standardised pre-test scores were created and defined as low, middle, and high attainers' groups. The lower tertile of this measure identifies low attainers at baseline, allowing us to represent the subgroup of disadvantaged pupils. Specifically, we investigate the following research questions: RQ1) What is the overall effect of EEF-funded interventions on low attainers' literacy and mathematics attainment? RQ2) Does the effect of EEF-funded interventions on literacy/mathematics attainment differ between low attainers and their peers?

2 Models and estimation

To summarise the intervention effect on low attainers, we applied the simplified meta-analysis method detailed in Ashraf et al. (2021). Specifically, to answer both research questions, we estimate the model:

$$Y^s_{ijk} = \begin{cases} \beta_{0k} + \beta_{1k}\text{Pret}^s_{ijk} + \beta_{2k}T_{ijk} + S_{jk} + \epsilon_{ijk}, \text{ for RQ1} \\ \beta_{0k} + \beta_{2k}T_{ijk} + \beta^l_{3k}\text{BAG}^l_{ijk} + \beta^l_{4k}T_{ijk}\text{BAG}^l_{ijk} + S_{jk} + \epsilon_{ijk}, \text{ for RQ2,} \end{cases}$$

where Y^s_{ijk} and Pret^s_{ijk} are standardised post-test and pre-test scores for pupils i in school j from trial k; β_{0k} is a fixed intercept, β_{1k} is a fixed gradient between the standardised post-test and pre-test scores, and β_{2k} is the average effect of the intervention in trial k; BAG^l is 0 if $\text{BAG} = l$ and 1 otherwise, using the notation $l = 1$ for middle and $l = 2$ for higher attainers categories of BAG, with low attainers forming the reference level. The parameters β^l_{4k} represent the attainment gap, i.e., the difference in the average effect of the interventions between BAG pupils (low attainers and pupils belonging to level l) in trial k; $S_{jk} \sim N(0, \omega^2_{jk})$, with ω_{jk} capturing between-school variability in trial k, and $\epsilon_{ijk} \sim N(0, \sigma^2_k)$, with σ_k capturing between-pupil variability in trial k.

Note, the pre-test variable was ignored in the attainment gap model since it forms the basis of the BAG variable and could lead to non-identifiabilities if both were included. Only the low attainers subgroup was considered to answer the first research question (RQ1). The effect of EEF-funded intervention on low attainers was then compared separately to that of middle and high attainers to answer the second research question (RQ2) regarding the attainment gap.

The pooled effect size for subgroup analysis and attainment gap can be calculated by:

$$\phi = \frac{\sum_{k=1}^{K} W_k \theta_k}{\sum_{k=1}^{K} W_k},$$

where $W_k = (\omega_{jk}^2 + \sigma_k^2)^{-1}$ captures within-trial variability, given that between-trial variability was pre-scaled to 1, with $\theta_k = \beta_{2k}$ for subgroup analysis and $\theta_k = \beta_{4k}$ for the attainment gap model.

A Bayesian framework with vague priors was used to fit the required multilevel models, from which the pooled estimates of effect sizes were computed. Credible intervals were obtained as 2.5% and 97.5% quantiles from posterior distributions of the pooled effect size estimates. All analyses were performed with manual implementations in the R package `R2jags`.

3 Results

At the time of the analysis (end of 2021) 100 projects were available in the EEF Data Archive through the Secure Research Service (SRS) environment. Since some trials have both maths and literacy outcomes, in total 52 trials with maths outcomes and 85 with literacy were available. The assessment of eligibility criteria resulted in 45 trials with literacy and 35 trials with maths outcomes to use in the final analysis. The results of the subgroup analysis are shown in Table 1, where positive effect size estimates mean that the EEF interventions has a positive effect on low attainers. Table 2 shows the attainment gap estimates, where positive estimates mean that due to the EEF interventions, lower attainers perform better than their peers.

TABLE 1. Overview of pooled effect sizes for maths and literacy outcomes on low attainers.

Outcome	Trials	Low attainers	ES (95% credible interval)
Literacy	45	70,819	0.033 (0.011, 0.055)
Maths	38	116,031	0.019 (-0.001, 0.038)

TABLE 2. Overview of the attainment gaps between low attainers and middle/high attainers. The estimates are the pooled attainment gaps and their 95% credible interval in parenthesis

Outcome	Trials	All pupils	Low vs *Middle* attainers	Low vs *High* attainers
Literacy	45	179,312	-0.001 (-0.020, 0.019)	0.003 (-0.017, 0.023)
Maths	38	270,979	0.010 (-0.007, 0.027)	-0.001 (-0.021, 0.020)

4 Discussion

The results indicate that EEF-funded interventions improved low attaining pupil's literacy outcomes with an effect size of 0.033 (0.011, 0.055). The improvement in the mathematics outcomes of low attainers was less pronounced, with an effect size of 0.019 (-0.001, 0.038). For both outcomes, there was no indication that EEF-funded interventions would widen the attainment gap. The evidence from this study can be used to support EEF stakeholders in assessing 'what worked' for these specific disadvantaged pupils.

Acknowledgments: Special thanks to the EEF for funding this research. We would like to thank FFT, part of the Fischer Family Trust, for enabling access to EEF's data archive. This work was produced using statistical data from ONS. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

References

Ashraf, B., Singh, A., Uwimpuhwe, G., Coolen-Maturi, T., Einbeck, J., Higgins, S. and Kasim, Adetayo (2021). *Individual participant data meta-analysis of the impact of EEF trials on the educational attainment of pupils on Free School Meals: 2011 - 2019.* London: EEF.

Gorard, S., See, B. H., and Siddiqui, N. (2014) *Switch-on Reading: Evaluation Report and Executive Summary February.* EEF.

Higgins, S., Katsipataki, M., Villanueva-Aguilera, A. B., Coleman, R., Henderson, P., Major, L. E., and Mason, D. (2016). *The Sutton Trust-EEF Teaching and Learning Toolkit.*

Macleod, S., Sharp, C., Bernardinelli, D., Skipp, A., and Higgins, S. (2015). *Supporting the attainment of disadvantaged pupils: articulating success and good practice: Research report November 2015.*

Estimating short-term air pollution effects on health via spectral methods

Massimo Ventrucci[1], Garritt L. Page[2]

[1] Department of Statistical Sciences, University of Bologna, Italy
[2] Department of Statistics, Brigham Young University, USA

E-mail for correspondence: massimo.ventrucci@unibo.it

Abstract: Environmental epidemiology estimation of air pollution effects on health is often conducted via generalized additive mixed models and observational data aggregated over space and time. Often times, the introduction of structured random effects can produce substantial changes in the slope coefficient which makes it challenging to interpret effect size. Spectral methods to model the slope as a function of spatial scale have recently been introduced by Guan et al (2022). We extend those methods to the spatio-temporal setting motivated by a case study on estimation of short-term air pollution effects on acute hospital admissions.

Keywords: GMRF; P-spline; spatial confounding; varying coefficient models.

1 Introduction

A fundamental task in environmental epidemiology is to estimate effects of environmental stressors such as air pollution on health. These studies are often conducted in a non-experimental setting using observational data routinely collected by governmental agencies. The effect of air pollution is known to be different depending on the length of the exposure time and a number of studies have been conducted to investigate both the short and long-term effects. Common practice in short-term studies is to use Poisson time series regression models with the daily count of hospitalisations as the outcome, the daily level of pollution as a linear predictor and smooth functions of weather variables and calendar time used to adjust for time-varying confounders (Peng et al., 2009). In this work we analyze the short-term effect of particulate matter (PM_{10}) on hospital admission for respiratory causes for the 315 municipalities in the province of Torino, Italy in 2004. In

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

total, there are 12743 residents hospitalized for respiratory causes, aggregated by municipality and day. Daily average temperature (Kelvin degrees) and PM_{10} ($\mu g/m^3$) are available at municipality level, the latter as estimates based on daily average PM_{10} concentration. We show the difficulty we encounter in interpreting the effect size of exposure on hospital admissions under various models with different space and time random effects specifications, which is symptomatic of confounding taking place between the random effects and the exposure variable. Then we describe spectral methods that have been proposed by Guan et al (2022) and propose an extension to the space-time scenario. We show how this can be useful in the context of short-term effects estimation to improve interpretation of the model output.

2 Accounting for confounding via spectral methods

Let $y_{i,t}$ and $E_{i,t}$ be the observed and expected number of hospitalizations in municipality $i = 1, \dots, 315$ and day $t = 1, \dots, 366$ respectively, we assume $y_{i,t} \sim \text{Poisson}(E_{i,t} \exp(\eta_{i,t}))$, where $\exp(\eta_{i,t})$ is the relative risk of hospitalization in municipality i at time t. We model the risks as:

$$\eta_{i,t} \quad = \quad \mu + v_t + u_i + \boldsymbol{c}_{i,t}^T \boldsymbol{\gamma} + \beta x_{i,t}, \tag{1}$$

where among the fixed effects we have observed confounders $\boldsymbol{c}_{i,t}$ such as temperature and weekend (indicator 0 or 1) and exposure $x_{i,t}$ taken as the sum of estimated daily concentrations of PM_{10} (μ/m^3) in the three days before t, in region i. The random effects $(v_1, ..., v_{366})$ and $(u_1, ..., u_{315})$ capture residual temporal and spatial structure and are modelled as a random walk (RW) of order two for time and an intrinsic conditional autoregressive (ICAR) process for space. The inferential target is β, with results being reported as $\exp(10\beta)$, i.e. the relative risk of hospitalization associated to $10\mu/m^3$ increase in PM_{10}.

Table 1 summarizes the posterior distribution for the exposure effect $\exp(10\beta)$ under several model assumptions on the random effects: spatial random effects (only $\boldsymbol{u}$), spatial and temporal random effects (both $\boldsymbol{v}$ and $\boldsymbol{u}$) and no random effects (i.e. $\boldsymbol{v}, \boldsymbol{u}$ omitted). The effect of air pollution on hospital admissions remains unclear as results varies according to the utilized model. In Table 1 we see that the estimated relative risk increase for additional $10\mu/m^3$ is 1.2% for the spatial model and 0.3% for the additive space-time model, with larger credible interval observed for the space-time model.

The situation in which inclusion of spatial random effects modifies the slope estimate is referred to as spatial confounding (Hodges and Reich, 2010; Page et al., 2017). Here we note that confounding may take place both at spatial and temporal level. In the following we describe an approach to deal with space-time confounding.

TABLE 1. Posterior summaries of $\exp(10\beta)$ under different model assumptions.

model	mean	sd	quant0.025	quant0.5	quant0.975
no random effects	0.996	0.001	0.994	0.996	0.998
only space	1.012	0.001	1.010	1.012	1.015
space + time	1.003	0.002	0.999	1.003	1.006

Guan et al (2022) introduces spectral methods to analyze spatial confounding bias as a function of spatial scale. Starting from the spectral decomposition of the precision matrix of the ICAR prior assumed on $\boldsymbol{u}$, this approach leads to modelling β in (1) as a smooth function of the obtained eigenvalues, $(\omega_1, \ldots, \omega_{315})$, ordered so that $\omega_1 \leq \ldots \leq \omega_{315}$. The slope is modelled as:

$$\beta(\omega_i) = \sum_{l=1}^{L} B_l(\omega_i) b_l \qquad i = 1, \ldots, 315 \tag{2}$$

where $(B_1(\omega_i), \ldots, B_L(\omega_i))$ are b-spline basis functions evaluated at eigenvalue ω_i and $(b_1, \ldots, b_L)$ the associated spline coefficients with an intrinsic autoregressive prior. Guan et al (2020) propose a method of adjusting for confounding based on the assumption that the relationship between exposure and outcome is only "confounded" at a broad spatial scale. In other words, it is assumed that an unbiased exposure-outcome association can only be detected when we focus on pairs of neighbors, i.e. at a small spatial scale. Under such assumption, the estimated $\beta(\omega_{315}) = \sum_{l=1}^{L} B_l(\omega_{315}) b_l$ (at the largest eigenvalue) is taken as the *unconfounded* exposure effect, i.e. unaffected by global-(spatial)scale confounding bias. This method is implemented in the `eCAR` `R` package.

The aim of the present work is to describe how the same idea can be extended to different types of spatio-temporal models for discrete data, including additive and interactions models based on intrinsic Kronecker product Gaussian Markov Random Fields. Figure 1 shows some preliminary results on the exposure effect varying over spatial frequencies ω (left) and temporal frequencies ϕ (right), under an additivity assumption $\beta(\omega, \phi) = \beta(\omega) + \beta(\phi)$. From Figure 1 we can see that both $\exp(10\beta(\omega))$ and $\exp(10\beta(\phi))$ are close to the naive estimate of 1.003, reported in Table 1, for small ω and ϕ. Interestingly, estimates are not constant over the spatial and temporal frequency domains; in particular $\exp(10\beta(\phi))$ is approximately 1 for large ϕ, suggesting that no effect can be detected at a small temporal scale. Similarly, there is no indication of a significant effect at small spatial scale.

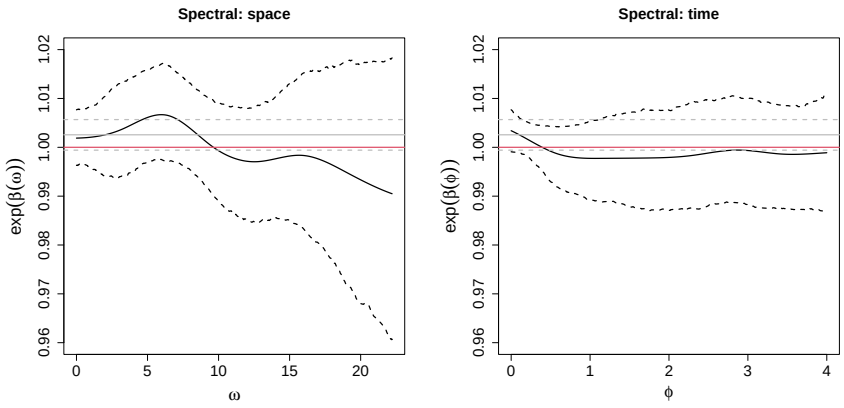


FIGURE 1. Exposure effect estimated over spatial and temporal frequencies.

3 Conclusion

While the literature on spatial confounding has flourished in recent years, less work has been done on space-time confounding. Here we have proposed an extension of spectral methods proposed by Guan et al. (2022) to space-time data with an application on short-term air pollution effects estimation in environmental epidemiology. The spectral adjustment can be extended to non-additive spatio-temporal models and this would lead to an estimated surface $\beta(\omega, \phi)$ that will inform about the effects size at small spatial and temporal scales jointly.

References

Besag., J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Stat. Soc. B*, **36**, 192 – 225.

Guan, Y., Page, G. L., Reich, B. J., Ventrucci, M., and Yang, S. (2022). Spectral adjustment for spatial confounding. *Biometrika, to appear.*

Hodges, J.S., Reich, B.J. (2010). Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. *The American Statistician*, **64**, 325 – 334.

Page, G.L., Liu, Y., He, Z., Sun, D. (2017). Estimation and prediction in the presence of spatial confounding for spatial linear models. *Scandinavian Journal of Statistics*, **44**, 780 – 797.

Peng, R.D., Dominici, F., Welty, L.J. (2009). A Bayesian hierarchical distributed lag model for estimating the time course of risk of hospitalization associated with particulate matter air pollution. *J. R. Stat. Soc. C.*, **58**, 3 – 24.

Examining quantiles of sensor outputs in structural health monitoring

Frederike Vogel[1]

[1] Department of Mathematics and Statistics, Helmut Schmidt University, Hamburg, Germany

E-mail for correspondence: vogelf@hsu-hh.de

Abstract: In structural health monitoring, high-dimensional sensor readings can be interpreted as functional data. In this paper, we apply an extended version of traditional functional principal component analysis to bridge inclination data to extend a center-based analysis by also examining conditional quantiles. This allows the extraction of observation-specific variability and patterns in the tails of the distributions of each signal.

Keywords: Functional Data Analysis; Dimension Reduction; Sensor Data

1 Introduction

Structural health monitoring (SHM) is the scientific field of analyzing signals from sensory systems attached to structures such as bridges in order to detect, localize and assess (potential) damage, and to predict the remaining service life of the structure. Accelerometers and inclinometers, for example, provide signals of vibration and inclination in which anomalies and damages should be detected to prevent a possible structural failure. In a testing scenario conducted by Jaelani et al. (2023), multiple sensors were attached to an experimental bridge at UniBw Munich (from now denoted by 'EBM'), and its inclination (in healthy condition) at ten different positions was measured during the course of $n = 21$ days with a resolution of 100Hz before damaging the bridge. Figure 1 shows the testing vicinity and example trackings in a downsampled version with a resolution of 1Hz where the frequency change was achieved by taking the median over values within the same second. For reasons of file compression, only every 10th value is plotted here. We will restrict ourselves to only using these single sensor (healthy) inclination data for further evaluation in this paper.
Due to the high resolution, sensor signals arise as almost continuous streams and can be treated as functional data over the course of the day. On account of environmental factors such as temperature and solar radiation, it

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

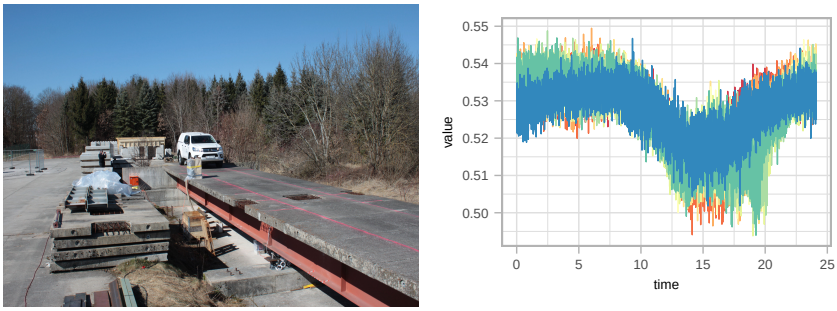


FIGURE 1. Left: The testing vicinity at EBM (Fabian Seitz, 2022). Right: Example of daily trackings of one of a total of ten inclinometers.

can be expected to see some recurring daily patterns in the sensor data, and treating the data as replicates across days $i = 1, \ldots, n$ seems to be a sensible approach. A common tool in functional data analysis (FDA) to reduce dimensionality is functional principal component analysis (FPCA) where the main directions of variability are identified. In the SHM field, FPCA extends traditional physics-based methods and has found numerous applications in which it has proven useful means of analysis, for a comprehensive review see Momeni and Ebrahimkhanlou (2022). Regular FPCA displays curves as deviations from the mean. However, under ambient influences and damages, it is reasonable to expect that the whole distribution of a signal is affected so that an analysis merely in the center of the data is likely to not be sufficient, see Tee et al. (2013). The FQPCA algorithm by Civieta et al. (2022) extends the traditional form of FPCA to also include functional (conditional) quantiles and capture sample-specific variability. In SHM, to the best of our knowledge, there is yet no work that includes quantiles as an option to summarize sensor signals in terms of dimension reduction but also include day-specific conspicuities, which may indicate damage. In this paper, we present the application of the FQPCA algorithm to the example data from EBM and show that it is able to capture differences among the samples.

2 Methods

Let $X_i(t)$, $t \in \mathcal{T}$, $i = 1, \ldots, n$, describe a set of sensor signals defined on some bounded closed interval $\mathcal{T}$, e.g., $\mathcal{T} = [0, 1]$, which is a standardized version of 'day' (24h) in this paper. The quantiles of X_i are defined as

$$Q_{X_i}(t, \tau) = \inf\{x \in \mathbb{R}, F_{X_i(t)}(x) \geq \tau\}, \tag{1}$$

where $F_{X_i(t)}$ denotes the cumulative distribution function of $X_i(t)$. FQPCA proposes a latent structure of

$$Q_{X_i}(t,\tau) = \phi_0(t,\tau) + \sum_{j=1}^{r} \lambda_{ij}(\tau)\phi_j(t,\tau)$$

to model the (day-specific) τ-quantile of sample curve i, with loadings $\phi(t,\tau) := (\phi_0(t,\tau),\ldots,\phi_r(t,\tau))'$, scores $\lambda_i(\tau) := (\mathbf{1},\lambda_{i1}(\tau),\ldots,\lambda_{ir}(\tau))'$, and $r \in \mathbb{N}$ denoting the number of principal components. Note that the intercept $\phi_0(t,\tau)$ models the general quantile trend, the matrix of loadings $\phi(t,\tau)$ is common to all observations, and the score vectors $\lambda_i(\tau)$ are signal-specific. This model is therefore comparable to traditional FPCA where functional components are used to describe deviations from the mean function of the data, but additionally promotes a tail-based analysis. To estimate the scores and loadings the objective function

$$M(\Lambda(\tau),\phi(t,\tau)) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \rho_\tau(X_i(t) - \lambda_i(\tau)'\phi(t,\tau))dt$$

is minimized, with $\Lambda(\tau) = (\lambda_1(\tau),\ldots,\lambda_n(\tau))'$ being the matrix of all scores, and $\rho_\tau(u) = u(\tau - I(u<0))$ is the quantile regression check loss function. The solution is found using an iterative algorithm based on a probabilistic approach (Civieta et al., 2022).

3 Results and Discussion

Each individual inclination signal as displayed in Figure 1 follows almost the same shaped curve/general trend with no clear outliers being present. Applying the FQPCA algorithm with a total number of $r = 6$ principal components, however, reveals some variability among trackings.

Figure 2 shows the .1, .5, and .9 quantile estimates for the example inclinometer data. The top-left quantiles (day 1), for instance, follow a u-shaped curve, while the two other samples reveal more w-shaped quantile estimates. Furthermore, the .9 quantile estimate of day 2 and day 3, as opposed to the almost equally leveled .1 and .5 quantiles, is located relatively high throughout, indicating an increased presence of peaks in the signals. It is to be expected that such sample specifics unfold to a much bigger extent once the structure sustains damage and it is therefore desirable to monitor deviation or changes in the signals over time. Several FPCA-based approaches have been proposed, compare Colosimo and Pacella (2010), including, for instance, the monitoring of the corresponding scores.

Figure 3 shows the quantile-specific scores for the first three components for the current application over the course of the tracking period. In the case of the .1 and .5 quantile estimates, the corresponding scores oscillate

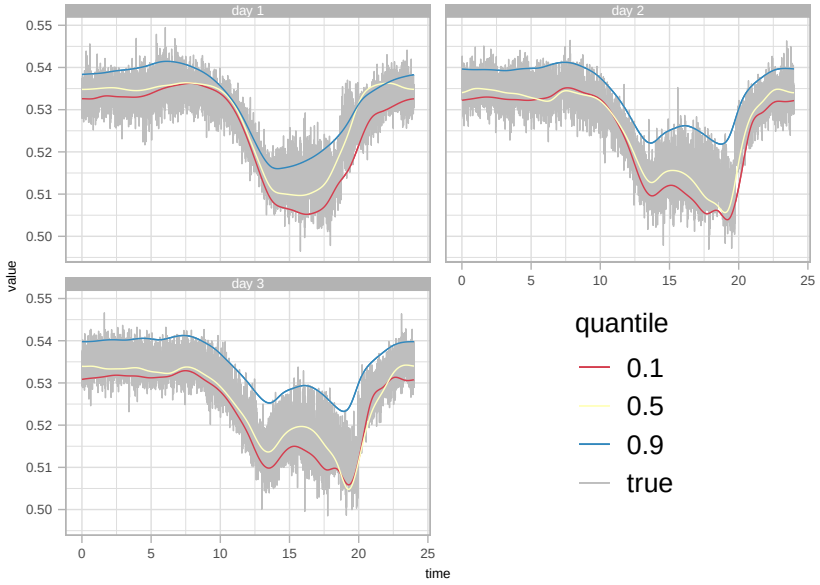


FIGURE 2. Example daily signals from EBM using trackings from one inclinometer and the corresponding .1, .5, and .9 quantile estimates (top row) and scores (bottom row) as provided by FQPCA.

around the zero mark with little variability, while in the .9 quantile case there is a presence of outliers for the first and second component indicating that a trend removal prior to monitoring is necessary. In the SHM field, one main factor of nuisance is the variation in temperature, compare Han et al. (2021). Future work could aim at incorporating the daily temperature functions into the latent quantile structure (1) by, e.g., adding a (non)parametric temperature-dependent shift to the intercept component, i.e.,

$$Q_{X_i}(t,\tau) = \phi_0(t,\tau) + f(T_i(t),\tau) + \sum_{j=1}^{r} \lambda_{ij}(\tau)\phi_j(t,\tau),$$

where $T_i(t)$ refers to the daily temperature function concurrently tracked alongside the sensor output $X_i(t)$, and f is a function that accounts for the confounding effect of the temperature. This way, both day-specific differences as well as overall structural health states over time can be assessed, providing a comprehensive analysis of sensor data in SHM.

Acknowledgments: This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

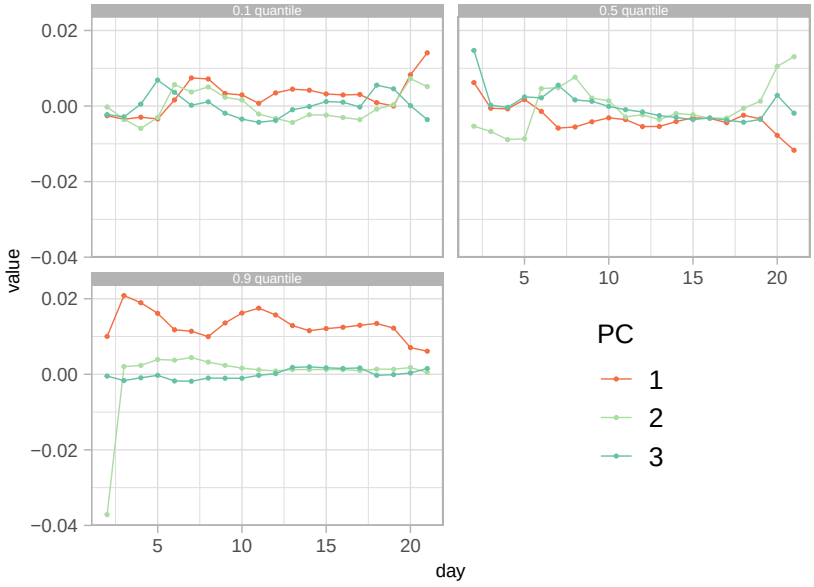


FIGURE 3. FQPCA scores for the .1, .5, and .9 quantile estimates over the cause of the entire tracking period of $n = 21$ days.

References

Civieta, Á.M., Wei, Y. and Goldsmith, J. (2023). Functional Quantile Principal Component Analysis. *Submitted for Publication.*

Colosimo, B.M. and Pacella, M. (2010). A Comparison Study of Control Charts for Statistical Monitoring of Functional Data. *International Journal of Production Research*, **48**, 1575 – 1601.

Han, Q., Ma, Q., Xu, J. and Liu, M. (2021). Structural Health Monitoring Research Under Varying Temperature Conditions: A Review. *Journal of Civil Structural Health Monitoring*, **11**, 149 – 173.

Jaelani, Y., Klemm, A., Wimmer, J., Seitz, F., Köhncke, M., Marsili, F., Mendler, A., von Danwitz, M., Keßler, S., Henke, S., Gündel, M., Braml, T. and Popp, A. (2023). Developing a Benchmark Study for Bridge Monitoring. Accepted for Publication in: *Steel Construction.*

Momeni, H. and Ebrahimkhanlou, A. (2022). High-dimensional Data Analytics in Structural Health Monitoring and Non-destructive Evaluation: A Review Paper. *Smart Materials and Structures*, **31**, 043001.

Tee, K.F., Cai, Y. and Chen, H-P. (2013). Structural Damage Detection Using Quantile Regression. *Journal of Civil Structural Health Monitoring*, **3**, 19 – 31.

Change point regression for estimated time series: An application to COVID-19 hospitalization data

Maximilian Weigert[1,2], Kai Becker[1], Helmut Küchenhoff[1,2]

[1] Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Germany
[2] Munich Center for Machine Learning, Munich, Germany

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

Abstract: Change point regression models were frequently used during the COVID-19 pandemic to analyze changes in temporal trends. When this modelling framework is applied to (partly) estimated time series, uncertainty arising from the estimation process of the time series should be taken into account. In this work, we present a general strategy to incorporate this additional uncertainty into the estimation of standard errors and confidence intervals for regression coefficients and for the location of the change points using variance decomposition based on Rubin's rules for multiple imputation. Our approach is illustrated with an application to COVID-19 hospitalization time series in Germany, which are based on nowcast estimates to enable real-time analyses.

Keywords: change point models; segmented regression; nowcasting; uncertainty quantification; COVID-19 data

1 Introduction

During the COVID-19 pandemic, change point regression was used in many studies (e.g. Küchenhoff et al., 2021) to analyze relevant changes in temporal trends in COVID-19 cases or other metrics of the pandemic. This modelling framework enables the estimation of standard errors and confidence intervals for linear effects as well as for the location of the change points. If the underlying time series, however, is not (fully) observed, but (partly) estimated, additional uncertainty arising from the estimation process of the time series is ignored. This is a common situation for public COVID-19 data, where e.g. new cases are often included with a delay of

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

a few days. With our work, we provide a general framework to easily incorporate additional uncertainty into the uncertainty estimates of change point models.

2 Data

Our study originates from the analysis of COVID-19 hospitalizations in Germany which are publicly provided by the Robert Koch Institute (Robert Koch Institute, 2023). Hospitalizations which are associated with a COVID-19 infection are not reported with regard to the hospitalization date, but the date when the infection was reported. This leads to incomplete data in a real-time setting as a relevant number of hospitalizations belonging to the latest reporting dates of infection has not yet occurred. Figure 1 illustrates how the shape of a time series for the same period changes over time. Thus, realistic estimates of hospitalizations are required to derive conclusions about temporal trends and potential change points in real-time analyses.

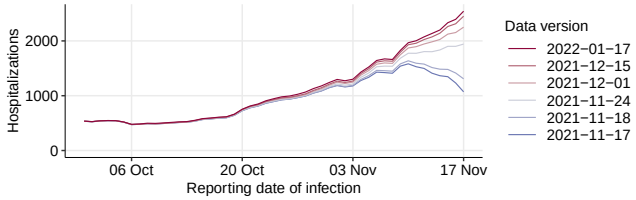


FIGURE 1. Reported rolling weekly sums of COVID-19 related hospitalizations for the period between October 1st, 2021 and November 17th, 2021 based on different public data versions.

3 Methods

We obtain estimated numbers of hospitalizations $\hat{Y}_t$ for time points $t = 1, ..., T$ by applying the nowcasting procedure by Fritz et al. (2022). This approach uses generalized additive regression to model the delay distribution between the reporting date of infection and the reporting date of hospitalization. Confidence intervals of the nowcast estimates are derived relying on the strategy by Krinsky and Robb (1986) who propose to draw bootstrap samples from the multivariate normal distribution of estimated regression coefficients. Within these bootstrap samples, individual nowcast time series $\hat{Y}_t^{(b)} (b = 1, ..., B)$ are estimated.
To detect change points in the nowcasted time series $\hat{Y}_t$, we estimate a segmented regression model based on the formulation of Muggeo (2003).

For exponential family responses with expected value μ and link function $g(\cdot)$, we use the model structure

$$g(\mu_t) = \beta_0 + \eta_t + \alpha t + \sum_{k=1}^{K} \gamma_k (t - \psi_k)_+, \tag{1}$$

where β_0 denotes the intercept, η a linear predictor containing additional covariate effects, $\gamma = (\alpha, \gamma_1, ..., \gamma_K)$ the vector of regression coefficents regarding the temporal trend, $\psi = (\psi_1, ..., \psi_K)$ the vector of change point locations and K the number of change points.

This model, however, does not take the uncertainty induced by the nowcasting procedure into account. The obtained variance estimate $\hat{\mathbb{V}}(\hat{\theta})$ for an arbitrary estimated coefficient $\hat{\theta}$ only reflects the parameter uncertainty according to the change point model. To incorporate the additional source of uncertainty, we make use of the already available bootstrap samples from the nowcasting step and fit change point models for all bootstrapped nowcast time series $\hat{Y}_t^{(b)}$. Building on Rubin's rules for multiple imputation (Rubin, 2004), we can consider the bootstrapped time series as imputed and not fully observed datasets. The overall variance $\hat{\mathbb{V}}_{\text{total}}(\hat{\theta})$ can then be expressed as a weighted sum of within- and between-sample variance through

$$\hat{\mathbb{V}}_{\text{total}}(\hat{\theta}) = \hat{\mathbb{V}}_{\text{within}}(\hat{\theta}) + \left(1 + \frac{1}{B}\right) \hat{\mathbb{V}}_{\texttt{between}}(\hat{\theta}), \tag{2}$$

where $\mathbb{V}_{\text{within}}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\mathbb{V}}(\hat{\theta}^{(b)})$ denotes the within-sample variance, $\mathbb{V}_{\text{between}}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \bar{\hat{\theta}})^2$ the between-sample variance with $\bar{\hat{\theta}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}$ and B the number of bootstrap samples. For large B, the overall variance converges to the unweighted sum of both components.

The number of required change points K is determined by means of the BIC. To ensure comparability between the change point models in the different bootstrap samples, the choice of K is based on the original sample and kept as a fixed parameter over all bootstrap samples. Furthermore, a matching scheme is applied to assign the estimated change points within the samples to the closest change point of the original nowcast time series.

4 Results

Figure 2 exemplarily visualizes the nowcast estimates including 95% confidence intervals for the rolling weekly sum of COVID-19 related hospitalizations in the German federal state of Bavaria on September 22nd, 2021 and November 17th, 2021. As the nowcasting model assumes all hospitalizations to occur within 40 days after the reporting date of infection, uncertainty in the resulting time series only exists for the last 40 days. While the results for September 22nd indicate a slight decline over the last two weeks with

rather high uncertainty, the nowcasting results for November 17th reveal a clear upward trend which would not be visible from the reported data.

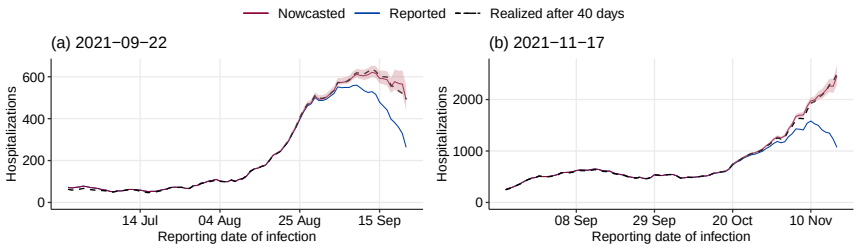


FIGURE 2. Comparison of reported and nowcasted COVID-19 related hospitalizations in Bavaria with 95% confidence intervals (red shades) over a period of 90 days based on data from September 22nd, 2021 (left) and November 17th, 2021 (right). The dashed black lines show the realized hospitalizations after 40 days.

To detect change points in the temporal course of hospitalizations, we use a negative binomial segmented regression model with a logarithmic link for a time series over a period of 90 days. Uncertainty adjustment is based on 100 bootstrap samples. Figure 3 illustrates the results obtained for the data available on September 22nd and November 17th, 2021 with four and six change points, respectively. For both considered periods, the presented strategy of uncertainty adjustment leads to higher standard errors and wider confidence intervals for the estimated locations of recent change points. Between-sample variance accounts for about 67% of the overall variance of the latest change point for September 22nd. For November 17th, the increase of uncertainty through the nowcasting procedure for the latest change point is rather moderate.

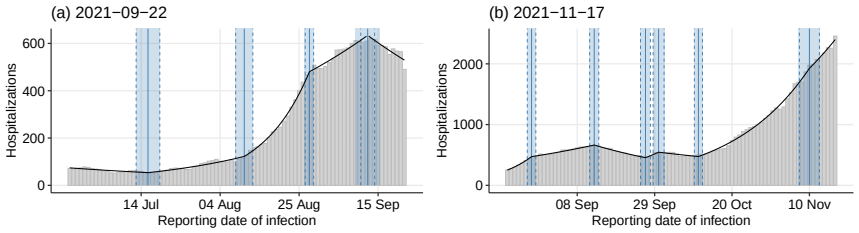


FIGURE 3. Fitted curves (black lines) and estimated change points (solid blue lines) for nowcasted COVID-19 hospitalizations in Bavaria (grey) over a period of 90 days based on data from September 22nd, 2021 (left) and November 17th, 2021 (right). Blue shades represent 95% confidence intervals for change points with uncertainty adjustment, dashed blue lines the borders without adjustment.

5 Conclusion

The presented framework aims to include additional estimation uncertainty into change point regression for estimated time series. We illustrated our concepts with an application to COVID-19 hospitalizations in Germany. In this context, we will retrospectively evaluate our outlined approach based on daily hospitalization data versions for all 16 German federal states over a period of one year. The proposed strategy of uncertainty adjustment, however, is more general and can be easily adapted to other research settings, also with fully estimated time series. Further, the strategy is not limited to bootstrap procedures for uncertainty quantification of the estimation of the underlying time series, but can also be applied to other settings such as Bayesian frameworks with posterior distributions.

References

Fritz, C., De Nicola, G., Rave, M., Weigert, M., Khazaei, Y., Berger, U., Küchenhoff, H., and Kauermann, G. (2022). Statistical modelling of COVID-19 data: Putting generalized additive models to work. *Statistical Modelling*.

Krinsky, I., and Robb, A.L. (1986). On approximating the statistical properties of elasticities. *The review of economics and statistics*, 715 – 719.

Küchenhoff, H., Günther, F., Höhle, M., and Bender, A. (2021). Analysis of the early COVID-19 epidemic curve in Germany by regression models with change points. *Epidemiology & Infection*, **149**.

Muggeo, V.M. (2003). Estimating regression models with unknown breakpoints. *Statistics in medicine*, **22(19)**, 3055 – 3071.

Robert Koch Institute (2023). COVID-19-Hospitalisierungen in Deutschland. Last retreived: 31/05/2023. `https://github.com/robert-koch-institut/COVID-19-Hospitalisierungen_in_Deutschland`.

Rubin, D.B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

A consistent way to define p-values

Paul Wilson[1], Jochen Einbeck[2]

[1] University of Wolverhampton, UK
[2] Durham University, UK

E-mail for correspondence: PaulJWilson@wlv.ac.uk

Abstract: p-values for discrete distributions have traditionally been defined as for those of continuous data. Mid p-values are sometimes regarded as an alternative form of p-value for use with discrete data. We show that that the mid p-value definition is preferable to the traditional one for discrete data, and is equivalent to the traditional definition of the p-value in the case of continuous data, thus we propose that it is used for all distributions, continuous or discrete.

Keywords: p-values; mid p-values; discrete test statistics.

1 Introduction

The "traditional" definition of a p-value as the "probability of observing a test statistic at least as extreme as that observed" is well suited to the case where the distribution of the test statistic is continuous, but not so when it is discrete. In this paper, for clarity of notation and ease of explanation, we will assume when discussing discrete distributions that they take values $0, 1, 2, \ldots$, but the arguments may easily be adapted for other discrete values. If X is a continuous random variable with (strictly increasing) cumulative distribution function $F_X(x) = P(X \leq x)$ then the equation $F(x) = p$ has an unique solution ξ_p, and $F_X(\xi_p) = p$, the (left-tailed) p-value associated with ξ_p, $P(X > \xi_p)$ often being referred to as the right-tailed p-value. ξ_p is referred to as the $100p^{th}$ percentile or the p-quantile of the distribution. It is well known that the distribution P_0 of p-values under a null hypothesis is uniformly distributed on $(0, 1)$, and hence $E(P_0) = \frac{1}{2}$ and $\mathsf{Var}(P_0) = \frac{1}{12}$. If however F_X is not continuous and strictly increasing, as is the case when X is discrete, then the equation $F(x) = p$ fails to have an unique solution. The traditional approach when X is discrete is to define the p^{th} quantile of a random variable X or of its corresponding distribution

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

as the least number ξ satisfying $F_X(\xi) \geq p$, (note that here $\xi \in \mathbb{N}$), and the corresponding p-value as $P(X \leq \xi_p)$, thus many quantiles share the same value; for example, we see from Table 1 that for a $\mathsf{Bin}(3, 0.3)$ distribution for any κ, $0.35 < \kappa < 0.78$, $\xi_\kappa = 1$. Let $P_0^\star$ denote the distribution of traditional p-values under the null hypothesis. $P_0^\star$ is itself discrete on $[p_0, 1)$ if X is infinite, or $[p_0, 1]$ if X is finite, (where $p_i = P(X = i)$). It may be shown (see for example, Berry and Armitage (1995)), that $E(P_0^\star) = 0.5 + 0.5 \sum p_i^2$. For example for a $\mathsf{Bin}(3, 0.3)$ distribution (see Table 1): $E(P_0^\star) = 0.674$ and $\mathsf{Var}(P_0^\star) = 0.063 \approx \frac{1}{16}$.

2 Mid-quantiles and Mid *p*-values

For discrete distributions, Franck (1986) strongly advocates the use of the mid p-value, drawing on previous research by Lancaster (1961), Dempster (1965) and Stone (1969). Mid p-values are defined by:

$$p_{\text{mid}} = P(X < x) + \tfrac{1}{2} P(X = x). \tag{1}$$

Table 1 summarises the traditional and mid p-values for a Bin$(3, 0.3)$ distribution. It may be shown (Berry and Armitage, 1995) that under the null hypothesis:

$$E(P_{\text{mid}}) = \frac{1}{2} \quad \text{and} \quad \mathsf{Var}(P_{\text{mid}}) = \frac{1}{12}\left(1 - \sum p_i^3\right), \tag{2}$$

where the random variable P_{mid} is the distribution of p_{mid}.

TABLE 1. $\mathsf{Bin}(3, 0.3)$

Traditional and mid p-values.

x	0	1	2	3
$P(X = x)$	0.343	0.441	0.189	0.027
Trad p-value	0.343	0.784	0.973	1.000
Mid p-value	0.172	0.564	0.878	0.986

Mid-quantiles are defined analogously Let X be a discrete random variable with distinct values $v_1 < v_2 < \cdots < v_d$, let $P(X = v_i) = p_i$.

$$\xi_{\text{mid}}(p) = \begin{cases} v_1 & \text{if } p < p_1/2 \\ v_k & \text{if } p = \pi_k,\ k = 1, \ldots, d \\ \lambda v_k + (1-\lambda) v_{k+1} & \text{if } p = \lambda \pi_k + (1-\lambda)\pi_{k+1} \\ & 0 < \lambda < 1,\ k = 1, \ldots, d-1 \\ v_d & \text{if } p > \pi_d \end{cases}$$

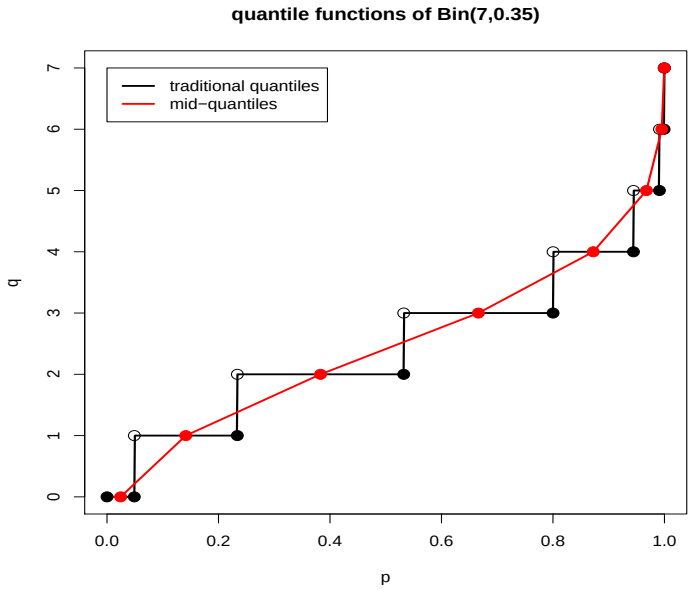


FIGURE 1. Attainment: Traditional and mid-quantiles

In essence, mid-quantiles and mid p-values align for values at which the mid p-values exist, and are obtained by linear interpolation at values inbetween. See Figure 1.

For a more detailed account of mid-quantiles see Ma et al. (2011) or Wilson and Einbeck (2018).

The employment of mid p-values has many advantages: under the null hypothesis the expected values of mid-p values for discrete test statistics and traditional p-values for continuous test statistics are equal, and their variances are approximately equal (see (2)), mirroring the continuous case. Also the power and attainment (type-one error) rate of such tests is improved by the use of mid p-values. Two such tests are the Wilcoxon-Mann-Whitney rank-sum test (Mann and Whitney, 1947), which henceforth we refer to as the Wilcoxon test, and the zero-modification test of Wilson and Einbeck (2018) (to which we refer as the Wilson-Einbeck test henceforth for terminological convenience), which uses the number of observed zeros as a test statistic for zero-inflation or deflation. Simulation studies, based upon one million resamples, show that for a Wilcoxon test on two samples of size seven, under the null hypothesis the distribution of traditional p-values has mean 0.518 and variance 1.001/12, whereas that of mid p-values has mean 0.500 and variance 1.000/12. Similarly simulation studies based upon one million resamples of size $n = 50$ show that under the null hypothesis of

a Poisson model with parameter $\mu = 3$ under a Wilson-Einbeck test of zero-inflation, the distribution of traditional p-values has mean 0.602 and a variance of 1.047/12 whereas that of mid p-values has a mean of 0.508 and a variance of 0.965/12. Figure 2 illustrates the attainment rate of a Wilson-Einbeck test of zero-inflation under the null hypothesis of a Poisson distribution with parameters varying from 0.5 to 4.5, and a sample size of $n = 500$. Clearly the attainment when mid p-values are employed is superior to that when traditional p-values are used. Figure 3 illustrates the power of a Wilson-Einbeck test of zero-inflation for samples of size $n = 50$, where under the null hypothesis of a Poisson distribution with parameter $\mu = 1$, we see that the use of mid p-values results in increased power. Figure 4 illustrates the attainment rates obtained when Wilcoxon rank sum tests, (using the exact test-statistic) are performed on two equally sized samples of sizes 4 to 25; again the mid p-values out-perform the traditional.

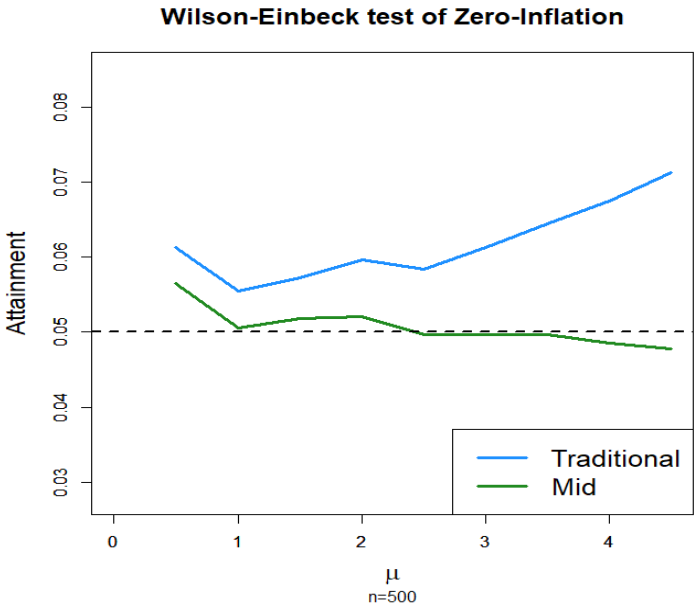


FIGURE 2. Attainment: Wilson-Einbeck test of zero-inflation

3 Conclusion: All p-values should be mid p-values

We have seen that for discretely distributed test statistics the adoption of mid p-values has many advantages, especially leading to superior attainment rates than traditional p-values. Note that for a continuous test statistic the definition of a traditional and mid p-value (1) are equivalent

FIGURE 3. Power: Wilson-Einbeck test of zero-inflation. (Poisson distribution, $\mu = 1$, $n = 50$)

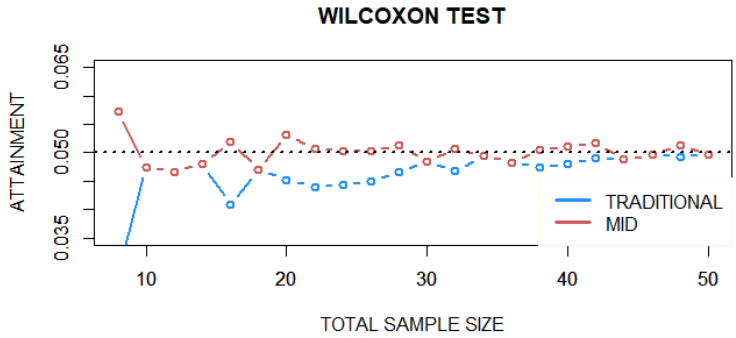


FIGURE 4. Attainment: Wilcoxon rank sum test

as for any continuous random variable X, one has $P(X = x) = 0$. Hence, traditional p-values for the continuous case *are* in fact mid $p-$values, and thus the definition of the mid p-value can be conveniently used to cover both the discrete and continuous case, thus we propose that the definition of the mid p-value should be use for *all* p-values.

References

Berry, G. & Armitage, P. (1995). Mid-P confidence intervals: a brief review. *The Statistician*, **44**, 417 -– 423.

Franck, W. (1986) P-values for discrete test statistics. *Biometr. J*, **28**, 403,– 406.

Lancaster, H.O. (1961) Significance tests in discrete distributions. *Journal of the American Statistical Association*, **60**, 233 – 234.

Ma, Y., Genton M. and Parzen, E. (2011) Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics* **63**, 227 – 243.

Mann, H. & Whitney, D. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*. **18 (1)**, 50 -– 60.

Stone (1969) The role of significance testing. Some data with a message. *Biometrika*, **56**, 485 – 493.

Wilson, P. & Einbeck, J. (2017) Sample quantiles corresponding to mid p-values for zero–modification tests. In: Grzegorczyk, M. and Ceoldo, G. (Eds). Proc's of the 32nd IWSM, Groningen, The Netherlands, Vol 1, pages 275 – 279.

Wilson, P. & Einbeck, J. (2018) A new and intuitive test for zero-modification. *Statistical Modelling*, **19 (4)**, 341 – 361. doi: 10.1177/1471082x18762277.

Modelling SHM sensor outputs: A functional data approach

Philipp Wittenberg[1], Jan Gertheiss[1]

[1] Dep. of Mathematics and Statistics, School of Economics and Social Sciences, Helmut Schmidt University, Hamburg, Germany

E-mail for correspondence: pwitten@hsu-hh.de

Abstract: Structural Health Monitoring (SHM) is increasingly applied in civil engineering. One of its primary purposes is detecting and assessing changes in structure conditions to reduce potential maintenance downtime. Recent advancements, especially in sensor technology, facilitate data measurements, collection, and process automation, leading to large data streams. We propose a function-on-function regression approach for modelling the sensor data and adjusting for confounder-induced variation.

Keywords: Functional Principal Component Analysis; Functional Regression

1 Introduction

Structural health monitoring (SHM) uses sensor data from buildings such as bridges to detect, localize and/or quantify damages. These measurements are not obtained under laboratory conditions, so the data also depends on environmental influences such as temperature. Therefore, a model to adjust for covariates is required when analysing the data. This paper considers the recent data set by Jaelani et al. (2023). It consists of sensor measurements of a test bridge in Munich, Germany. Among other variables, the strain was measured with six strain gauges in 100 hertz and one external temperature sensor in 1 hertz over 22 days in 2022 between March 11 and April 1. The strain data were downsampled to 1 hertz to use the same frequency in both types of measurements, resulting in 1,900,800 observations per sensor. Figure 1 shows a digital representation of the bridge. The strain sensors were evenly distributed on both sides of the bridge. The external temperature sensor inside a meteorological station was positioned on the north side of the bridge.
A recent and comprehensive review of methods for SHM under varying temperature conditions is provided by Han et al. (2021). Regarding forecasting and separating temperature-induced from structural responses, it

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

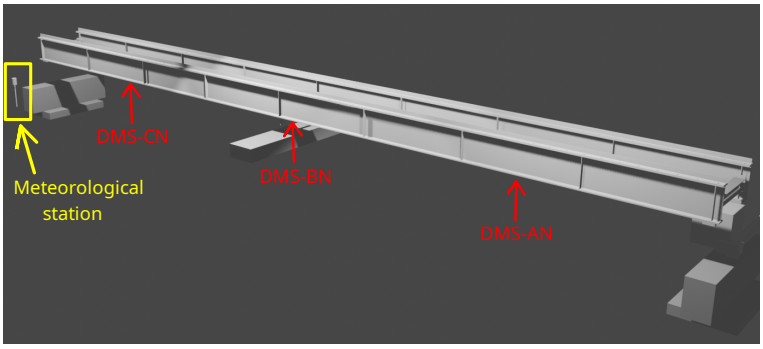


FIGURE 1. Digital representation of the Munich test bridge with the positions of the three strain sensors (DMS) on the north side of the bridge.

is typically distinguished between input-output methods and output-only approaches. In the first case, both observations of the sensor output and confounding variables such as temperature are considered, while in the latter case, only the vibration or static (such as strain) responses of the structures are used, often using projection methods such as principal component analysis (PCA). Among input-output methods, a very popular approach is regressing sensor measurements on environmental and/or operational variables, also known under the name *response surface* modelling. Then, following the so-called "subtraction method" the predicted sensor data is subtracted from the observed data, and the residuals are used for further analysis. Various methods exist for fitting regression function(s) to the data, ranging from simple linear or polynomial regression to advanced machine learning approaches such as artificial neural networks.
This paper presents a functional data approach where we consider the sensor measurements as functional data across days rather than each measurement as a single sampling instance. Also, our method can be seen as a combination of subtraction and projection methods.

2 A Functional Data Approach

The model we assume has the following basic form. To keep things simple, we restrict ourselves to a single, functional covariate $z_j(t)$, e.g., denoting the temperature at time $t \in \mathcal{T}, \mathcal{T} = [0\text{h}, 24\text{h}]$, and day j, and a single sensor output $u_j(t)$. Then,

$$u_j(t) = \alpha(t) + f_u(z_j(t)) + E_j(t), \tag{1}$$

where $\alpha(t)$ is a fixed functional intercept, $f_u(z_j(t))$ is a fixed, potentially non-linear effect of temperature, and $E_j(t)$ is a day-specific, functional error

term. This model has (at least) two advantages over response surface modelling as used so far: (i) The functional intercept $\alpha(t)$ captures recurring daily patterns that cannot be explained through the available environmental or operational variables, e.g., because the factors causing them are not monitored. In the case of a longer monitoring period, we may extend the one-dimensional $\alpha(t)$ to a two-dimensional, smooth surface $\alpha(t, d_j)$, where d_j denotes the time/date of the year corresponding to the day j in the data set. (ii) The error term/process $E_j(t)$ is typically correlated over time, i.e., in t-direction. However, ignoring this correlation when fitting α and f_u through ordinary least squares, common maximum likelihood, or a similar approach that assumes conditional independence between measurements will typically lead to less accurate estimates and, more importantly, biased measures of statistical uncertainty such as confidence or prediction intervals. In SHM, the latter can be particularly harmful if those intervals are used for determining if measurements are "out of control".
In the framework of SHM, there is another essential aspect to note concerning $E_j(t)$: This process contains the relevant information for SHM, since it captures deviations from the sensor output $\alpha(t) + f_u(z_j(t))$ that would be expected for a specific temperature if the structure/bridge is "in control". For exploiting this information, it is necessary to decompose this process into a more structural component, say $w_j(t)$, and pure noise $\epsilon_j(t)$, i.e.

$$E_j(t) = w_j(t) + \epsilon_j(t). \tag{2}$$

For doing so, we make use of the Karhunen-Loeve expansion, which allows us to expand the random function $E_j(t)$ as

$$E_j(t) = \sum_{r=1}^{\infty} \xi_{rj}\phi_r(t) = \sum_{r=1}^{m} \xi_{rj}\phi_r(t) + \epsilon_j(t), \tag{3}$$

where ϕ_r are orthonormal eigenfunctions of the covariance operator of the E-process and ξ_{rj} are the respective "scores". Estimating the eigenfunctions and scores is known under the name *functional* principal component analysis (FPCA). The obtained scores $\hat{\xi}_{1j}, \ldots, \hat{\xi}_{mj}$ can be used, e.g., as input for control charts and further monitoring schemes; see Centofanti et al. (2021).

3 Results

Figure 2 illustrates the model output for the strain sensor DMS-BN (compare Figure 1), with the functional intercept, as a two-dimensional surface, being allowed to vary across days. The model was fit using R packages `mgcv` (Wood, 2017) and `refund` (Goldsmith et al., 2022). For modelling the functional intercept, we used a tensor product approach. The temperature effect was fit as a one-dimensional cubic regression spline with the

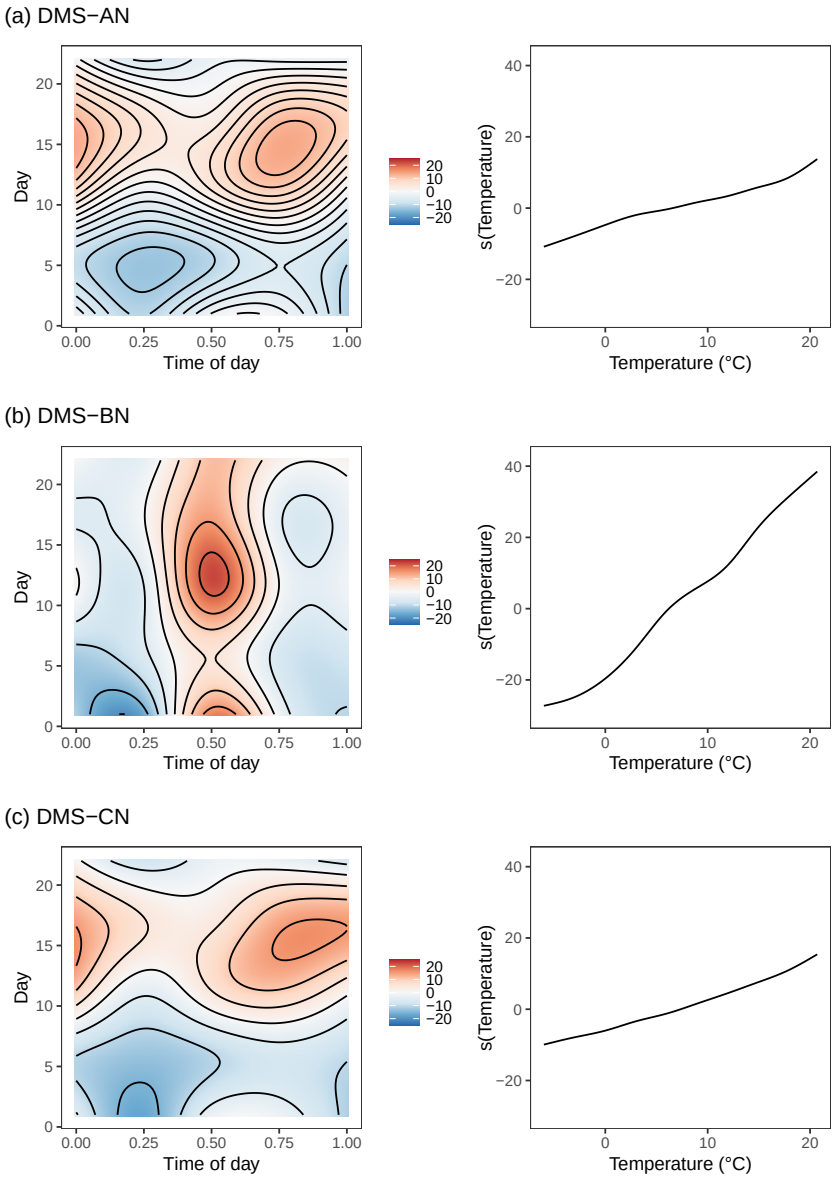


FIGURE 2. Results of the proposed functional modelling approach with two-dimensional functional intercept (left column) and the effect of temperature (right column) on strain sensor (a) DMS-AN, (b) DMS-BN and (c) DMS-CN; time of day was standardized to $[0, 1]$.

usual penalty on curvature. We see that the daily pattern (Figure 2, left) differs between the sensors. For DMS-BN (middle), the values of the strain sensor tend to be highest around the middle of the day and this effect is rather constant across days, whereas for DMS-AN (top) and DMS-CN (bottom), variation is rather across days than across time of the day. The (partial) effect of temperature (Figure 2, right) appears to be quite linear for all sensors, with the strongest effect found for sensor DMS-BN (middle).

Acknowledgments: This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

References

Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B., Vantini, S. (2021). Functional Regression Control Chart. *Technometrics*, **63**, 281 – 294.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M.W., Swihart, B., Xiao, L., Crainiceanu, C., Reiss, P.T. (2023). *refund: Regression with Functional Data*. R package version 0.1-30.

Han, Q., Ma, Q., Xu, J., Liu, M. (2021). Structural health monitoring research under varying temperature condition: a review. *Journal of Civil Structural Health Monitoring*, **11**, 149 – 173.

Jaelani, Y., Klemm, A., Wimmer, J., Seitz, F., Köhncke, M., Marsili, F., Mendler, A., von Danwitz, M., Keßler, S., Henke, S., Gündel, M., Braml, T. and Popp, A. (2023). Developing a Benchmark Study for Bridge Monitoring. Accepted for Publication in: *Steel Construction*.

Wood, S.N. (2017). *Generalized Additive Models*, 2nd ed., London: Chapman & Hall.

Analyzing blended learning education with eye tracking and deep learning methods

Hilal Yagimli [1], Julia Berginski [1], Alexander Silbersdorff [1]

[1] Georg-August Universität Göttingen, Germany

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

Abstract: This paper proposes a low-cost, non-intrusive eye tracker based on Convolutional Neural Networks (CNNs) for analyzing blended learning education, especially the effectiveness of the design of lecture videos. The input to our CNN model is an image of the eye region of a person looking at a desktop screen and the output is the predicted gaze point. Our network achieves an average prediction error of around 9.7 pixels on the train set, 44.3 pixels on the test set and 311.1 pixels for an unseen person on a $W1920 \times H1080$ pixels screen.

Keywords: Eye Tracking; Deep Learning; Blended Learning Education.

1 Introduction

Eye tracking involves capturing the movement of a person's eyes and predicting the gaze point. It is increasingly used in various fields, including engineering, epidemiology and psychology (David et al. 2021). Recently it has also been employed to analyze learning processes since learners take most information in through their eyes (Gruber et al. 2017). In particular, the use of blended learning tools and the development of digital educational resources yields an array of potential applications - both in terms of analyzing learning behavior and in terms of using insights gained by eye tracking to further learning outcomes.
As learning takes place in different places depending on the students themselves, a non-device-dependent eye tracking method is required, which works on laptops, tablets, and computers to gain insight into the real-life learning environment of students.
Remote Video Occulography (VOG), also called remote video-based eye tracking, uses images of eyes recorded with one or more video cameras to detect the gaze point. Remote eye trackers film the subject with a camera

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

attached to the desktop, such as laptop webcams, where the subject is performing a task or from a different but stationary position.
Thus, remote video-based eye trackers that process low-resolution images offer a non-intrusive approach for eye tracking. Additionally, they can be used on standard hardware with a webcam such as laptops and tablets, so there is no need for additional hardware. This offers a way to track the eye movement of students watching lecture videos or performing tasks on learning management systems at home or in their usual learning environment.
Remote video-based eye tracking can be subdivided into appearance-based and feature-based approaches (Villanueva et al. 2009). Feature-based approaches rely on geometric information of the eye. Feature-based eye trackers use features such as eye corners, contours, and reflections to determine the gaze point. Depending on the model, these approaches need a subject-specific calibration procedure, multiple (high-resolution) cameras, multiple lighting sources, and some image feature detectors to have high accuracy.
The other way to predict the gaze point with remote cameras is an appearance-based approach which only uses the images and no additional information to learn a mapping between the appearance of the eye and the gaze point (Liu et al. 2019).
Recently, appearance-based approaches increasingly rely on deep learning methods, and Convolutional Neural Networks (CNNs) are mainly used to predict the gaze point (Corcoran and Kar 2017). Using an appearance-based approach, we find that such eye trackers do not need specific or additional hardware, and low-resolution images such as those provided by webcams can be processed while performing sufficiently well for the purposes of detecting broad gazing points.

2 Eye Tracking for Educational Research

For educational research, an eye tracker can be used to analyze how students view lecture videos to further improve them by comparing fixation points and eye movement with the action in the video. In order to gain as realistic an insight as possible into the learning behavior of students, they need to be captured in their everyday learning environments and not under laboratory conditions. Thus, it is advisable to outsource the eye tracking process to the students' devices. Given the resultant variations in the recording conditions on the students' devices CNN-based eye trackers are best suited. Thus, input with different lighting conditions, head/eye positions, etc. can be contemplated in a flexible modelling manner by sufficient training episodes.
Naturally in adherence with the GDPR any recording must be by informed consent and is strictly voluntary. The sample may yield selection issues and those students that participate could follow the lecture more atten-

tively than they usually would, and this should be considered for further statistical analysis of the eye tracking data.
Concerning the technical requirements it should be noted that, depending on the number of participants, there will be a large amount of images to process and predict the gaze points, which will need considerable computational capacity and storage. Given only a very limited 2 minute video sequence used in our setting, we already require roughly 2.7 MB of storage capacity per student. While training the CNN for the whole sample, we required a computation time of several days using over a dozen cores. For larger sample sizes, both in terms of the number of participants and video duration, we thus are required to select sequences of interest and downsize the frame numbers used in the analysis.
Despite these challenges, we find that analyzing learners' activities while watching lecture videos with CNN-based eye trackers offers a feasible research method that provides deeper insight into the learning process.

3 Preliminary Methodology

3.1 Data

The data used to train the model is self-collected. The dataset contains pictures of ten students looking at a set of known point on the screen and the corresponding gaze point on the screen as the target variable. To collect these pictures, the students were recorded with a laptop webcam with 25 frames per second for 109 seconds yielding a total of 27250 frames for the dataset. Both for the training phase and the subsequent evaluation phase, we asked the students to restrict head movement to a minimum and to leave the lighting conditions unchanged. Furthermore, we also restricted the screen size to a constant size of $W1920 \times H1080$ pixels for this trial phase. The dot moved almost continuously over the screen with some jumps starting at the top left corner.
Last but not least, the recordings were pre-processed with the OpenCV Cascade Classifiers (Bradski, G. 2000) to find the face and eye regions and reduce the CNN-input to a $W230 \times H80$ pixel image of the eye region.

3.2 Model Architecture

The eye tracker model is displayed in Figure 1. It is a five-layer Convolutional Neural Network with Leaky ReLU as the activation function. The last layer is a fully connected layer that produces the two-dimensional output. A batch normalization and max pooling layer follow each convolutional layer.
Given the RGB encoded input picture of the eye regions, the input to the network is a $3 \times 80 \times 230$ dimensional array. The output is a two dimensional vector entailing the predicted gaze coordinate on the desktop.

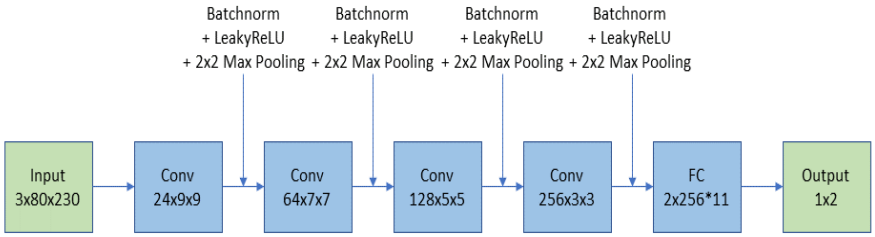


FIGURE 1. Architecture of eye tracking model proposed in this paper

3.3 Model Performance

The network was trained for 1000 epochs with a batch size of 128 and a learning rate of 0.01 using the Adam optimizer and the MSE loss as the error function. The data was split into 80% for training and 20% for testing. To evaluate the model performance not only on the test set, in which the same students from the train set are covered but on unseen data, the images of one student were taken out of the dataset before splitting. Provided that sufficient data for every student is available, we consider it feasible to calibrate CNNs for every person individually. Given the joint modelling of nine participants, the network achieves an average error of around 9.7 pixels on the train set, 44.3 pixels on the test set, and 311.1 pixels for an unseen person on a $W1920 \times H1080$ pixels screen.

4 Further Research

4.1 Methodology

For future work and to use the eye tracker for real-life situations where natural head movement and different lighting conditions should be allowed, our used dataset and our model will require some adjustments.
On the one hand, the dataset should include images of people in different lighting conditions and from different angles and distances since the desktop position can vary a lot depending on the device used. This will also lead to different image sizes, so we pursue additional pre-processing steps that allow for the handling of more heterogeneous input images.
On the other hand, more people and more data are needed to train a more robust network or individualized adjustments are needed. So, the diversity of how human beings and their eye regions can look should be covered in the dataset. The GazeCapture dataset could be used for this. It was introduced by Krafka et al.(2016) and contains data from over 1450 people with almost 2.5M images.
Additionally, our network should handle situations where the subject does not look at the screen, e.g., while taking notes, and in general, head movement should be handled. We plan to do this by extending the dataset and

training a larger and deeper model. This could also be done by additionally giving the network information about landmark points from the face as Liu G., Odobez J. and Yu, Y. (2019) did in their paper, but this requires a second model which can detect those landmark points. This will also require more computational capacity, especially in the training phase.
Furthermore, the proposed network was only tested on one desktop. This could also lead to difficulties when working with different devices and needs to be researched.
Lastly, instead of using only the eye regions as the input, we will try to process the whole face image with the expectation that the network learns different head positions without additional information.

4.2 Further Analysis

In future work, we plan to use the eye tracker for analyzing students learning behaviors in order to improve lecture videos according to students' needs and identify the learning strategies of successful students, measuring success based on their performance on tasks to be solved during and after the lecture video and their exam grade.
Therefore, we aim to analyze whether students are looking in the areas we would expect based on the current content, if not, we analyze where they are looking and whether nudging students' attention while watching the videos can improve learning outcomes. Furthermore, we aim to explore whether their gazing behavior can allow for one of the following two educative insights:
If, on the one hand, several students occur to be distracted in the same part of the lecture, this can be an indication that this part should be revised.
On the other hand, the additional consideration of information from so-called logfiles may give us insights into the learning process of the students. In learning management systems such as Moodle and ILIAS, data on user behavior is stored in these logfiles (Park and Petri 2021). We can gain information about students' learning behavior by analyzing how often they watch the lecture, whether they additionally use the script and, if tasks are offered, how they solve them after watching the lecture, and compare this with the eye tracking data.
Finally, with the information about learning behaviors and corresponding performance on assignments and the exam, we can build a model to predict the success of other students in future semesters, identify knowledge gaps, and thus provide appropriate individualized support.

5 Conclusion

In this paper, a CNN model for predicting the gaze point of students in an educative setting is outlined and underwent a preliminary test. Using

such models, we aim to explore the students' behavior in blended learning tools to analyze and further the effectiveness of lecture videos for tertiary education and to further the competencies of future student generations using the scopes of artificial intelligence.

References

Bradski, G. (2000). The OpenCV Library. In: *Dr. Dobb's Journal of Software Tools.*

Corcoran, P. and Kar, A. (2017). A Review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms. In: *IEEE Access 5*, 16495 -– 16519.

David, F., Kanade, P., and Kanade, S. (Apr. 2021). Convolutional Neural Networks (CNN) based Eye-Gaze Tracking System using Machine Learning Algorithm. In: *European Journal of Electrical Engineering and Computer Science* **5**, 36 -– 40.

Gruber, H., Holmqvist, K., and Jarodzka, H. (Feb. 2017). Eye tracking in Educational Science: Theoretical frameworks and research agendas. In: *Journal of Eye Movement Research* **10**.

Krafka, K. et al. (2016). Eye Tracking for Everyone. In: *CoRR abs/1606.05814*.

Liu, G., Odobez, J., Yu, Y. (2019). Deep Multitask Gaze Estimation with a Constrained Landmark-Gaze Model. In: *Computer Vision – ECCV 2018 Workshops.* Ed. by Laura Leal-Taix é and Stefan Roth. Cham: Springer International Publishing, 456 — 474.

Park, S. and Petri, P. (June 2021). Was Logfiles uns verraten können – Nutzung multipler Datenquellen zur Evaluation digitaler Lehre. In: *69*, 31 -– 38.

Villanueva, A. et al. (Nov. 2009). A geometric approach to remote eye tracking. In: *Universal Access in the Information Society* **8.4**, 241 -– 257.

Crime prediction models in the metropolitan area of São Paulo - Brazil

Wellington Yuanhe Zhao[1], Luis Gustavo Nonato[2], Cibele M. Russo[2]

[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 Applied Mathematics and Statistics, University of São Paulo, São Carlos, Brazil

E-mail for correspondence: wellington.zhao@usp.br

Abstract: Public security is a main challenge for Brazilian society, as crime is a major problem in large Brazilian cities such as São Paulo. An important issue in this context is to model and predict crime patterns considering historical data from each particular spatial location. In this work, we evaluate and compare different prediction models such as spatial autoregressive (SAR) and artificial neural network models (ANN). For ANN models we use as input data ranging from population, economic, and education indications to historical data of crimes in each geolocation. The SAR model takes into account the covariates, as well as the underlying spatial dependence of the data.

Keywords: Crime modelling; Geo-spatial data; Temporal data.

1 Introduction and Exploratory Analysis

Urban crime is one of the most critical social problems worldwide, being even more prevalent in Latin America's big cities. Particularly, in the metropolitan area of São Paulo, urban crimes vary substantially in intensity and type of occurrences depending on the characteristics and geolocation of each city. This scenario demands a systematic investigation of the determinant factors that lead to such variability.

In order to investigate the factors that most impact the intensity and type of crime over the different cities in metropolitan area of São Paulo, we build statistical and machine learning models from multiple sources of

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

data. Specifically, crime data were obtained from the open repository of the São Paulo State Department of Public Safety - *Secretaria de Segurança Pública de São Paulo, 2021*, in Portuguese, with data available up until the year 2021. We combined all types of crimes into a single variable called "total_of_crime", which will be the predictive variable for the models. To avoid bias due to the population size, we divided the "total of crime" by the city's population, generating the "crime_per_capita" indicator. Figure 1 depicts the distribution of "crime_per_capita" in the metropolitan area of the city of São Paulo.

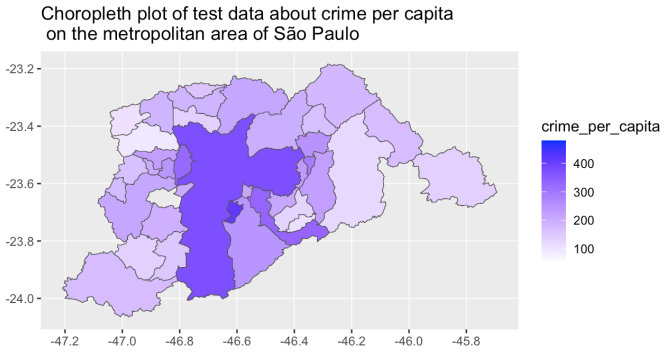


FIGURE 1. Choropleth plot of "crime_per_capita" in the metropolitan area of the city of São Paulo. The metropolitan area comprise the cities with the higher crime rate in the State. The x and y axes show latitudes and longitudes, respectively.

For illustration purposes, Figure 2 shows a correlogram of different types of crimes in the metropolitan area of São Paulo. It shows that there is a strong correlation between different types of rape and murder, as well as a relatively strong correlation between murder, assault, and rape. In this work, for initial analyses, we considered the total crime per capita in the metropolitan area of São Paulo.

2 SAR model

The SAR (Spatial Autoregressive) model (see, for instance, Kazar and Celik 2012), also known in the literature as spatial lag model or mixed regressive model, is an extension of the linear regression model, given by the equation

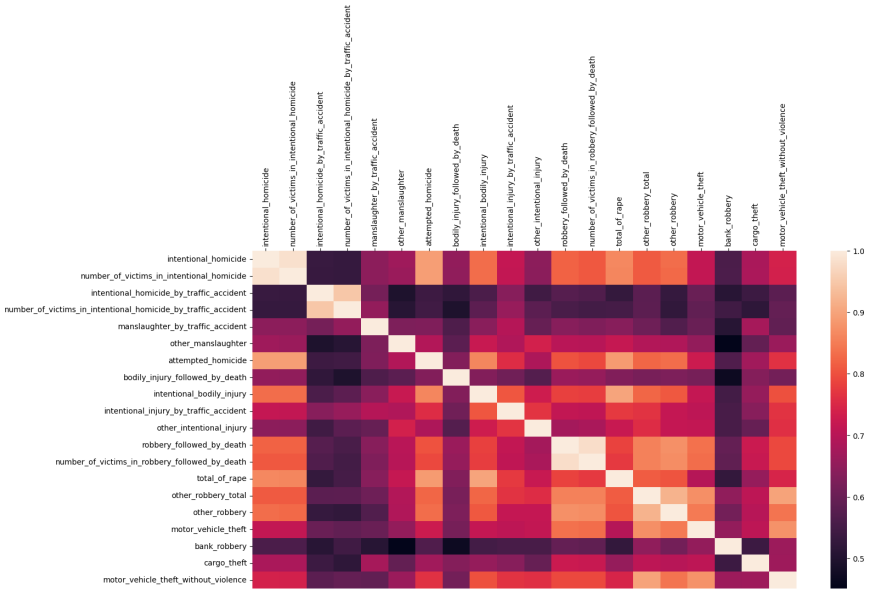


FIGURE 2. Correlogram's plot of different type of crimes in the metropolitan area of São Paulo.

$$y_i = \rho \sum_{j=1}^{n} W_{ij} y_j \sum_{q=1}^{Q} X_{iq} \beta_q + \epsilon_i, \text{ where} \tag{1}$$

ρ is the spatial autoregression parameter, W is an n-by-n neighborhood matrix that accounts for the spatial relationships among the spatial data. ϵ_i is independently and identically distributed with mean zero and variance σ^2, y_i is the dependent variable, X_{iq} are independent variables, β_q is the coefficient related to X_{iq} and n is the number of observations. The index i indicates the n successive observations in the municipality, with data observed between 2002 to 2017 and Q the number of independent variables. A derivative-based optimization that can be used in SAR model solution is Newton-Raphson (root-finding) algorithm. In this algorithm, we need to compute the first derivative of the log-likelihood function. This gives the location of optimal solution for ρ parameter. The first derivative of log-likelihood function is given by the following equation.

$$\frac{\partial lnL(y)}{\partial \rho} = tr((I - \rho W)^{-1} \frac{\partial (I - \rho W)}{\partial \rho})) - \tag{2}$$

$$\frac{n}{2}(\frac{-y^T M^T M W y - y^T W^T M^T M y + 2\rho y y^T W^T M^T M W y}{y^T (I - \rho W)^T M^T M (I - \rho W) y}) \tag{3}$$

The term M corresponds to $[I - X(X^TX)^{-1}X^T]$, where X is the design matrix that contains the predictive variables.

3 Artificial neural networks

Artificial neural network models are often used for regression or classification tasks. This type of model is inspired by the structure and functioning of the human brain and it uses layers and neurons to transform the information into predictions (see, for instance, Hastie, Tibshirani and Friedman 2009).
In this paper, we relied on an ANN whose first layer has 32 units (neurons) with the ReLU activation function. The second and third layers have 16 units each and they also use ReLU activation. Dropout regularization is applied after each of the first two dense layers to prevent overfitting. The output layer has one unit and no activation function, as we are handling crime prediction as a regression problem. The model's loss function is the mean squared error and we use the Adam optimizer with a learning rate of 0.001. Finally, the model is trained on the normalized training data for 50 epochs with a batch size of 32 and a validation split of 0.2.

4 Application

The data utilized in this study extends beyond the domain of public security and encompasses various facets such as economy, education, and geographic information from each city, with a total of 102 available variables. Data was divided into training and test sets for predicting-focused analysis. The training data consisted of crimes per capita between 2002 and 2017 with 123,000 observations, while the test data the crimes in 2018 with 7,728 observations. The crime per capita was computed by dividing the total crime by the population size. To account of certain intrinsic characteristics of the data, for instance São Paulo and Guarulhos having the highest values across all variables, a logarithmic function was applied to each variable.
Subsequently, the SAR model was applied to the data, resulting in the findings presented in Table 1. To optimize space, we have used an abbreviated notation, where X_1 denotes the total crime 1m lag, X_2 represents the total crime 2m lag, X_3 corresponds to the population, X_4 denotes the population of age 60+, X_5 represents the registration at the municipal pre-school, X_6 denotes the registration at the state elementary school (beginning), X_7 represents the registration at the municipal elementary school (final), and finally, X_8 corresponds to the registration of municipal high school.
An artificial neural network model was also fitted. The main difference between the fitted SAR and ANN covariates lies in the variables used and the consideration of geographical factors. SAR models utilize variables like "total crime 1m lag" and "total crime 2m lag" as explanatory factors, whereas

TABLE 1. Fitted SAR model for crimes at the region of greater São Paulo after applying the logarithmic function to each variable.

Variables and Intercept	Estimate	Std. Error	z value	p-value
intercept	3.122	0.153	20.396	<0.001
X_1	0.004	0.002	2.670	0.008
X_2	0.006	0.001	5.180	<0.001
X_3	-0.740	0.028	-26.082	<0.001
X_4	0.786	0.027	28.918	<0.001
X_5	0.044	0.008	5.785	<0.001
X_6	-0.003	0.001	-3.388	0.001
X_7	0.008	0.001	10.103	<0.001
X_8	-0.014	0.001	-11.437	<0.001
ρ	0.506			<0.001
σ^2	0.090			

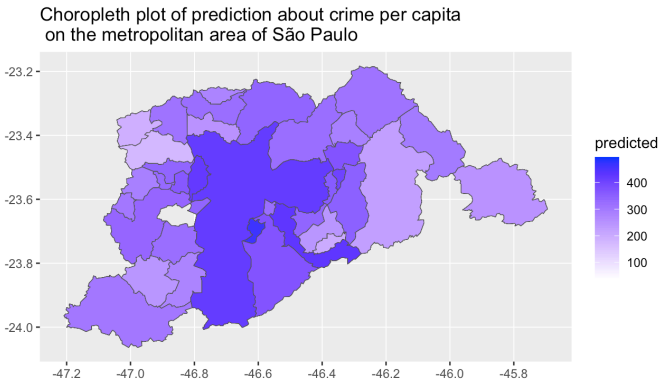


FIGURE 3. Choropleth plot of crime per capita in the metropolitan area of São Paulo, based on the predicted data.

ANN models exclude those variables. Additionally, ANN models incorporate the populational density as a covariate to account for geographic influences. This distinction allows ANN models to capture spatial dependencies and variations in crime patterns more effectively compared to SAR models. Based on the test data, the results in Table 2 demonstrate the superior predicting performance of the ANN model when compared to SAR model for predicting crimes in the metropolitan area of São Paulo state.

To evaluate the SAR model, a simulation study was conducted by fixing theoretical values for the fixed coefficients, similar to those estimated for

TABLE 2. Predictive performance comparison between SAR model and artificial neural network model.

Measure	SAR model	Artificial neural network
RMSE	0.3503	0.0016
MAE	0.3022	0.0349

São Paulo crime data, as well as combinations of $\sigma^2 \in (0.01, 0.1)^\top$ and $\rho \in (0.1, 0.3, 0.7)^\top$. The simulation study assessed the bias and mean squared error of the estimates and allowed verifying an adequate recovery of the theoretical parameters used, which shows that the SAR model correctly estimates the regression coefficients. The detailed results will be omitted here due to space limitations in the abstract.

Finally, based on the artificial neural networks model, a choropleth plot illustrates the predicted crime per capita in the metropolitan area of São Paulo (Figure 3).

As depicted in Table 2, the artificial neural network demonstrates superior predicting performance when compared to the estimated SAR model. However, it is important to note that, as a regression model, the SAR model can estimate the relationship between the response variable and the explanatory variables, thus providing interpretability to the regression coefficients. Additionally, the SAR model enables the computation of p-values when considering the marginal significance of each parameter.

Acknowledgments: The authors W. Y. Zhao, L. G. Nonato and C. M. Russo acknowledge the finantial support from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and FAPESP eScience and Data Science Project 2022/09091-8.

References

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

Kazar, B. M and Celik, M. (2012). *Spatial autoregression (SAR) model: Parameter estimation techniques.* Springer Science & Business Media.

Secretaria de Segurança Publica de São Paulo (SSP) (2021). *Dados estatísticos da Secretaria de Segurança Pública do Estado de São Paulo. In Portuguese.* Available at https://basedosdados.org/dataset/dbd717cb-7da8-4efd-9162-951a71694541?table=a2e9f998-e2c2-49b7-858a-ae1daef46dc0

A scalable and embedded diachronic sense change model

Schyan Zafar[1], Geoff K. Nicholls[1]

[1] Department of Statistics, University of Oxford, Oxford, UK

E-mail for correspondence: schyan.zafar@jesus.ox.ac.uk

Abstract: Meanings or *senses* of a word like 'bank' (riverbank or institution) change over time. The recent DiSC model measures temporal word-sense changes via careful statistical modelling to quantify uncertainty in the sense-change estimates. We introduce EDiSC, an embedded version of DiSC, which combines word embeddings and DiSC to improve model performance, both in terms of accuracy and sampling efficiency of MCMC methods, via embedded representations of senses. The resulting model is better adapted to scaling up for larger datasets.

Keywords: Sense change models; Word embeddings; Bayesian inference.

1 Introduction

Diachronic lexical semantics is the study of temporal meaning change in languages. Areas of interest include words with multiple meanings or *senses*, e.g. 'mouse' (a rodent or a computer pointing device). Computational models of meaning change aid in this study. The recent DiSC model introduced in Zafar and Nicholls (2022), building on the framework of related earlier models, represents distinct senses of a target word as distinct distributions over context words, and sense prevalence as a distribution over senses. Given a set of unlabelled text snippets, we use careful statistical modelling to fit the data, predict the target-word sense in each snippet, and obtain credible intervals for the evolving senses and sense prevalence.
Using analogous ideas to Dieng et al. (2019), we now introduce EDiSC, an embedded version of DiSC, by combining it with word embeddings, whereby context words are represented as vectors in an embedding space. This has two main advantages over DiSC. Firstly, embeddings exploit the wider text corpus to capture useful semantic information about the context words, which is otherwise lost if we focus only on the context of a given target

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

word. Secondly, the dimension of the embedding space is typically much lower than the vocabulary size. These two features lead to, respectively, improved accuracy and more efficient Monte Carlo sampling as the data size is scaled up. We demonstrate this on real test cases from English and ancient Greek as well as synthetic data.

2 Notation and Models

At time $t \in \{1, \ldots, T\}$, target-word sense $k \in \{1, \ldots, K\}$ is represented as a distribution $\tilde{\psi}^{k,t}$ over context words $1, \ldots, V$, and sense-prevalence for genre $g \in \{1, \ldots, G\}$ is represented as a distribution $\tilde{\phi}^{g,t}$ over senses $1, \ldots, K$. Our goal is to infer $\tilde{\psi}$ and $\tilde{\phi}$ given the data W comprising D snippets. Each snippet $d \in \{1, \ldots, D\}$ is a context of L_d words around the target word, with genre label γ_d and time label τ_d. If true sense labels $o_1, \ldots, o_D$ are available, we refer to a subset $\{1', \ldots, D'\}$ as 'collocates', i.e. snippets where a human could identify the sense from context alone.

Both DiSC and EDiSC are generative bag-of-words models, comprising the same observation model but different prior models. Under the observation model, each snippet d is generated independently by first sampling the sense $z_d|\tilde{\phi}^{\gamma_d,\tau_d} \sim \text{Mult}(\tilde{\phi}^{\gamma_d,\tau_d})$ and then sampling the words given the sense $w_{d,i}|z_d, \tilde{\psi}^{z_d,\tau_d} \sim \text{Mult}(\tilde{\psi}^{z_d,\tau_d})$ independently for each position $i \in \{i_1, \ldots, i_{L_d}\}$.

Under both prior models, probability vectors $\tilde{\psi}^{k,t}$ and $\tilde{\phi}^{g,t}$ are softmax transforms of real vectors $\psi^{k,t}$ and $\phi^{g,t}$ respectively, i.e. $\tilde{\psi}^{k,t} = \frac{\exp(\psi^{k,t})}{\sum_{v=1}^{V}\exp(\psi_v^{k,t})}$ and $\tilde{\phi}^{g,t} = \frac{\exp(\phi^{g,t})}{\sum_{k=1}^{K}\exp(\phi_k^{g,t})}$. The prior on $\phi^{g,t}$, for each g, is an AR(1) time series with stationary distribution $\mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\phi}{1-(\alpha_\phi)^2}\right)\right)$. Under DiSC, $\psi^{k,t} = \chi^k + \theta^t$, with a $\mathcal{N}(0, \text{diag}(\kappa_\chi))$ prior on χ^k and an AR(1) prior on θ^t with stationary distribution $\mathcal{N}\left(0, \text{diag}\left(\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)\right)$. In contrast, under EDiSC, we now define $\psi^{k,t} = \rho\xi^{k,t} + \varsigma$, where ρ is a $V \times M$ matrix of context-word embeddings learnt using GloVe (Pennington 2014), $\xi^{k,t}$ is an M-dimensional sense-time embedding, and ς is a V-dimensional bias or correction parameter with prior $\mathcal{N}(0, \text{diag}(\kappa_\varsigma))$. ξ is decomposed as $\xi^{k,t} = \chi^k + \theta^t$, where χ^k and θ^t are M-dimensional sense and time embeddings respectively. We place priors on χ^k and θ^t, functionally the same as in DiSC, whilst noting that these are now vectors in the M-dimensional embedding space rather than the larger V-dimensional space in DiSC. The full EDiSC generative model is given in Table 1.

3 Data and Evaluation

We have a simple test case from COHA (Davies 2010) annotated by Zafar and Nicholls (2022):

TABLE 1. EDiSC generative model

PRIOR MODEL

get word embeddings matrix ρ
fix hyperparameters $\kappa_\phi, \kappa_\theta, \kappa_\chi, \kappa_\varsigma, \alpha_\phi, \alpha_\theta$ (with $|\alpha_\phi| < 1, |\alpha_\theta| < 1$)
draw bias or correction parameter $\varsigma|\kappa_\varsigma \sim \mathcal{N}(0, \mathrm{diag}(\kappa_\varsigma))$
initialise at time $t = 1$
  for genre $g \in 1 : G$ **do**
    draw sense prevalence parameter $\phi^{g,1}|\kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(0, \mathrm{diag}\left(\frac{\kappa_\phi}{1-(\alpha_\phi)^2}\right)\right)$
  end for
  draw time embedding $\theta^1|\kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(0, \mathrm{diag}\left(\frac{\kappa_\theta}{1-(\alpha_\theta)^2}\right)\right)$
for time $t \in 2 : T$ **do**
  for genre $g \in 1 : G$ **do**
    draw sense prevalence parameter $\phi^{g,t}|\phi^{g,t-1}, \kappa_\phi, \alpha_\phi \sim \mathcal{N}\left(\alpha_\phi \phi^{g,t-1}, \mathrm{diag}(\kappa_\phi)\right)$
  end for
  draw time embedding $\theta^t|\theta^{t-1}, \kappa_\theta, \alpha_\theta \sim \mathcal{N}\left(\alpha_\theta \theta^{t-1}, \mathrm{diag}(\kappa_\theta)\right)$
end for
for sense $k \in 1 : K$ **do**
  draw sense embedding $\chi^k|\kappa_\chi \sim \mathcal{N}(0, \mathrm{diag}(\kappa_\chi))$
end for
for sense $k \in 1 : K$ and time $t \in 1 : T$ **do**
  set sense-time embedding $\xi^{k,t} = \chi^k + \theta^t$
  set context-word probability parameter $\psi^{k,t} = \rho \xi^{k,t} + \varsigma$
end for
transform ϕ and ψ into probabilities $\tilde{\phi}$ and $\tilde{\psi}$ using softmax

OBSERVATION MODEL

fix probabilities of drawing stopwords q^{SW} and uninformative words q^{U}
for snippet $d \in 1 : D$ **do**
  draw number of context words $L_d|L, q^{\mathrm{SW}}, q^{\mathrm{U}} \sim \mathrm{Bin}(L, 1 - q^{\mathrm{SW}} - q^{\mathrm{U}})$
  draw a random subset $\{i_1, \ldots, i_{L_d}\}$ of size L_d from $\{1, \ldots, L\}$
  draw sense assignment $z_d|\tilde{\phi}^{\gamma_d, \tau_d} \sim \mathrm{Mult}\left(\tilde{\phi}_1^{\gamma_d, \tau_d}, \ldots, \tilde{\phi}_K^{\gamma_d, \tau_d}\right)$
  for context position $i \in \{i_1, \ldots, i_{L_d}\}$ **do**
    draw context word $w_{d,i}|z_d, \tilde{\psi}^{z_d, \tau_d} \sim \mathrm{Mult}\left(\tilde{\psi}_1^{z_d, \tau_d}, \ldots, \tilde{\psi}_V^{z_d, \tau_d}\right)$
  end for
end for

'bank' (riverbank or institution) [$D = 3\,685, D' = 3\,525, V = 973, K = 2, G = 1, T = 10$]; and three more challenging test cases from ancient Greek (Vatri et al. 2018, 2019): **'kosmos'** (decoration, order, world) [$D = 1\,469$, $D' = 1\,144, V = 2\,904, K = 3, G = 2, T = 9$], **'mus'** (mouse, muscle, mussel) [$D = 214, D' = 118, V = 899, K = 3, G = 2, T = 9$], and **'harmonia'** (abstract, concrete, musical) [$D = 653$, $D' = 451$,

$V = 1\,607, K = 3, G = 2, T = 12$].

The model posteriors are defined by $\pi(\phi, \psi|W) \propto \pi(\phi)\pi(\psi)p(W|\phi, \psi)$, where the likelihood $p(W|\phi, \psi) = \prod_{d=1}^{D} \sum_{k=1}^{K} \tilde{\phi}_k^{\gamma_d, \tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{k, \tau_d}$ is common to both DiSC and EDiSC, since the models differ only in the prior structure for ψ. We infer the posteriors $\phi, \psi|W$ using MCMC with HMC and MALA sampling.

Given true sense labels o for the collocates, we assess model predictive accuracy using the Brier score $BS = \frac{1}{D} \sum_{d=1'}^{D'} \sum_{k=1}^{K} \left(\hat{p}(z_d = k) - \mathbb{I}(o_d = k)\right)^2$, a proper scoring rule for multi-category probabilistic predictions $\hat{p}(z_d = k)$, ranging from 0 (best) to 2 (worst). Here, $\hat{p}(z_d = k)$ is the estimated value of $\mathbb{E}_{\phi,\psi|W}\big(p(z_d = k|W_d, \phi, \psi)\big)$ computed by normalising $\tilde{\phi}_{z_d}^{\gamma_d, \tau_d} \prod_{i=i_1}^{i_{L_d}} \tilde{\psi}_{w_{d,i}}^{z_d, \tau_d}$ over $z_d \in \{1, \ldots, K\}$ using the MCMC output. Table 2 shows the Brier scores for DiSC and EDiSC. In the case of 'bank', 'kosmos' and 'mus', EDiSC with an appropriately chosen embedding dimension M offers a clear improvement over DiSC. However, for 'harmonia', both models fail to converge to any meaningful senses.

Table 2 also shows the WAIC (Watanabe 2010, Vehtari et al. 2017) for the different models and data. In general, when the true labels are not available, we can select the model that minimises the WAIC. In our experiments, using the WAIC for model selection results in the optimal or near-optimal model based on Brier scores.

TABLE 2. Brier scores and WAIC for test data using different models

	'bank'		'kosmos'		'mus'	
	BS	WAIC	BS	WAIC	BS	WAIC
DiSC	0.150	154782	0.371	138869	0.204	20058
EDiSC ($M = 50$)	0.140	154682	0.352	137243	0.135	19434
EDiSC ($M = 100$)	0.139	154440	<u>0.327</u>	136866	<u>0.093</u>	<u>19417</u>
EDiSC ($M = 200$)	<u>0.133</u>	<u>154165</u>	0.332	<u>136510</u>	0.099	19450

We assess true-model recovery. True sense-prevalence is unknown. However, we can use independent well-informed estimates $\tilde{\phi}|(z = o)$ of the ground truth given the *labelled* data $o_d, d \in \{1, \ldots, D\}$ for assessment. Model posteriors $\tilde{\phi}|W$ given the *unlabelled* data W are compared against these independent estimates of the ground truth. Figure 1 shows the comparison for 'kosmos'. Whilst both models perform well, EDiSC generally does better: the EDiSC $\tilde{\phi}|W$ HPD intervals (blue bars) have higher overlap with the $\tilde{\phi}|(z = o)$ HPD intervals (dashed bars) compared to DiSC (red bars), indicating better ground truth recovery, and the EDiSC posterior means (blue circles) are also generally closer to the $\tilde{\phi}|(z = o)$ posterior means (black circles). We see similar results for the other test cases.

We assess the MCMC sampling efficiency for DiSC and EDiSC on the 'bank' data, using the effective sample size (ESS) per hour of CPU time

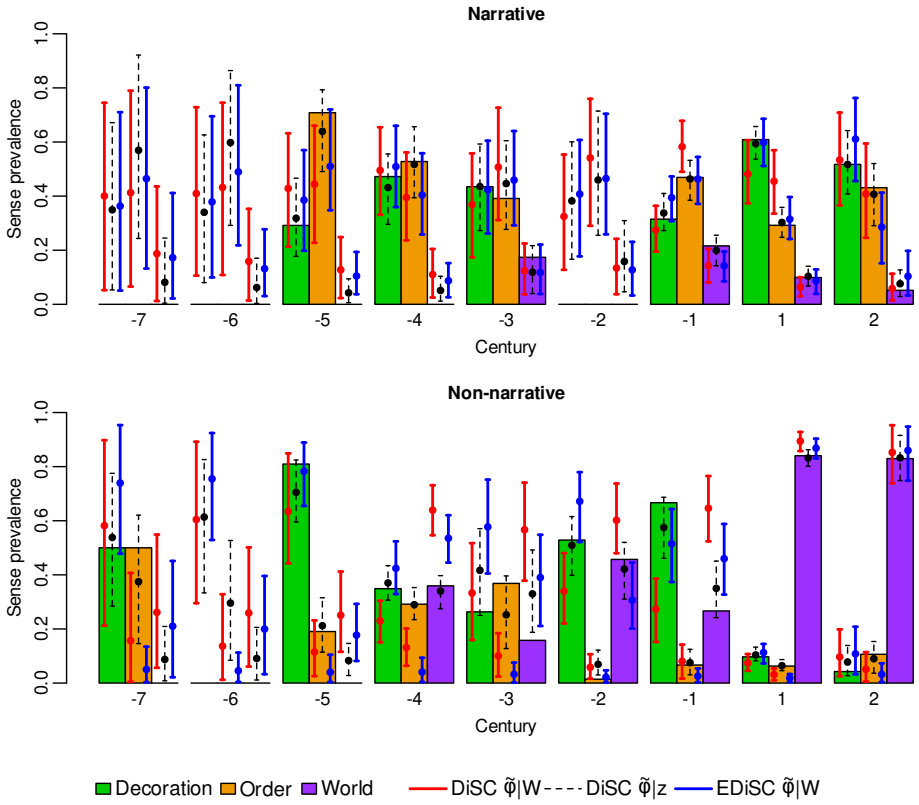


FIGURE 1. 'Kosmos' 95% HPD intervals (error bars) and posterior means (circles). Empirical sense prevalence (coloured bars) also shown for interest.

TABLE 3. Median ESS per hour of CPU time from MALA MCMC sampling

Model	Burn-in* (sims)	ESS for $\tilde{\phi}$	ESS for $\tilde{\psi}$
DiSC	700	375	391
EDiSC ($M = 50$)	100	1,916	391
EDiSC ($M = 100$)	250	2,192	344
EDiSC ($M = 200$)	500	2,237	303

* Defined as the approximate MCMC sample after which the trace plots for the variables of interest appear flat

as the metric. We see that, whilst the ESS for $\tilde{\psi}$ is of the same order, the ESS for $\tilde{\phi}$ is many times better under EDiSC than under DiSC. EDiSC also benefits from smaller burn-in times. Table 3 shows the results.
Finally, using synthetic data experiments, we see in Figure 2 that, for vocab-

ulary $V > 500$ (which is typical), MCMC run times increase with snippets D much more slowly for EDiSC than for DiSC. Thus, EDiSC is a lot better suited than DiSC to scaling up for larger and more complex data.

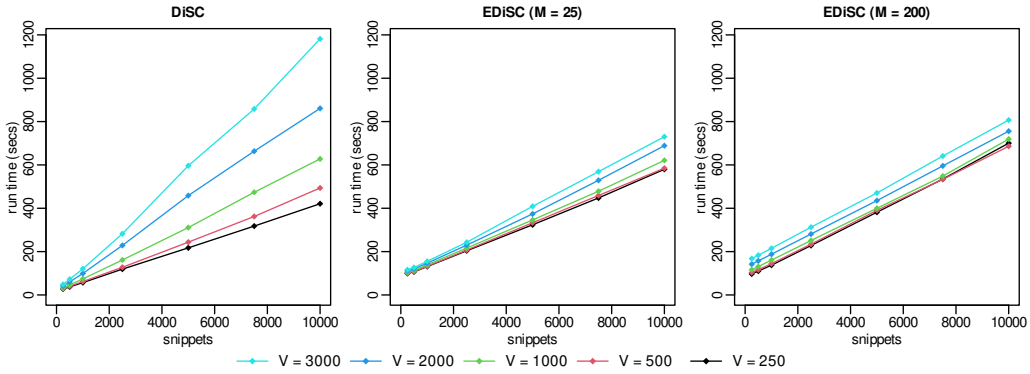


FIGURE 2. Mean run times in CPU seconds for 500 MCMC samples on synthetic data using different models, vocabulary sizes (V) and number of snippets (D)

References

Davies, M. (2010) The Corpus of Historical American English: 400 million words, 1810-2009.

Dieng, A.B., Ruiz, F.J., and Blei, D.M. (2019) The Dynamic Embedded Topic Model. *arXiv:1907.05545*.

Vatri, A. and McGillivray, B. (2018) The Diorisis Ancient Greek Corpus. *Research Data Journal for the Humanities and Social Sciences*, **3**(1): 55 – 65.

Vatri, A., Lähteenoja, V., and McGillivray, B. (2019) Ancient Greek semantic change – annotated datasets and code. *figshare.*

Vehtari, A., Gelman, A., and Gabry, J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**(5): 1413 – 1432.

Watanabe, S. (2010) Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, **11**(116): 3571 – 3594.

Zafar, S. and Nicholls, G.K. (2022) Measuring Diachronic Sense Change: New Models and Monte Carlo Methods for Bayesian Inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **71**(5): 1569 – 1604.

Bayesian nonparametric inference for the three-class covariate-specific overlap coefficient

Zhaoxi Zhang[1], Vanda Inácio[1]

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

E-mail for correspondence: Z.Zhang-156@sms.ed.ac.uk

Abstract: We extend the concept of overlap coefficient, and its conditional counterpart, to the case where there are three disease stages. We propose to estimate the covariate-specific overlap coefficient using a Bayesian nonparametric covariate-dependent mixture model that relies on a logit stick-breaking prior formulation. Our methods are motivated by an application to the diagnosis of Alzheimer's disease where the goal is to study how the accuracy of a potential biomarker for distinguishing between subjects with normal cognition, mild cognitive impairment, and dementia, changes with age.

Keywords: Diagnostic test; Overlap coefficient; Covariate-dependent mixture model; Logit stick-breaking prior; Alzheimer's disease.

1 Introduction

Evaluating the performance of diagnostic tests is of great importance in public health, clinical practice, and medical research. Before a diagnostic marker is approved for use in clinical practice, its ability of making diagnostic classification must be rigorously assessed through statistical analysis. In the two-class case, the major goal of a diagnostic marker is to classify subjects into a diseased or a nondiseased group. However, an intermediate transitional stage usually exists prior to disease onset in the process of several diseases, which is especially true for neurological disorders. Traditional summary measures of the diagnostic accuracy in the three-class setting, such as the volume under the ROC surface and the Youden index, assume an order between the three classes and their respective conditional counterparts, assume that the same order is maintained across all covariates levels, which may not hold in practice. In this work, we extend the

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023.

concept of overlap coefficient to the three-class disease case. In this setting, the overlap coefficient is defined as the proportion of area between the three density functions in each disease stage group. An advantage of the overlap coefficient over the volume under the surface and the Youden index, besides its intuitive interpretation, is that in addition of taking into account both the location and shape of the distributions of test outcomes in the two populations, it is 'non-directional', meaning that it does not need to assume an order between the three disease classes. We further propose a flexible Bayesian model for estimating the three-class covariate-specific overlap coefficient that relies on estimating the conditional density in each group using a covariate-dependent mixture model that relies on a logit stick-breaking prior for the mixing measure. The resulting model is widely applicable to a wide range of continuous diagnostic tests and for a wide range of diseases.

2 Bayesian nonparametric inference for the three-class covariate-specific overlap coefficient

Covariates (e.g.,age and gender) can impact the performance of a diagnostic test 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 overlap coefficient. Let Y_1, Y_2, and Y_3 be three independent continuous random variables representing the diagnostic test outcomes in the normal cognition, mild impairment, and dementia group, with covariate vectors $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$. For a given covariate vector value $\mathbf{x}$, the covariate-specific overlap coefficient is defined as:

$$\mathrm{OVL}(\mathbf{x}) = \int_{-\infty}^{+\infty} \min\left\{f_1(y \mid \mathbf{X}_1 = \mathbf{x}), f_2(y \mid \mathbf{X}_2 = \mathbf{x}), f_3(y \mid \mathbf{X}_3 = \mathbf{x})\right\} dy, \tag{1}$$

where $f_d(y \mid \mathbf{x})$ denotes the conditional (on $\mathbf{x}$) density of Y_d, for $d \in \{1, 2, 3\}$. Using the well known formula $min\{u, v\} = \frac{1}{2}(u + v) - \frac{1}{2}|u + v|$, the OVL expression in (1) can be shown to be equivalent to

$$\begin{aligned}\mathrm{OVL}(\mathbf{x}) =& 1 - \frac{1}{4}\int_{-\infty}^{+\infty} |f_1(y \mid \mathbf{x}) - f_2(y \mid \mathbf{x})| + |f_1(y \mid \mathbf{x}) + f_2(y \mid \mathbf{x}) - \\ & |f_1(y \mid \mathbf{x}) - f_2(y \mid \mathbf{x})| - 2f_3(y \mid \mathbf{x})| dy. \end{aligned} \tag{2}$$

When there is a complete overlap of the distributions of test outcomes in the three groups, thus corresponding to a useless diagnostic test, the overlap coefficient takes the value zero. On the other extreme case, when the three distributions are completely separated, the overlap coefficient is equal to one. Values between zero and one correspond to different degrees of overlap between the distributions of the test outcomes in the three

groups. Essentially, accurately estimating the covariate-specific overlap coefficient reduces to accurately estimate the conditional density function of the marker in each of the three disease groups. Within the Bayesian nonparametric framework, we consider the general class of covariate-dependent infinite mixture of normals model

$$f_d(y \mid \mathbf{x}) = \sum_{l=1}^{\infty} \omega_l(\mathbf{x}) \phi(y|\theta_l(\mathbf{x})), \quad d \in \{1, 2, 3\}, \tag{3}$$

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, 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 (2) can be approximated numerically by the trapezoidal rule.

3 Application

Our methods are motivated and applied to a dataset derived from the Alzheimer's Disease Neuroimaging Initiative. The dataset consistes 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 overlap coefficient, of the hypometabolic convergence index (HCI) to distinguish (simultaneously) between the three groups

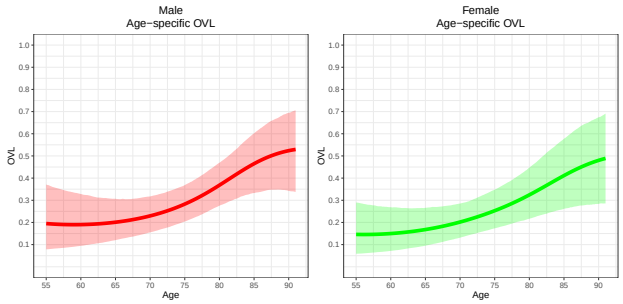


FIGURE 1. Posterior mean and 95% credible intervals for the age and gender specific overlap coefficient.

In Figure 1 we show the estimated age and gender specific overlap coefficient. We can see that the overlap coefficient generally increases with age for both genders, and the overlap coefficient of females is generally lower than that of males of the same age. Our results seem then to suggest that the diagnostic accuracy of HCI to simultaneously distinguish between subjects with normal cognition, mild impairment, and dementia decreases as age increases and its performance is slightly better for females than males.

References

Polson, Nicholas G., James G. Scott, and Jesse Windle (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables *Journal of the American statistical Association*, **108**, 1339 – 1349.

Rigon, Tommaso, and Daniele Durante (2021). Tractable Bayesian density regression via logit stick-breaking priors *Journal of Statistical Planning and Inference*, **211**, 131 – 142.

fakultät statistik

1973–2023

50 JAHRE STATISTIK

SMS
Statistical Modelling Society

IWSM
International Workshop on Statistical Modelling

Thanks to:

SIGNAL IDUNA

GdF Gesellschaft der Freunde der Technischen Universität Dortmund e.V.

ISBN: 978-3-947323-42-5