22nd European Young Statisticians Meeting – Proceedings

Published by: Department of Psychology & Department of Sociology, School of Social Science, 136, Syggrou Ave 17671 Athens, Greece

For publisher: Panteion University of Social and Political Sciences

Editors: Andreas Makridis, Fotios S. Milienos, Panagiotis Papastamoulis, Christina Parpoula & Athanasios Rakitzis

Place of publication: Athens, Greece

Year of publication: 2021

ISBN: 978-960-7943-23-1

22nd European Young Statisticians Meeting

6 – 10 September 2021, Athens, Greece

Proceedings

Eds. Andreas Makridis, Fotios S. Milienos, Panagiotis Papastamoulis, Christina Parpoula & Athanasios Rakitzis

Bernoulli Society
for Mathematical Statistics
and Probability

Πάντειον Πανεπιστήμιο Κοινωνικών & Πολιτικών Επιστημών
Panteion University of Social and Political Sciences

Preface

European Young Statisticians Meetings are organized every two years under the auspices of the European Regional Committee of the Bernoulli Society for Mathematical Statistics and Probability. The aim is to provide a scientific forum for the next generation of European researchers in probability theory and statistics. It represents an excellent opportunity to promote new collaborations and international cooperation. Participants are less than 30 years old or have 2 to 8 years of research experience, and are invited on the basis of their scientific achievements, in a uniformly distributed way in Europe (at most 2 participants per country). The International Organizing Committee (IOC) is responsible for their selection.

There were thirty two European countries participating at the 22nd EYSM. The scientific part of the Conference was organized as follows:

- [-] five eminent scientists from the field of mathematical statistics and probability gave 60-minutes keynote lectures
- [-] fifty seven invited young scientists gave 20-minutes lectures.

The topics presented include, but are not limited to:

- Applied statistics in biology, medicine, etc.
- Bayesian inference
- Change-point detection
- Characterizations of probability distributions
- Extreme and record value theory
- Functional statistics
- Goodness-of-fit testing
- High-dimensional statistics
- Markov chain Monte Carlo (MCMC) methods

- Regression models
- Robust estimation
- Spatial statistics
- Stochastic processes
- Survival analysis
- Time series analysis

More information about the Conference, such as the scientific program, abstracts of all given lectures, the list of participants together with their affiliations and contact information, is available in the Book of Abstracts, and at the Conference website https://www.eysm2021.panteion.gr.

These Proceedings contain short papers that went through the peer review process organized by the IOC, in the way that the IOC representatives personally acted as a referee or proposed reviewers for papers of participants they invited.

We would like to thank the European Regional Committee of the Bernoulli Society for giving us the opportunity to organize this lovely event. We are also thankful to the members of the International Organizing Committee for selecting prominent young scientists to attend this conference, as well as to the reviewers of the papers published in the conference proceedings. We also appreciate very much the help of the administrative staff of the Panteion University of Social and Political Sciences. A special thanks goes to our Sponsors for their assistance. Last, but not least, we thank all Keynote Speakers and Young Participants for providing an excellent scientific program, and great vibes that made this event special giving invited young scientists the opportunity to present their recent research results, exchange experience, gain new knowledge and establish contacts, in the hope that this event will be a driving force for their future academic achievements.

Athens, September 2021 Local Organizing Committee

22nd European Young Statisticians Meeting

Co-Organized by

Depts. of Psychology & Sociology, Panteion University of Social & Political Sciences (Host Institution)

Dept. of Statistics & Actuarial-Financial Mathematics, University of the Aegean

Dept. of Statistics, Athens University of Economics & Business

Under the auspices of

Bernoulli Society for Mathematical Statistics and Probability

International Organizing Committee

Conference Structure: keynote lectures, invited lectures.

Conference Language: English

Contents

Papers **1**

Different ordering behavior with respect to two closely related stochastic orders
I. Arab, M. Hadjikyriakou, P.E. Oliveira and B. Santos 3

About one nonparametric estimate of the Bernoulli regression function
P. Babilua 8

On the stochastic orders of extremes order statistics
L.I. Catana 13

Clustering multivariate functional data using unsupervised binary trees
S. Golovkine, N. Klutchnikoff and V. Patilea 18

Comparison of statistics for testing changed segment in a sample
J. Gudan 24

COVID-19: Study of the spread of the pandemic in Bulgaria
S.M. Gurova 30

Reversible genetically modified mode jumping MCMC
A. Hubin, F. Frommlet and G. Storvik 35

A lasso-type estimation for the Lorenz regression
A. Jacquemain, C. Heuchenne and E. Pircalabelu 41

On some aspects of discrete-time semi-Markov switching models
E.N. Kalligeris, A. Karagrigoriou, A. Makrides, C. Parpoula and V.S. Barbu 46

A note on parameter estimation of thinned random intersection graphs
J. Karjalainen 51

Techniques from functional data analysis adaptable for spatial point patterns
K. Koňasová and J. Dvořák 56

Look-ahead screening rules for the Lasso
J. Larsson 61

Adversarial attacks against Bayesian forecasting dynamic models
R. Naveiro 66

Yield curve modelling in insurance
M. Padyšák 71

An application of a geometric process model for debugging and testing costs
M.H. Pekalp and H. Aydoğdu 76
A regime switching on Covid-19 analysis and prediction in Romania
M. Petrica, R.D. Stochitoiu, M. Leordeanu and I. Popescu 81
An analogue of the Feynman-Kac formula for higher order evolution equations
M. Platonova 86
The limit order book model
D. Radojičić 91
Linear regression for uplift modeling
K. Rudaś and S. Jaroszewicz 96
Generalized Gaussian model for EEG data
Z. Salinger, N.N. Leonenko, A. Sikorskii, N. Šuvak and M.J. Boivin 101
Study of neural networks to predict survival in oncology
M. Sautreuil, S. Lemler and P.H. Cournède 107
Testing for the Markov property in sequential decision making
C. Shi 112
Mixed moment estimator for space-time heteroscedastic extremes: semi-parametric inference on extreme rainfall
J. Silva Lomba, M.I. Fraga Alves and C. Neves 117
Combinatorial regression in abstract simplicial complexes
A. Srakar and M. Verbic 123
Application of survival techniques to establish environmental and operational controls on road bridge deterioration
N.A. Stevens, M. Lydon, A.H. Marshall and S.E. Taylor 128
Estimation of in vitro bactericidal potency based on colony counting method
M. Szalai and P. Kevei 133
Modelling block maxima with the blended generalised extreme value distribution
S. M. Vandeskog, S. Martino and D. Castro-Camilo 138
Modeling the superposition of dependent binary signals with hidden Markov models
L.J. Vanegas 143
Clustering based on multivariate mixed type longitudinal data with an application to the EU-SILC database
J. Vávra and A. Komárek 148

Author index **153**

Sponsors **154**

Papers

Different ordering behavior with respect to two closely related stochastic orders

Idir Arab,[1] Milto Hadjikyriakou[2], Paulo Eduardo Oliveira[1] and Beatriz Santos[1*]

[1] *CMUC, Department of Mathematics, University of Coimbra, Portugal*
[2] *University of Central Lancashire, Cyprus*

Abstract: The lifetime of complex systems with heterogeneous components is modelled by distributions depending on several parameters. We exhibit a few applications regarding the ordering of these types of systems with respect to two closely related stochastic orders, namely the star and convex transform orders. Despite the closeness of these orders, different ordering behaviors are obtained.

Keywords: Star-shaped order, convex transform order, parallel systems, series systems.

AMS subject classification: 60E15, 60E05, 62N05.

1 Introduction

The problem of comparing the ageing rates of lifetime of complex systems is addressed by studying two popular notions of stochastic ordering, the convex and star-shaped transform orders. These are closely related, however there are several significant models for which the ageing behavior does not coincide. In this paper, based on previous characterizations for these orderings, we exhibit a few applications with contrasting ageing patterns, depending on the reference ordering criterion.

Let $\mathcal{F}$ denote the family of distributions vanishing at 0, and X be a nonnegative random variable with distribution function $F_X \in \mathcal{F}$, density function f_X, and survival function $\overline{F}_X$. Recall the relevant definitions:

Definition 1. Let X and Y be two nonnegative random variables with distribution functions $F_X, F_Y \in \mathcal{F}$, respectively.

1. The random variable X is said to be smaller than Y in the star-shaped order, denoted by $X \leq_* Y$, if $F_Y^{-1}(F_X(x))$ is star-shaped, i.e., $\frac{1}{x}F_Y^{-1}(F_X(x))$ is increasing with $x > 0$.

2. The random variable X is said to be smaller than Y in the convex transform order, denoted by $X \leq_c Y$, if $F_Y^{-1}(F_X(x))$ is convex, $x > 0$.

*Corresponding author: b14796@gmail.com

The verification of the star-shapedness and convexity of the desired functions can be technically difficult. In Marshall and Olkin [4] or Shaked and Shantikumar [6], a general characterization of the above transform order relations, based on a sign variation technique, may be found.

Theorem 1. *Let X and Y be nonnegative random variables with distribution functions $F_X, F_Y \in \mathcal{F}$.*

1. *$X \leq_* Y$ if and only if, for every real number a, $F_X(x) - F_Y(ax)$ changes sign at most once, and if the sign change occurs it is in the order "$-, +$" as x traverses from 0 to $+\infty$.*
2. *$X \leq_c Y$ if and only if, for every real numbers a and b, $F_X(x) - F_Y(ax + b)$ changes sign at most twice, and if the sign change occurs twice it is in the order "$+, -, +$" as x traverses from 0 to $+\infty$.*

It is obvious from Theorem 1, that the convex transform and star-shaped order are related. In fact, the convex transform order implies the star-shaped order. The main difference between these two orders, since they both capture the notion of a system ageing faster than another, relies on the requirement that the systems under comparison start operating at the same time or not.

An alternative characterization to solve the technical difficulties for the star-shaped order, well adapted for distributions for which we do not have an explicit description of the distribution functions, was proved by Saunders and Moran [5]. The following extension for families depending on multidimensional parameters was proved by Arab et al. [2].

Theorem 2. *Let $\{F_\lambda : \lambda \in I \subseteq \mathbb{R}^n\}$ be a family of distributions such that $F_\lambda \in \mathcal{F}$ and has a density function f_λ, which does not vanish on any subinterval of its support. Let $\mu \in I$, $v = (v_1, v_2, \ldots, v_n) \in \mathbb{R}^n$ and $J \subseteq I$. Then $F_{\lambda_t} \leq_* F_{\lambda_{t'}}$, for every $\lambda_t, \lambda_{t'} \in L_{(\mu,v)} \cap J$, for $t \leq t'$, if and only if $R(x) = \frac{\langle v, \nabla F_\lambda(x) \rangle}{x f_\lambda(x)}$ is decreasing with $x > 0$, for every $\lambda \in L_{(\mu,v)} \cap J$, where $L_{(\mu,v)} = \{\lambda_t \in I \subseteq \mathbb{R}^n : \lambda_t = \mu + tv, t \in \mathbb{R}\}$, $\nabla F_\lambda(x)$ is the gradient of $F_\lambda(x)$ with respect to the parameter λ and $\langle v, \nabla F_\lambda(x) \rangle$ denotes the inner product between v and $\nabla F_\lambda(x)$.*

Remark 1. The argument used for the proof of Theorem 2 reduces the variation of the parameters to moving along a line, hence converting the n-dimensional variation into a one dimensional problem, which is handled using the Saunders and Moran's [5] characterization of the star-shaped order. Note that, we may replace the line $L_{(\mu,v)}$ by any other parametric curve, leading to an obvious extension of Theorem 2.

2 Parallel systems with dependent components

In Proposition 1, a parallel system with dependent exponentially distributed components is compared in the sense of the star-shaped order, with a parallel system that consists of its independent duplicates. Arab et al. [1] proved that these systems are non-comparable with regard to the convex transform order. We assume

that the joint distribution of the dependent components $(X_1, \ldots, X_n)$ follows an n-dimensional FGM (Farlie-Gumbel-Morgenstern, cf. [3]) distribution, that is

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i) \left(1 + \sum_{1 \le j < k \le n} a_{jk} \overline{F}(x_j) \overline{F}(x_k) \right). \tag{1}$$

The distribution function of $X_{(n)} = \max(X_1, X_2, \ldots, X_n)$ is given by

$$F_c(x) = F^n(x)(1 + c\overline{F}^2(x)), \tag{2}$$

where $c = \sum_{1 \le j < k \le n} a_{jk} \in [-1, 1]$. Note that the constant c describes the strength of dependence among the random variables, while its sign reveals the direction of the dependence, i.e., if $c > 0$ ($c < 0$), the components are positively (negatively) dependent. In the following result, the star-shaped comparability follows from the characterization given by Saunders and Moran in [5].

Proposition 1. *Let $X_1, \ldots, X_n$ be independent exponentially distributed random variables with hazard rate $\lambda > 0$ and let $Y_1, \ldots, Y_n$ be exponentially distributed random variables with hazard rate $\lambda > 0$, such that their joint distribution is described by the FGM model defined in (1), with $F(x) = 1 - e^{-\lambda x}$. Then $X_{(n)}$ and $Y_{(n)}$ are not comparable with respect to the convex transform order. If $0 < c < \frac{n}{n+2}$, then $X_{(n)} \le_* Y_{(n)}$.*

3 Complex systems with heterogeneous components

In what follows, we present cases of complex systems with heterogeneous components for which, although the convex transform ordering fails, the star-shaped ordering is established as it is derived from Theorem 2.

Arab et al. [1] proved that two parallel systems with independent and Weibull non-identically distributed lifetime components are not comparable w.r.t. the convex transform order. The following proposition shows a different ordering behavior when considering ordering w.r.t. the star-shaped order.

Proposition 2. *Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be vectors of independent random variables with Weibull distributions with common shape parameter $\alpha > 0$ and scale parameters $0 < \lambda_1 \le \cdots \le \lambda_n$ and $0 < \theta_1 \le \cdots \le \theta_n$, respectively.*

1. *If $\frac{\theta_i}{\theta_j} \le \frac{\lambda_i}{\lambda_j}$, for $i < j$, $i, j = 1, \ldots, n$, then $X_{(n)} \le_* Y_{(n)}$.*
2. *If $\alpha \ge 1$, $\lambda_1 < \lambda_2 \le \cdots \le \lambda_n$, $\theta_1 < \theta_2 \le \cdots \le \theta_n$, then $X_{(n)}$ and $Y_{(n)}$ are not comparable with respect to the convex transform order.*

A similar result to Proposition 2 may be established for parallel systems with Gamma distributed lifetime components.

Proposition 3. *Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be vectors of independent random variables with Gamma distributions with common shape parameter $\alpha > 0$ and scale parameters $0 < \lambda_1 \leq \cdots \leq \lambda_n$ and $0 < \theta_1 \leq \cdots \leq \theta_n$, respectively.*

1. *If $\alpha > 1$ and $\frac{\theta_i}{\theta_j} \leq \frac{\lambda_i}{\lambda_j}$, for $i < j$, $i, j = 1, \ldots, n$, then $X_{(n)} \leq_* Y_{(n)}$.*

2. *If $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$, $\theta_1 < \theta_2 \leq \cdots \leq \theta_n$ and $\lambda_n = \theta_n$, then $X_{(n)}$ and $Y_{(n)}$ are not comparable with respect to the convex transform order.*

Example 1. In Proposition 2 (resp., Proposition 3) assume that $n = 2$, and λ_1 and λ_2 fixed. Choosing θ_1 and θ_2 above the line that goes through $(0, 0)$ and has slope $\frac{\lambda_2}{\lambda_1}$, we may define a system that ages slower w.r.t. the star-shaped order than the system whose distribution depends on the parameters initially fixed. According to Propostion 2 these systems are not comparable w.r.t. the convex transform order. Taking into account Proposition 3, the same conclusion regarding the convex transform order yields, if θ_2 is also chosen so that $\theta_2 = \lambda_2$.

Again, the ordering behavior found for the star-shaped order contrasts with the one obtained by Arab et al. [1] in Proposition 3, where non-comparability, w.r.t. the convex transform order, was proved for parallel systems with Gamma distributed components. These authors also established non-comparability w.r.t. the same ordering notion between series systems when components have exponentiated exponential distributions (cf. Proposition 4 in Arab et al. [1]) as described below. The following result states a different ordering behavior when considering the star-shaped order. Given $X_1, \ldots, X_n$ random variables, we denote by $X_{(1)} = \min(X_1, \ldots, X_n)$.

Proposition 4. *Let $X_1, \ldots, X_n$ be independent random variables where X_i has distribution function $F_i(x) = (1 - e^{-\lambda x})^{\alpha_i}$, for $\lambda, \alpha_i > 0$, $i = 1, \ldots, n$ and $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$. Let $Y_1, \ldots, Y_n$ be independent random variables such that Y_i has distribution function $G_i(x) = (1 - e^{-\lambda x})^{\beta_i}$, for $\beta_i > 0$, $i = 1, \ldots, n$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$.*

1. *If $\frac{\beta_i}{\beta_j} \geq \frac{\alpha_i}{\alpha_j}$, for $i < j$, $i, j = 1, \ldots, n$, $X_{(1)} \geq_* Y_{(1)}$.*

2. *If $\prod_{i=1}^n \alpha_i = \prod_{i=1}^n \beta_i$, then $X_{(1)}$ and $Y_{(1)}$ are not comparable with respect to the convex transform order.*

Example 2. In the assumptions of Proposition 4, for $n = 2$, given α_1 and α_2 fixed, we may choose β_1 and β_2 between the lines that goes through $(0, 0)$ and have slopes 1 and $\frac{\alpha_2}{\alpha_1}$, to define a system that ages slower w.r.t. the star-shaped order than the system whose distribution depends on the parameters initially fixed. However, the systems considered are not comparable w.r.t. the convex order, if β_2 and β_1 are also chosen to satisfy $\beta_2 = \frac{\alpha_1 \alpha_2}{\beta_1}$.

Remark 2. The direction of the star-shaped ordering in Propositions 2, 3 and 4 may be reversed, if the inequalities assumed for the parameters are also reversed.

4 Conclusions

Arab et al. [1] established non-comparison results regarding the convex transform order relation, within parallel and series systems depending on a large number of parameters. However, the characterizations provided were not sufficient to draw any conclusions with regard to star-shaped order for the same type of systems. The result proved by Arab et al. in [2], gives a new criterion for the star-shaped order to hold between families of distributions indexed by multidimensional parameters. Based on the new criterion, we obtain a different ordering behavior from the one found when considering ordering w.r.t. the convex transform order, for the referred systems.

Since the star-order implies other stochastic orders, such as the second stochastic dominance and Lorenz orders (cf. Shaked and Shantikumar [6]), which have several applications in economics, we may conclude that the results remain true for the aforementioned orders.

Acknowledgements: The authors IA, PEO, and BS are partially supported by the Centre for Mathematics of the University of Coimbra - UIDB/00324/2020, funded by the Portuguese Government through FCT/MCTES. BS was also supported by FCT, through the grant PD/BD/150459/2019, co-financed by the European Social Fund.

Bibliography

[1] I. Arab, M. Hadjikyriakou and P.E. Oliveira. Non comparability with respect to the convex transform order with applications. *J. Appl. Prob.* 57(4), 1339–1348, 2020.

[2] I. Arab, M. Hadjikyriakou, P.E. Oliveira and B. Santos. Star-shaped order for distributions with multidimensional parameters and some applications. *Pré-publicações do Departamento de Matemática da Universidade de Coimbra, 21-05,* 2021.

[3] S. Kotz, N. Balakrishan and N. L. Johnson. *Continuous Multivariate Distributions: Models and Applications.* John Wiley, New York, 2000.

[4] A. W. Marshall I. Olkin. *Life Distributions.* Springer, New York, 2007.

[5] I. W. Saunders and P. A. P. Moran. On the quantiles of the gamma and F distributions. *J. Appl. Prob.*, 15(2), 426–432, 1978.

[6] M. Shaked and J. G. Shantikumar. *Stochastic orders.* Springer, New York, 2007.

About one nonparametric estimate of the Bernoulli regression function

Petre Babilua[1]*

[1]*Faculty of Exact and Natural Sciences, Department of Mathematics, Ivane Javakhishvili Tbilisi State University, University Str., 3 Tbilisi 0143, Georgia*

Abstract: The estimate for the Bernoulli regression function is constructed using the Bernstein polynomial. The question of its consistency and asymptotic normality is studied. Testing hypothesis is constructed on the form of the Bernoulli regression function.

Keywords: Bernstein polynomial, Bernoulli regression function, consistency, testing hypothesis

AMS subject classification: 62G10, 62G20.

1 Introduction

Our results obtained for an estimate of the Bernoulli regression function are new. For us the existence of similar results for the Bernoulli regression function is unknown. We were the first to introduce Bernstein polynomials as an estimate of the Bernoulli regression function and moreover, the estimate created by us is free from the boundary effect, which motivated us to study such kind of estimates. Note that the kernel-type estimates for the Bernoulli regression function, which we have considered in [7], do not have such a good property.

Let a random variable Y take two values 1 and 0 with probabilities p ("success") and $1-p$ ("failure"). Assume that the probability of "success" p is the function of an independent variable $x \in [0,1]$, i.e. $p = p(x) = \mathbb{P}(Y = 1 \mid x)$ [1, 2, 3, 4]. Assume that $x_k = \frac{k}{n}$, $k = 0, \ldots, n$, are the points of division of the interval $[0,1]$ and we have Y_i, $i = 0, \ldots, n$, which are independent Bernoulli random variables with $\mathbb{P}(Y_i = 1 \mid x_i) = p(x_i)$, $\mathbb{P}(Y_i = 0 \mid x_i) = 1 - p(x_i)$. The problem consists in estimating a function $p(x)$, $x \in [0,1]$, by means of the sampling $Y_0, Y_1, \ldots, Y_n$. A problem like this one arises, for instance, in biology [1, 3, 4], also when studying corrosion processes [5], and so on.

As an estimate for $p(x)$ let us consider the following statistic

$$\widehat{p}_n(x) = \sum_{k=0}^{n} Y_k b_k(n,x), \tag{1}$$

*Corresponding author: petre.babilua@tsu.ge

where $b_k(n,x) = \binom{n}{k}x^k(1-x)^{n-k}$, $k = 0,\dots,n$ is a binomial distribution with probability of "success" p, $x \in (0,1)$.

Note that

$$E\widehat{p}_n(x) = B_n(x) = \sum_{k=0}^{n} p\Big(\frac{k}{n}\Big)b_k(n,x),$$

where $B_n(x)$ is a Bernstein polynomial of order n of the function $p(x)$. It is well known that if $p(x)$ is continuous on $[0,1]$, then $\lim\limits_{n\to\infty} B_n(x) = p(x)$ uniformly with respect to $x \in [0,1]$. Moreover, the order of bias $E\widehat{p}_n(x) - p(x)$ is established from the result given in Lorentz' monograph [6, Section 1.6]. According to Lorentz monograph, following assertion is true.

Lemma 1. *Let $p(x)$, $x \in [0,1]$ have a bounded derivative of second order (we denote the class of such functions by $W[0,1]$). Then*

(a)

$$E\widehat{p}_n(x) - p(x) = O\Big(\frac{1}{n}\Big) \tag{2}$$

uniformly with respect to $x \in [0,1]$.

(b) *Let $p''(x)$ satisfies the Lipshitz condition, i.e. there exists $c > 0$ such that $|p''(x) - p''(y)| \le c|x-y|$ for all $x, y \in [0,1]$, then*

$$E\widehat{p}_n(x) = p(x) + n^{-1}x(1-x)p''(x)2^{-1} + O(n^{-3/2})$$

uniformly with respect to $x \in [0,1]$.

Moreover, the estimate $\widehat{p}_n(x)$ is free from the boundary effect, which motivates us to study estimates like (1) for the Bernoulli regression function $p(x)$. Note that the kernel-type estimates of the function $p(x)$, which we have considered in [7], do not have such a good property.

Theorem 1. *Let $p(x) \in W[0,1]$. Then, for $x \in (0,1)$*

1^0. *$\widehat{p}_n(x)$ is a consistent estimate of $p(x)$.*

2^0. $\sqrt{n}\,(\widehat{p}_n(x) - p(x))\sigma^{-1}(x) \xrightarrow{d} N(0,1)$,

$$\sigma^2(x) = p(x)(1-p(x))\big[4\pi x(1-x)\big]^{-\frac{1}{2}},$$

where $\xrightarrow{d}$ denotes converges in distribution and $N(0,1)$ is a random variable that has a standard normal distribution $\Phi(x)$.

$$\sqrt{n}\,\big(\widehat{p}_n(x) - p(x)\big)\sigma_n^{-1}(x) \xrightarrow{d} N(0,1), \quad x \in (0,1),$$

where

$$\sigma_n^2(x) = \widehat{p}_n(x)(1-\widehat{p}_n(x))\big[4\pi x(1-x)\big]^{-\frac{1}{2}}.$$

This makes it possible to construct the confidence interval for $p(x)$:

$$p_n^{\pm}(x) = \widehat{p}_n(x) \pm \frac{\sigma_n(x)}{\sqrt{n}}\,\lambda_\alpha, \;\; \lambda_\alpha = \Phi^{-1}\Big(\frac{1+\alpha}{2}\Big), \;\; 0<\alpha<1.$$

Theorem 2.

(a) *Let $p(x)$ have the bounded first derivative. Then*

$$\sqrt{n}\,\overline{T}_n \xrightarrow{d} N(0,\sigma^2(p)),$$

where $\overline{T}_n = \int\limits_0^1 [\widehat{p}_n(x) - E\widehat{p}_n(x)]\,dx$, $\sigma^2(p) = \int\limits_0^1 p(x)(1-p(x))\,dx$.

(b) *Let $p(x) \in W[0,1]$. Then*

$$\sqrt{n}\,T_n \xrightarrow{d} N(0,\sigma^2(p)), \tag{3}$$

where $T_n = \int\limits_0^1 [\widehat{p}_n(x) - p(x)]\,dx$, $\sigma^2(p) = \int\limits_0^1 p(x)(1-p(x))\,dx$.

Testing the specified of the Bernoulli regression function

We give several comments on the application of T_n as a test statistic for the testing hypothesis H_0: $p(x) = p_0(x)$ ($p_0(x)$, $x \in [0,1]$, is the well-known dose-response curve). This statistic is informative because the sign of T_n may carry information on the character of an alternative when the hypothesis H_0 is not true, i.e. the sign of the test statistic indicates the direction of deviation of the alternative from H_0. It can be shown that

$$E\int\limits_0^1 \big(\widehat{p}_n(x) - p_0(x)\big)\,dx \sim \int\limits_0^1 (p(x) - p_0(x))\,dx.$$

Thus, for an alternative of the form $H_1^+ : \; p(x) > p_0(x)$ the statistic T_n will have the tendency to deviate to the right from zero, while for the alternative $H_1^- : \; p(x) < p_0(x)$ it deviates to the left. Hence it is natural to use the statistic T_n in problems of testing the hypothesis H_0 against the one-sided alternatives H_1^+ and H_1^-.

The assertion (b) of Theorem 2 enables us to construct the test of the asymptotic level α, $0<\alpha<1$, for testing the hypothesis H_0, according to which $p(x) = p_0(x)$, $x \in [0,1]$:

Test I. Reject the hypothesis H_0 against the right-side alternative H_1^+: $p(x) > p_0(x)$, $x \in [0,1]$ when $T_n \geq \frac{\lambda_\alpha \sigma(p_0)}{\sqrt{n}}$ for $\lambda_\alpha = \Phi^{-1}(1-\alpha)$.

Test II. Reject the hypothesis H_0 against the left-side alternative H_1^-: $p(x) < p_0(x)$ when $T_n \leq \frac{\lambda_\alpha}{\sqrt{n}}\,\sigma(p_0)$ for $\lambda_\alpha = \Phi^{-1}(\alpha)$.

Tests I and II are consistent against the one-sided alternatives H_1^+ and H_1^-, respectively. As an example we show that this is so for Test I. Let $p(x)$ and $p_0(x) \in W[0,1]$.

It is obvious that

$$\begin{aligned}\gamma_n(p) &= \mathbb{P}_{H_1^+}\Big(T_n \geq \lambda_\alpha \frac{\sigma(p_0)}{\sqrt{n}}\Big) \\ &= \mathbb{P}_{H_1^+}\Bigg(\sqrt{n}\int_0^1 \big(\widehat{p}_n(x) - p(x)\big)\,dx\,\sigma^{-1}(p) \\ &\qquad \geq -\sqrt{n}\int_0^1 (p(x) - p_0(x))\,dx \cdot \sigma^{-1}(p) + \lambda_\alpha \sigma(p_0)\sigma^{-1}(p)\Bigg).\end{aligned}$$

Since

$$\sqrt{n}\int_0^1 \big(\widehat{p}_n(x) - p(x)\big)\,dx\,\sigma^{-1}(p) \xrightarrow{d} N(0,1)$$

for the hypothesis H_1^+, we have $\gamma_n(p) \longrightarrow 1$ as $n \to \infty$.

However if n changes, the alternative changes too, getting closer to the basic hypothesis H_0, which means that the power of the test does not necessarily converge to 1. As an example, let us consider a sequence of Pitmen-type alternatives that are close to the hypothesis H_0:

$$H_n^+ : \;\; p_1^{(n)}(x) = p_0(x) + n^{-\frac{1}{2}} u(x), \tag{4}$$

where $u(x) > 0$ and $u(x) \in W[0,1]$. Then

$$\mathbb{P}_{H_n^+}\Big(T_n \geq \frac{\lambda_\alpha}{\sqrt{n}}\,\sigma(p_0)\Big) \longrightarrow 1 - \Phi\Big(\lambda_\alpha - \frac{c}{\sigma(p_0)}\Big), \quad c = \int_0^1 u(x)\,dx > 0.$$

Indeed,

$$\begin{aligned}&\mathbb{P}_{H_n^+}\Big(T_n \geq \frac{\lambda_\alpha}{\sqrt{n}}\,\sigma(p_0)\Big) \\ &\quad = \mathbb{P}_{H_n^+}\Bigg(\sqrt{n}\int_0^1 \big(\widehat{p}_n(x) - p_1^{(n)}(x)\big)\,dx\,\sigma^{-1}(p_1^{(n)}) \geq \frac{\sigma(p_0)}{\sigma(p_1^{(n)})}\,\lambda_\alpha - \frac{c}{\sigma(p_1^{(n)})}\Bigg) \\ &\qquad\qquad \longrightarrow 1 - \Phi\Big(\lambda_\alpha - \frac{c}{\sigma(p_0)}\Big).\end{aligned}$$

Hence it follows that for alternative (4) Test I for testing the hypothesis H_0 is asymptotically strictly unbiased since $c > 0$, and is equal to 0 if and only if $u(x) = 0$ (For Test II, the argumentation is analogous).

Bibliography

[1] S. Efromovich. *Nonparametric Curve Estimation. Methods, Theory, and Applications*. Springer Series in Statistics. Springer-Verlag, New York, 1999.
[2] J. B. Copas. Plotting p against x. *Appl. Statist.*, 32(1):25–31, 1983.
[3] H. Okumura and K. Naito. Weighted kernel estimators in nonparametric binomial regression. *J. Nonparametr. Stat.*, 16(1-2):39–62, 2004.
[4] M. Aerts and N. Veraverbeke. Bootstrapping a nonparametric polytomous regression model. *Math. Methods Statist.*, 4(2):189–200, 1995.
[5] K. V. Mandzhgaladze. On some estimate of the distribution function and its moments. *Soobshch. Akad. Nauk Gruz. SSR*, 124:261–263, 1986 (in Russian).
[6] G. G. Lorentz. *Bernstein Polynomials*. Second edition. Chelsea Publishing, New York, 1986.
[7] E. Nadaraya, P. Babilua and G. Sokhadze. About the nonparametric estimation of the Bernoulli regression. *Comm. Statist. Theory Methods*, 42(22):3989–4002, 2013.
[8] A. Leblanc. On estimating distribution functions using Bernstein polynomials. *Ann. Inst. Statist. Math.*, 64(5):919–943, 2012.
[9] H. Cramér. *Mathematical Methods of Statistics*. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946.
[10] A. N. Shiryaev. *Probability*. Nauka, Moscow, 1989 (in Russian).
[11] W. Feller. *An Introduction to Probability Theory and its Applications*. Vol. II. John Wiley & Sons, Inc., New York–London–Sydney, 1966.

On the stochastic orders of extremes order statistics

Luigi-Ionut Catana[1*]

[1]*Faculty of Mathematics and Computer Science, Mathematical Doctoral School, University of Bucharest, Str. Academiei nr. 14, sector 1, C.P. 010014, Bucharest, Romania*

Abstract: In this paper we prove that we can have hazard rate and reversed hazard rate orders of extremes order statistics in the case of quadratic transmuted distributions for a family parameters, although the parameters are not comparable in some sense.

Keywords: Stocahastic order, Transmuted distribution, Order statistics
AMS subject classification: 60E15

1 Introduction

An important problem in risk analysis is the ordering of two risks. Stochastic dominance has become a topic of great interest which is widely studied due to its various applications in different domains: economics, finances, banking, statistics, risk theory, medicine and others. For introduction in the field, we refer to the following books that address different types of stochastic dominance and the links between them: Levy (2015) and Shaked and Shanthikumar (2007).

Order statistics are of great interest in operations research, reliability theory, data analysis, statistical inference and other areas of applied probability. They have received a lot of attention from many researchers. Let us consider $X_1, X_2, ..., X_n$ with $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ where $X_{k:n}$ represents the $k-$th order statistic which is related to the lifetimes of $(n-k+1)$-out-of-n system. In particular, $X_{n:n}$ (when $k=n$) and $X_{1:n}$ (when $k=1$) denote the lifetimes of parallel and series systems, respectively. Balakrishnan and Rao (1998) offered an important course in this field. Some recent applications have included: empirical studies of price dispersion on the Internet (Warin and Leiter (2012)); utility maximization frameworks for fair and efficient multicasting in multicarrier wireless cellular networks (Liu et al. (2013)); degradation pattern prediction of a polymer electrolyte membrane fuel cell stack (Bae et al. (2014)). Results about stochastic order of smallest or highest order statistics were given by Balakrishnan and Torrado (2016), Balakrishnan et al. (2020), Chen et al. (2019), Khaledi and Kochar (2006).

*Corresponding author: luigi_catana@yahoo.com

The work in this case uses transmuted distribution, a recent class of distribution. A transmuted distribution obtained from another distribution. A simple extension of the probability distributions is proposed by Shaw and Buckley (2007).

We present the structure of this paper. In the section 2 there are presented the preliminaries. In the section 3 we present sufficient conditions of transmutation parameters for some stochastic orders from Catana and Preda (2021). In the section 4 we prove that we can have hazard rate and reversed hazard rate orders of extremes order statistics in the case of quadratic transmuted distributions for a family parameters, although the parameters are not comparable in some sense. In the last section we discuss the conclusions.

2 Preliminaries

Let $(\Omega, \mathcal{F}, P)$ be a probability space and λ the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let $X : \Omega \to \mathbb{R}$ be an absolutely continuous nonnegative random variable. We consider $F_X(x) = P(X \leq x)$ its distribution function, survial function $\overline{F}_X(x) = P(X > x) = 1 - F_X(x)$, f_X density function, $r_X(x) = \frac{f_X(x)}{\overline{F}_X(x)}$, $x \in Supp(\overline{F}_X)$ hazard rate function and $\widetilde{r}_X(x) = \frac{f_X(x)}{F_X(x)}$, $x \in Supp(F_X)$ reversed hazard rate function, where, for a function $g : \mathbb{R} \to \mathbb{R}$ we denote $Supp(g) = \{x \in \mathbb{R} : g(x) \neq 0\}$.

In this paper all the random variables are absolutely continuous with respect to the Lebesgue measure.

For $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ real numbers, we denote by $x_{(i)}$ the i-th ordered number.

Let $X_1, X_2, ..., X_n$ be independent random variables.

We denote $X_{1:n} = \min(X_1, X_2, ..., X_n)$ and $X_{n:n} = \max(X_1, X_2, ..., X_n)$.

Then it is known that

$$\overline{F}_{X_{1:n}}(x) = \prod_{i=1}^{n} \overline{F}_{X_i}(x) \text{ and } F_{X_{n:n}}(x) = \prod_{i=1}^{n} F_{X_i}(x)$$

$$f_{X_{1:n}}(x) = \prod_{i=1}^{n} \overline{F}_{X_i}(x) \cdot \left(\sum_{i=1}^{n} r_{X_i}(x) \right), \; f_{X_{n:n}}(x) = \prod_{i=1}^{n} F_{X_i}(x) \cdot \left(\sum_{i=1}^{n} \widetilde{r}_{X_i}(x) \right),$$

$$r_{X_{1:n}}(x) = \sum_{i=1}^{n} r_{X_i}(x) \text{ and } \widetilde{r}_{X_{n:n}}(x) = \sum_{i=1}^{n} \widetilde{r}_{X_i}(x).$$

Let H be a real distribution function, $\overline{H} = 1 - H$ and $h = H'$. For $\lambda \in [-1, 1]$ we denote the quadratic transmuted distribution $T - H(\lambda)$ with distribution function $F = (1 + \lambda)H - \lambda H^2$.

Definition 1. *(Shaked and Shanthikumar, 2007) Let $x, y \in \mathbb{R}^d$.*

(i) We say that x is smaller than y in the sense 1 (and denote $x \leq_{(1)} y$) if $x_i \leq y_i \; \forall \, i \in \{1, 2, ..., d\}$.

Definition 2. *(Shaked and Shanthikumar, 2007) Let X, $Y : \Omega \to \mathbb{R}$ random variables. We say that X is smaller than Y in the*

(i) hazard rate order (written as $X \prec_{hr} Y$) if $r_X(x) \geq r_Y(x) \; \forall \, x \in Supp(\overline{F}_X) \cap Supp(\overline{F}_Y)$;

(ii) reversed hazard rate order (written as $X \prec_{rh} Y$*) if* $\widetilde{r}_X(x) \leq \widetilde{r}_Y(x)$ $\forall$ $x \in Supp(F_X) \cap Supp(F_Y)$;

3 Stochastic order when the transmutation parameters are comparable in sense 1

The following four theorems give sufficient conditions for hazard rate and reverse hazard rate orders between extremes order statistics using transmuted distributions:

Theorem 1. *(Catana and Preda, 2021, Theorem 3.2) Let* $X_1, X_2, ..., X_n$ *independent random variables,* $X_i \sim T - H(\alpha_i)$ *and* $Y_1, Y_2, ..., Y_n$*, independent random variables,* $Y_i \sim T - H(\beta_i)$*. Then* $(\alpha_1, ..., \alpha_n) \geq_{(1)} (\beta_1, ..., \beta_n) \Rightarrow X_{1:n} \prec_{hr} Y_{1:n}$.

Theorem 2. *(Catana and Preda, 2021, Theorem 3.4) Let* $X_1, X_2, ..., X_n$ *be independent random variables,* $X_i \sim T - H(\alpha_i)$ *and* $Y_1, Y_2, ..., Y_n$*, independent random variables,* $Y_i \sim T - H(\beta_i)$*. Then* $(\alpha_1, ..., \alpha_n) \geq_{(1)} (\beta_1, ..., \beta_n) \Rightarrow X_{n:n} \prec_{rh} Y_{n:n}$.

More results with more general transmuted distributions classes are given in Catana and Preda (2021).

4 Main results

$r_{X_{1:2}}(x) = r_{X_1}(x) + r_{X_2}(x) = \frac{(1+\alpha_1)h(x) - 2\alpha_1 h(x)\overline{H}(x)}{(1+\alpha_1)\overline{H}(x) - \alpha_1\left(\overline{H}(x)\right)^2} + \frac{(1+\alpha_2)h(x) - 2\alpha_2 h(x)\overline{H}(x)}{(1+\alpha_2)\overline{H}(x) - \alpha_2\left(\overline{H}(x)\right)^2} = \frac{h(x)}{\overline{H}(x)}\left(2 - \frac{\alpha_1\overline{H}(x)}{1+\alpha_1-\alpha_1\overline{H}(x)} - \frac{\alpha_2\overline{H}(x)}{1+\alpha_2-\alpha_2\overline{H}(x)}\right)$

Thus $X_{1:2} \prec_{hr} Y_{1:2} \Leftrightarrow \frac{\alpha_1}{1+\alpha_1-\alpha_1\overline{H}(x)} + \frac{\alpha_2}{1+\alpha_2-\alpha_2\overline{H}(x)} \leq \frac{\beta_1}{1+\beta_1-\beta_1\overline{H}(x)} + \frac{\beta_2}{1+\beta_2-\beta_2\overline{H}(x)}$ $\forall x \in Supp(\overline{F}_{X_{1:2}}) \cap Supp(\overline{F}_{Y_{1:2}})$

$\widetilde{r}_{X_{2:2}}(x) = \widetilde{r}_{X_1}(x) + \widetilde{r}_{X_2}(x) = \frac{(1-\alpha_1)h(x) + 2\alpha_1 h(x)H(x)}{(1-\alpha_1)H(x) + \alpha_1(H(x))^2} + \frac{(1-\alpha_2)h(x) + 2\alpha_2 h(x)H(x)}{(1-\alpha_2)H(x) + \alpha_2(H(x))^2} = \frac{h(x)}{H(x)}\left(2 + \frac{\alpha_1 H(x)}{1-\alpha_1+\alpha_1 H(x)} + \frac{\alpha_2 H(x)}{1-\alpha_2+\alpha_2 H(x)}\right)$

$X_{2:2} \prec_{rh} Y_{2:2} \Leftrightarrow \frac{\alpha_1}{1-\alpha_1+\alpha_1 H(x)} + \frac{\alpha_2}{1-\alpha_2+\alpha_2 H(x)} \leq \frac{\beta_1}{1-\beta_1+\beta_1 H(x)} + \frac{\beta_2}{1-\beta_2+\beta_2 H(x)}$ $\forall x \in Supp(F_{X_{2:2}}) \cap Supp(F_{Y_{2:2}})$.

Proposition 1 and 2 below show that there exists a family transmutation parameters that are not comparable in sense 1 we can have stochastic order in the sense of hazard rate and reversed hazard rate.

Proposition 1. *Let* $n = 2$, $r \in [0, \infty)$, $\alpha_1 = -r$, $\alpha_2 = r$, $\beta_1 \in [r, \infty)$, $\beta_2 \in [-r, r]$. *Then there is no order in sense 1 between* (α_1, α_2) *and* (β_1, β_2) *but* $Y_{1:2} \prec_{hr} X_{1:2}$.

Proof. According to the definition of order in sense 1 between two points we have that there is no order in sense 1 between (α_1, α_2) and (β_1, β_2). Let $X_1 \sim T - H(-r), X_2 \sim T - H(r)$ independent random variables, $Y_1 \sim T - H(r), Y_2 \sim T - H(-r)$ independent random variables. Then $r_{X_{1:2}}(x) = r_{Y_{1:2}}(x)$, $\forall$ $x \in Supp(\overline{F}_{X_{1:2}}) \cap Supp(\overline{F}_{Y_{1:2}})$, thus $X_{1:n} =_{hr} Y_{1:n}$. Let $X_1' \sim T - H(r), X_2' \sim T - H(-r)$

independent random variables, $Y_1' \sim T - H(\beta_1), Y_2' \sim T - H(\beta_2)$ independent random variables. From theorem 1 it results $Y_{1:n}' \prec_{hr} X_{1:n}'$. Thus $X_{1:n} =_{hr} Y_{1:n} =_{hr} Y_{1:n}' \prec_{hr} X_{1:n}'$. □

Proposition 2. *Let $n = 2$, $r \in [0, \infty)$, $\alpha_1 = -r$, $\alpha_2 = r$, $\beta_1 \in [r, \infty)$, $\beta_2 \in [-r, r]$. Then there is no order in sense 1 between (α_1, α_2) and (β_1, β_2) but $Y_{2:2} \prec_{rh} X_{2:2}$.*

Proof. According to the definition of order in sense 1 between two points we have that there is no order in sense 1 between (α_1, α_2) and (β_1, β_2). Let $X_1 \sim T - H(-r), X_2 \sim T - H(r)$ independent random variables, $Y_1 \sim T - H(r), Y_2 \sim T - H(-r)$ independent random variables. Then $\widetilde{r}_{X_{2:2}}(x) = \widetilde{r}_{Y_{2:2}}(x)$, $\forall\ x \in Supp(F_{X_{2:2}}) \cap Supp(F_{Y_{2:2}})$, thus $X_{2:2} =_{rh} Y_{2:2}$. Let $X_1' \sim T - H(r), X_2' \sim T - H(-r)$ independent random variables, $Y_1' \sim T - H(\beta_1), Y_2' \sim T - H(\beta_2)$ independent random variables. From theorem 1 it results $Y_{2:2}' \prec_{rh} X_{2:2}'$. Thus $X_{2:2} =_{rh} Y_{2:2} =_{rh} Y_{2:2}' \prec_{rh} X_{2:2}'$. □

5 Conclusion

In this paper we have shown that for Theorems 1 and 2 presented in Section 3, the reciprocal is not generally valid. In a future research we will extend the results in a more general context.

Acknowledgements: Thanks to professors Vasile Preda and Gheorghita Zbaganu for my introduction in this field and advice.

Bibliography

[1] Bae, S. J., Kim, S. J., Lee, J. H., Song, I., Kim, N. I., Seo, Y., et al. (2014). Degradation pattern prediction of apolymer electrolyte membrane fuel cell stack with series reliability structure via durability data of single cells. *Applied Energy,* 131, 48–55.

[2] Balakrishnan, N., Rao C.R. (1998). *Order Statistics: Applications.* Handbook of Statistics 17, North-Holland, Amsterdam.

[3] Balakrishnan, N., Torrado N. (2016). Comparisons between largest order statistics from multiple-outlier models. *Communications in Statistics-Theory and Methods,* 50, 176-189.

[4] Balakrishnan, N., Barmalzan, G., Haidari, A. (2020). Exponentiated models preserve stochastic orderings of parallel and series systems. *Communications in Statistics-Theory and Methods*, 49, 1592-1602.

[5] Catana, L. I., Preda, V. (2021). Comparing the extremes order statistics between two random variables sequences using transmuted distributions. *Communications in Statistics-Theory and Methods*, 1-18, DOI: 10.1080/03610926.2021.1898641.

[6] Chen, J., Zhang, Y., Zhao, P. (2019). Comparisons of order statistics from heterogeneous negative binomial variables with applications. *Communications in Statistics-Theory and Methods,* 53, 990-1011.

[7] Khaledi, B.E., Kochar, S. C. (2006). Weibull distribution: Some stochastic comparisons results. *Journal of Statistical Planning and Inference,* 136, 3121-3129.
[8] Levy, H. (2015). *Stochastic Dominance: Investment Decision Making under Uncertainty*, 3rd ed.; Springer: Berlin/Heidelberg, Germany.
[9] Liu, J., Chen, W., Zhang, Y. J., Cao, Z. (2013). A utility maximization framework for fair and efficient multicasting in multicarrier wireless cellular networks. *IEEE/ACM Transactions on Networking,* 21, 110-120.
[10] Shaked, M., Shanthikumar, J. G. (2007). *Stochastic orders*. New York: Springer.
[11] Shaw, W.T., Buckley, I.R.C. (2009). The alchemy of probability distributions: beyond Gram-Charlier expansions, and a skew-kurtotic-normal distribution from a rank transmutation map.
[12] Warin, T., Leiter, D. (2012). Homogenous goods markets: An empirical study of price dispersion on the internet. *International Journal of Economics and Business Research,* 4, 514-529.

Clustering multivariate functional data using unsupervised binary trees

Steven Golovkine,[1*] **Nicolas Klutchnikoff**[2] **and Valentin Patilea**[3]

[1]*Groupe Renault & CREST - UMR 9194, Rennes, France*
[2]*Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France*
[3]*Ensai, CREST - UMR 9194, Rennes, France*

Abstract: We propose a model-based clustering algorithm for a general class of functional data. The random functional data realizations could be measured with error at discrete, and possibly random, points in the definition domain. The idea is to build a set of binary trees by recursive splitting of the observations. The number of groups are determined in a data-driven way. The algorithm provides easily interpretable results and fast predictions for online data sets. Results on simulated datasets reveal good performance in various complex settings. The open-source implementation of the algorithm can be accessed at https://github.com/StevenGolovkine/FDApy. Complete version of the work is available at arxiv:2012.05973.

Keywords: Gaussian mixtures, Model-based clustering, Multivariate functional principal components analysis

AMS subject classification: 62R10

1 Introduction

Sensors are more and more present in our daily life. These provide a large amount of data that can be modeled as functional data. The amount of such collected data grows rapidly as does the cost of their labeling. Thus, there is an increasing interest in methods that aim to identify homogeneous groups within functional datasets.

Assume a sample of N curves as being values on the realizations of a stochastic process, possibly recorded with some error, at discrete random times. We aim to define a procedure, based on a sample of N noisy curves, to build groups of similar curves.

*Corresponding author: steven_golovkine@icloud.com

2 Model

The structure of our data, referred to as *multivariate functional data*, is very similar to that presented in [3]. The data consist of independent trajectories of a stochastic process $X = (X^{(1)}, \ldots, X^{(P)})^\top$, $P \geq 1$. For each $1 \leq p \leq P$, let $\mathcal{T}_p = [0,1]^{d_p}, d_p \geq 1$. Each coordinate $X^{(p)} :_p \to \mathbb{R}$ is assumed to belong to 2_p endowed with the usual inner product that we denote by $\cdot,\cdot$. Thus X is a stochastic process indexed by $= (t_1, \ldots, t_P)$ belonging to $:=_1 \times \cdots \times_P$ and taking values in $:= 2_1 \times \cdots \times 2_P$. Consider the function $\cdot,\cdot : \times \to \mathbb{R}$,

$$f, g := \sum_{p=1}^{P} f^{(p)}, g^{(p)}, \quad f, g \in .$$

is a Hilbert space with respect to the inner product $\cdot,\cdot$ [3].

Let $K \geq 1$ be an integer, and let Z be a random variable taking values in $\{1, \ldots, K\}$ such that $\mathbb{P}(Z = k) = p_k$ with $p_k > 0$ and $\sum_{k=1}^{K} p_k = 1$. The variable Z represents the cluster membership of the realizations of the process. We consider that the stochastic process X follows a *functional mixture model with K components*:

$$X() = \sum_{k=1}^{K} \mu_k() \mathbb{1}_{\{Z=k\}} + \sum_{j \geq 1} \xi_j \phi_j(), \quad \in, \tag{1}$$

where $\mu_1, \ldots, \mu_K \in$ are the mean curves per cluster, $\{\phi_j\}_{j \geq 1}$ is an orthonormal basis of and ξ_j, $j \geq 1$ are real-valued random variables which are conditionally independent given Z. For each $1 \leq k \leq K$, $\xi_j \mid Z = k \sim \mathcal{N}(0, \sigma_{kj}^2)$ for all $j \geq 1$.

Lemma 1. *Let X be defined as in (1) for some orthonormal basis $\{\phi_j\}_{j \geq 1}$. Let $\{\psi_j\}_{j \geq 1}$ be another orthonormal basis in and consider*

$$c_j = X - \mu, \psi_j, \quad j \geq 1 \quad \textit{where} \quad \mu(\cdot) = \sum_{k=1}^{K} p_k \mu_k(\cdot).$$

Then $c_j \mid Z = k \sim m_{kj} \tau_{kj}^2$, where

$$m_{kj} = \mu_k - \mu, \psi_j \ \textit{ and } \ \tau_{kj}^2 = \sum_{l \geq 1} \phi_l, \psi_j{}^2 \sigma_{kl}^2.$$

Remark 3. Lemma 1 shows that, no matter what the user's choice may be for orthonormal basis $\{\psi_j\}_{j \geq 1}$, the clusters will be preserved after expressing the realizations of the process into this basis. However, some bases might be more suitable than others. In particular, the basis based on a multivariate functional principal component analysis (MFPCA) developed in [3] is very interesting in this context.

In real data applications, the realizations of X are usually measured with error at discrete, and possibly random, points in the definition domain. For each $1 \leq n \leq N$, and given a vector of positive integers $\boldsymbol{M}_n = (M_n^{(1)}, \ldots, M_n^{(P)})$, let $T_{n,\boldsymbol{m}} =$

$(T^{(1)}_{n,m_1}, \ldots, T^{(P)}_{n,m_P}), 1 \leq m_p \leq M^{(p)}_n, 1 \leq p \leq P$, be the random observation times for the curve X_n. These times are obtained as independent copies of a variable $\boldsymbol{T}$ taking values in $\mathcal{T}$. The vectors $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_N$ represent an independent sample of an integer-valued random vector $\boldsymbol{M}$ with expectation $\boldsymbol{\mu_M}$ which increases with N. We assume that the variables X, $\boldsymbol{M}$ and $\boldsymbol{T}$ are mutually independent. The observations associated with a curve, or trajectory, X_n consist of the pairs $(Y_{n,\boldsymbol{m}}, T_{n,\boldsymbol{m}}) \in \mathbb{R}^P \times \mathcal{T}$, where $\boldsymbol{m} = (m_1, \ldots, m_P), 1 \leq m_p \leq M^{(p)}_n$, $1 \leq p \leq P$, and $Y_{n,\boldsymbol{m}}$ is defined as

$$Y_{n,\boldsymbol{m}} = X_n(T_{n,\boldsymbol{m}}) + \varepsilon_{n,\boldsymbol{m}}, \quad 1 \leq n \leq N, \tag{2}$$

with the $\varepsilon_{n,\boldsymbol{m}}$ being independent copies of a centered error random vector $\varepsilon \in \mathbb{R}^P$ with finite variance. The mean and covariance functions of $X^{(p)}, 1 \leq p \leq P$ of X can be estimated using, *e.g.*, [7]. Concerning the estimation of the eigenfunctions and the eigenvalues for the MFPCA, as well as for the projection of the observations on the eigenfunctions basis, we use [3].

3 The fCUBT algorithm

Let $\mathcal{S}$ be a sample of realizations of the process X, defined in (1). We consider the problem of learning a partition $\mathcal{U}$ such that every element U of $\mathcal{U}$ gathers similar elements of $\mathcal{S}$. Our clustering procedure follows the idea of the Clustering using Unsupervised Binary Trees (CUBT) algorithm, considered by [2], which we adapt to functional data. In the following, we describe in detail the Functional Clustering Using Unsupervised Binary Trees (`fCUBT`) algorithm.

Building the maximal tree

In the following, a tree $\mathfrak{T}$ is a full binary tree which represents a nested partition of the sample $\mathcal{S}_N$, and $\mathfrak{D} \geq 1$ its depth. Let $\mathfrak{S}_{0,0}$ be the root node to which we assign the whole space sample $\mathcal{S}_N$. Every node $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}} \subset \mathcal{S}_{N_0}$ is indexed by the pair $(\mathfrak{d}, \mathfrak{j})$ where $0 \leq \mathfrak{d} < \mathfrak{D}$ is the depth index of the node and $0 \leq \mathfrak{j} < 2^{\mathfrak{d}}$ is the node index. A non-terminal node $(\mathfrak{d}, \mathfrak{j})$ has two children $\mathfrak{S}_{\mathfrak{d}+1,2\mathfrak{j}}$ and $\mathfrak{S}_{\mathfrak{d}+1,2\mathfrak{j}+1}$ such that $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}} = \mathfrak{S}_{\mathfrak{d}+1,2\mathfrak{j}} \cup \mathfrak{S}_{\mathfrak{d}+1,2\mathfrak{j}+1}$.

A tree $\mathfrak{T}$ is thus defined using a top-down procedure by recursively splitting. At each stage, a node $(\mathfrak{d}, \mathfrak{j})$ is possibly split into two subnodes provided it fulfills some condition. A MFPCA with $\mathtt{n_{comp}}$ components, $\mathtt{n_{comp}} \leq J$, is then conducted on the elements of $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}$. This results in a set of eigenvalues associated with a set of eigenfunctions. The matrix of scores $C_{\mathfrak{d},\mathfrak{j}}$ is then defined with the columns built with the projections of the elements of $\mathcal{S}_{\mathfrak{d},\mathfrak{j}}$ onto the set of eigenfunctions. For each $K = 1, \ldots, K_{max}$, we fit a Gaussian mixture model to the columns of the matrix $C_{\mathfrak{d},\mathfrak{j}}$ using an EM algorithm. The resulting models are denoted as $\{\mathcal{M}_1, \ldots, \mathcal{M}_{K_{max}}\}$. The number of groups within a node is determine using the BIC, $\widehat{K}_{\mathfrak{d},\mathfrak{j}} = \arg\max_{K=1,\ldots,K_{max}} \mathrm{BIC}(\mathcal{M}_K)$. If $\widehat{K}_{\mathfrak{d},\mathfrak{j}} > 1$, we split $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}$ using the model $\mathcal{M}_2$. Otherwise, the node is considered to be a terminal node and the construction of the tree is stopped for this node.

The recursive procedure continues downward until one of the following stopping rules are satisfied: there are less than `minsize` observations in the node or the estimation $\widehat{K}_{\mathfrak{d},\mathfrak{j}}$ of the number of clusters in the mode $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}$ is equal to 1. When the algorithm ends, a label is assigned to each leaf.

Joining step

In a perfect case, this tree will have the same number of leaves as the number of mixture components of X. In practice, it is rarely the case, and the number of leaves may be much larger than the number of clusters. That is why a joining step, which joins terminal nodes which do not necessarily share the same direct ascendant, should also be considered.

Let $\mathcal{G} = (V, E)$ be a graph where $V = \{\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}, 0 \leq j < 2^{\mathfrak{d}}, 0 \leq \mathfrak{d} < \mathfrak{D} \mid \mathfrak{S}_{\mathfrak{d},\mathfrak{j}} \text{ is a terminal node}\}$ is a set of vertices and

$$E = \left\{(\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}, \mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'}) \mid \mathfrak{S}_{\mathfrak{d},\mathfrak{j}}, \mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'} \in V,\ \mathfrak{S}_{\mathfrak{d},\mathfrak{j}} \neq \mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'} \text{ and } \widehat{K}_{(\mathfrak{d},\mathfrak{j})\cup(\mathfrak{d}',\mathfrak{j}')} = 1\right\}$$

is a set of edges. $\widehat{K}_{(\mathfrak{d},\mathfrak{j})\cup(\mathfrak{d}',\mathfrak{j}')}$ is the estimation of the number of clusters in $\mathfrak{S}_{\mathfrak{d},\mathfrak{j}} \cup \mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'}$ using the same methodology than the one in the previous step.

For each element $(\mathfrak{S}_{\mathfrak{d},\mathfrak{j}}, \mathfrak{S}_{\mathfrak{d}',\mathfrak{j}'})$ of E, we associate the value of the BIC that corresponds to $\widehat{K}_{(\mathfrak{d},\mathfrak{j})\cup(\mathfrak{d}',\mathfrak{j}')}$. The edge of $\mathcal{G}$ that corresponds to the maximum of the BIC is then removed and the associated vertices are joined. Thus, there is one cluster less. This procedure is run recursively until no pair of nodes can be joined or only one node in the tree remains.

Once the partition $\mathcal{U}$ has been created, we can classify new observations. This classification is performed by descending the tree $\mathcal{T}$ and by computing the probabilities to belong to each of the classes at each node.

4 Empirical analysis

We show the performance of our algortihm on an example. Let $K = 5$, $P = 2$, $_1 =_2= [0, 1]$. An independent sample of $N = 1000$ bivariate curves is simulated according to the following model : for $t_1, t_2 \in [0, 1]$,

$$\begin{array}{ll}
\text{Cluster 1: } X^{(1)}(t_1) = h_1(t_1) + b_{0.9}(t_1), & X^{(2)}(t_2) = h_3(t_2) + 1.5 \times b_{0.8}(t_2), \\
\text{Cluster 2: } X^{(1)}(t_1) = h_2(t_1) + b_{0.9}(t_1), & X^{(2)}(t_2) = h_3(t_2) + 0.8 \times b_{0.8}(t_2), \\
\text{Cluster 3: } X^{(1)}(t_1) = h_1(t_1) + b_{0.9}(t_1), & X^{(2)}(t_2) = h_3(t_2) + 0.2 \times b_{0.8}(t_2), \\
\text{Cluster 4: } X^{(1)}(t_1) = h_2(t_1) + 0.1 \times b_{0.9}(t_1), & X^{(2)}(t_2) = h_2(t_2) + 0.2 \times b_{0.8}(t_2), \\
\text{Cluster 5: } X^{(1)}(t_1) = h_3(t_1) + b_{0.9}(t_1), & X^{(2)}(t_2) = h_1(t_2) + 0.2 \times b_{0.8}(t_2),
\end{array}$$

where $h_1(t) = (6 - |20t - 6|)_+/4$, $h_2(t) = (6 - |20t - 14|)_+/4$ and $h_3(t) = (6 - |20t - 10|)_+/4$, for $t \in [0, 1]$. The functions b_H are defined, for $t \in [0, 1]$, by $b_H(t) = (1+t)^{-H} B_H(1+t)$ where $B_H(\cdot)$ is a fractional Brownian motion with Hurst parameter H. The mixing proportions are set to be equal. The data to which we apply the clustering is obtained as in (2). Each component curve is observed at 101 equidistant points in $[0, 1]$. The bivariate error vectors have zero-mean Gaussian

independent components with variance 1/2. For each $n \in \{1, \dots, N\}$, we observe a realization of the vector $X = (X^{(1)} + \alpha X^{(2)}, X^{(2)})^\top$, where $\alpha = 0.4$.

Our procedure is compared to `FunHDDC` [6], `Funclust` [5] and k-means [4] on the curves (k-`means`-d_1) and their derivatives (k-`means`-d_2). We also compare our algorithm with a GMM on the coefficients of a MFPCA (`FPCA+GMM`) and also our algorithm without the joining step (`Growing`). Our algorithm exhibits good performance in terms of Rand index (cf. Figure 1) and estimation of the number of clusters (cf. Table 1).

Method	3-	4	5	6	7+
`fCUBT`	-	-	0.664	0.238	0.098
`Growing`	-	-	0.604	0.182	0.214
`FPCA+GMM`	-	-	0.414	0.396	0.19
`FunHDDC`	1	-	-	-	-
`Funclust`	0.248	0.192	0.200	0.196	0.164
k-`means`-d_1	-	-	0.034	0.144	0.822
k-`means`-d_2	0.014	0.094	0.874	0.010	0.008

Table 1: Number of clusters

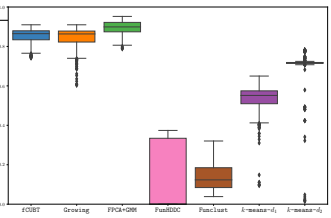


Figure 1: Rand index

The `fCUBT` algorithm was introduced above for multivariate functional data which could be defined on different domains, possibly of different dimensions, *e.g.* $\mathcal{T} = [0, 1]^2$. In such situations, the eigendecomposition of image data can be performed using the FCP-TPA algorithm for regularized tensor decomposition [1] and be used in the MFPCA.

Acknowledgements: The authors wish to thank Groupe Renault and the ANRT for their financial support via the CIFRE convention no. 2017/1116.

Bibliography

[1] G. I. Allen. Multi-way functional principal components analysis. In 2013, *5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 220-223, 2013.

[2] R. Fraiman, B. Ghattas and M. Svarc. Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification*, 7, 2013.

[3] C. Happ and S. Greven. Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, 113 649-659, 2018.

[4] F. Ieva, A. M. Paganoni, D. Pigoli, and V. Vitelli. Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *Journal of the Royal Statistical Society* Series C (Applied Statistics), 62(3):401-418, 2013.

[5] J. Jacques and C. Preda Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, 71 92-106, 2014.

[6] A. Schmutz, J. Jacques, C. Bouveyron, L. Cheze and P. Martin. Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, 2020.

[7] F. Yao, H.-G. Müller and J.-L. Wang. Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, 100 577-590, 2005.

Comparison of statistics for testing changed segment in a sample

Jovita Gudan[1*]

[1]*Department of Mathematics and Informatics, Vilnius University*

Abstract: We propose a new test statistic for detecting a changed segment in the mean of a sample at unknown dates. Test statistic is based on the adaptive self-normalized partial sums process. Our aim is to detect a short length changed segment. The new test statistic is compared with maximal ratio statistic by generating samples from symmetrized Pareto and Log-Gamma distributions.

Keywords: change-point detection, changed segment in the mean, epidemic change, Hölder norm statistics, regularly varying random variables.

AMS subject classification: 62G10, 60F17.

1 Introduction

For each $n \geq 1$ we consider the model

$$Y_{nk} = \mu_{nk} + X_k, \quad k = 1, 2, \ldots, n, \tag{1}$$

where $(\mu_{nk}) \subset \mathbb{R}$ and $(X_k, k \geq 1)$ is a sequence of independent identically distributed (i.i.d.) zero mean random variables with its generic element denoted by X. We want to test the null hypothesis

$$H_0 : \quad \mu_{n1} = \cdots = \mu_{nn} = 0 \tag{2}$$

against the alternative

$$H_A : \quad \mu_{nk} = \mu_n(k) 1_{I^*}(k), \quad k = 1, \ldots, n, \tag{3}$$

where $I^* := \{k^* + 1, \ldots, k^* + \ell^*\}$ for some $k^* = k_n^* \geq 0$, $\ell^* = \ell_n^* \in (0, n)$ and $1_{I^*}(k) = 1$ if $k \in I^*$ and 0 otherwise. The interval I^* represents a changed segment in the sample. We are interested in detection short length changed segment thus assuming ℓ^*/n to be small as n is large.

The test statistics are constructed from moving sums

$$S_{k,m} = X_{n,k+1} + \cdots + X_{n,m}, \quad 0 \leq k < m \leq n,$$

*Corresponding author: jovita.gudan@mif.vu.lt

and moving sums of squares

$$V_{k,m}^2 = X_{n,k+1}^2 + \cdots + X_{n,m}^2, \quad 0 \le k < m \le n.$$

We consider the class of functions $\{\rho_{\gamma,\beta}, 0 \le \gamma < 1, \beta > 0\}$, where

$$\rho_{\gamma,\beta}(h) = h^\gamma \log^\beta(1 + 1/h), \quad h \in (0,1)$$

and following [2] we define

$$T_{\gamma,\beta,n}^{ad} := V_{0,n}^{-1} \max_{0 \le k < m \le n} \frac{|S_{k,m}|}{\rho_{\gamma,\beta}(V_{k,m}^2/V_{0,n}^2)}. \tag{4}$$

The upper script ^{ad} means adaptive (see [2] for such reason). Asymptotic behaviour of $T_{\gamma,\beta,n}^{ad}$ is contained in the following results.

Theorem 1. *Let either $0 \le \gamma < 1/2$ and $\beta \ge 0$ or $\gamma = 1/2$ and $\beta > 1/2$. Assume for the model (1) that either X is symmetric and $EX^2 < \infty$ or $E|X|^q < \infty$ for some $q > 2$. Then under the null hypothesis (2), for any $x > 0$,*

$$\lim_{n\to\infty} P(T_{\gamma,\beta,n}^{ad} > x) = P(T_{\gamma,\beta} > x), \tag{5}$$

where

$$T_{\gamma,\beta} = \sup_{0 \le s < t \le 1} \frac{|W(t) - W(s)|}{\rho_{\gamma,\beta}(|t-s|)} \tag{6}$$

and $(W(t), t \in [0,1])$ is a standard Wiener process.

Proof. Denote $\tau_{nk} = V_{0,k}^2/V_{0,n}^2, k = 0, 1, \ldots, n$. Consider the adaptive polygonal line process $\zeta_n^{ad}(t), 0 \le t \le 1$, which is obtained by linear interpolation of the points $(0,0), (S_{0,k}, \tau_{nk}), \quad k = 1, \ldots, n$. Under H_0 we have $X_{ni} = X_i, i = 1, \ldots, n$. In this case it follows from [2], that

$$V_{0,n}^{-1} \zeta_n^{ad} \xrightarrow[n\to\infty]{\mathscr{D}} W$$

in the Hölder space $H_{\gamma,\beta}^o[0,1]$ (see, [2] for definitions). By the continuous mapping theorem it follows then

$$V_{0,n}^{-1} ||\zeta_{\gamma,\beta,n}^{ad}||_\gamma \xrightarrow[n\to\infty]{\mathscr{D}} ||W||_{\gamma,\beta}. \tag{7}$$

It is known (see, [2]), that the Hölder norm of polygonal line function is attained at some vertexes. Hence the result follows from (7). □

Suppose that $\widehat{T} := \widehat{T}_{\gamma,\beta,n}^{ad}$ is the observed value of a test statistic $T := T_{\gamma,\beta,n}^{ad}$. Then the P-value of $\widehat{T}$ is

$$p(\widehat{T}) = 1 - P(T \le \widehat{T}) = P(T > \widehat{T}).$$

If we knew the distribution of the statistics T, we would simply calculate $p(\widehat{T})$ and reject the null whenever $p(\widehat{T}) < \alpha$. Since we do not know the probability $P(T > x)$ it is common to use either an asymptotic approximation $P(T > x) \approx P(T^{ad}_{\gamma,\beta} > x)$, where the probability $P(T_{\gamma,\beta} > x)$ can be calculated by Monte-Carlo. So that

$$p(\widehat{T}) \approx \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}\{T^{(j)}_{\gamma,\beta} > \widehat{T}),$$

where $(T^{(j)}_{\gamma,\beta}, j = 1, \dots, N)$ are simulated by using independent standard Wiener processes $W_1, \dots W_N$. The distribution of $T^N_{\gamma,\beta}$ was evaluated on a grid of size 1000 and by running a Monte-Carlo simulations with 400 runs. The results are presented in Table 1 given in the Appendix.

Assuming $n = 4m$ we divide the sample $X_{n1}, \dots, X_{nn}$ into four equal parts and following [1] define for $\gamma \geq 0$ the maximal ratio statistics as follows

$$\mathrm{MR}_{\gamma,n} := \max\left\{\frac{T_{n1}}{T_{n3}}, \frac{T_{n3}}{T_{n1}}, \frac{T_{n2}}{T_{n4}}, \frac{T_{n4}}{T_{n2}}\right\}, \tag{8}$$

where

$$T_{nj} = T^{(\gamma)}_{nj} := \max_{1\leq \ell \leq m} \ell^{-\gamma} \max_{(j-1)m \leq k \leq jm-\ell} |S_{k,\ell}|\,, \quad j = 1, \dots, 4.$$

The asymptotic behavior in distribution of $\mathrm{MR}_{\gamma,n}$ depends on the threshold $\gamma = 1/2 - 1/a$. As proved in [1]:

- when $1/2 - 1/a < \gamma < 1/2$, then

$$\lim_{n\to\infty} P(\mathrm{MR}_{\gamma,n} > x) = \begin{cases} 1 & \text{if } x < 1, \\ 4x^a(1+x^a)^{-2} & \text{if } x \geq 1; \end{cases} \tag{9}$$

- when $0 < \gamma < 1/2 - 1/a$, then

$$\lim_{n\to\infty} P(\mathrm{MR}_{\gamma,n} > x) = \begin{cases} 1 & \text{if } x < 1 \\ 1 - \left[1 - 2\int_0^\infty F_\gamma(xy)\,\mathrm{d}F_\gamma(y)\right]^2 & \text{if } x \geq 1, \end{cases} \tag{10}$$

where $F_\gamma(x) = P(T_{\gamma,0} > x)$.

2 Comparison of statistics by simulation study

To compare the finite sample behavior of tests $T^{ad}_{\gamma,\beta,n}$ assuming $\beta = 0$ and $\gamma < 1/2$ and $\mathrm{MR}_{\gamma,n}$ assuming $\gamma < 1/2$ for independent random variables, we conduct an experiment using Monte-Carlo simulations. We use the symmetrized Pareto and Log-Gamma distributions to generate X regularly varying with index a. Data are generated by the model $X_{nk} = \mu^*_n 1_{\{k^*+1,\cdots,k^*+\ell^*\}}(k) + X_k$.

The performance of $T^{ad}_{\gamma,\beta,n}$ depends on a parameter γ, which is examined under variant values. Figure 1 portrays the empirical power of test (4) under various $\gamma = 0.02, 0.04, \ldots, 0.45$ values with tail index $a = 5$, nominal size $\alpha = 0.05$, sample size $n = 1000$, changed mean segment $\mu^*_n = 0.2$ and 1000 Monte-Carlo simulations. Consequently, without loss of generality, the remaining Monte-Carlo simulations for the Adopted test (4) are conducted with parameter β equal to 0.

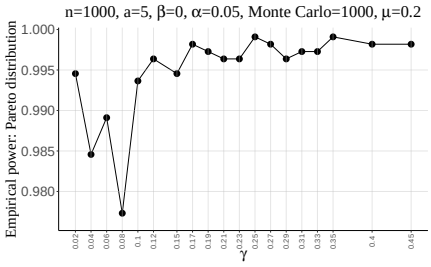
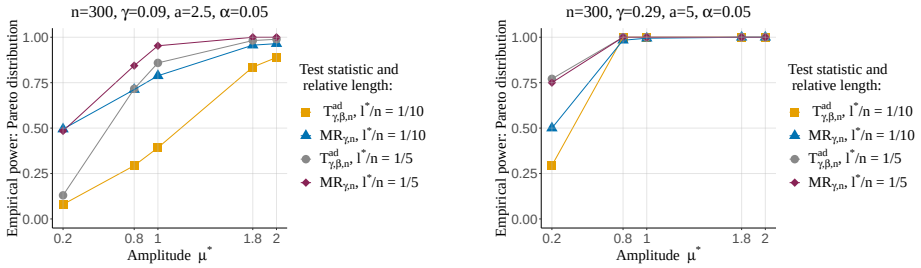


Figure 1: Empirical power of $T^{ad}_{\gamma,\beta,n}$ test under the similar settings $n = 1000, \alpha = 0.05$ from Pareto distribution

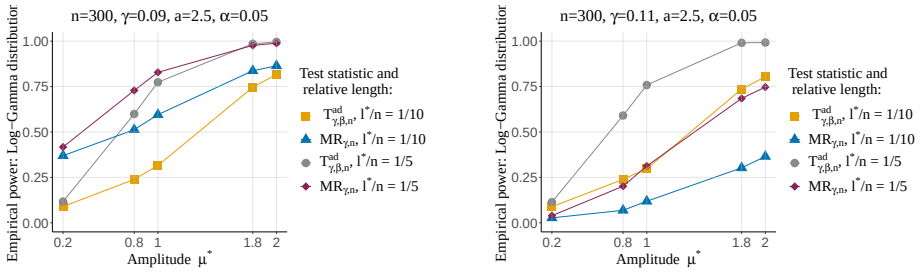


(a) Case: $a = 2.5$ (b) Case: $a = 5$

Figure 2: Empirical power of the tests under the similar settings with $n = 300, \alpha = 0.05$ from Pareto distribution

Now, we investigate the relationship between $T^{ad}_{\gamma,\beta,n}$ and $\mathrm{MR}_{\gamma,n}$ tests under various settings. First, to analyze the effect of the tail index a on empirical power scores, we reflect on only two values of parameter of regular variation $a = 2.5, 5$ as even bigger values of a cause higher empirical power of tests (4) and (8). Figure 2 depicts the empirical power for different values of a for the samples generated from Pareto distribution with $n = 300$, $\ell^* = 30, 60, \mu^*_n = 0.2, 0.8, 1, 1.8, 2$ and $\alpha = 0.05$. For both tests $T^{ad}_{\gamma,\beta,n}$ and $\mathrm{MR}_{\gamma,n}$, it can be seen that the greater the index a, the amplitude of the change μ^*_n or the length of the changed segment ℓ^*, the bigger the empirical power of the test. These two tests (4) and (8) are linked by threshold $\gamma = 1/2 - 1/a$ as introduced in (9) and (10). Therefore, in order to investigate the effect at the different two settings, we introduce the results of the empirical power

of the tests (4) and (8) under the events of $0 < \gamma < 1/2 - 1/a$ with $\gamma = 0.09$ and $1/2 - 1/a < \gamma < 1/2$ with $\gamma = 0.11$ for both $a = 2.5, n = 300, \alpha = 0.05$, $\ell^* = 30, 60$ and $\mu_n^* = 0.2, 0.8, 1, 1.8, 2$ as illustrated in Figure 3. In this case, the threshold $1/2 - 1/a = 0.1$ divides $\mathrm{MR}_{\gamma,n}$ limiting distribution to (9) and (10) and for $T^{ad}_{\gamma,\beta,n}$ limiting distribution (5) is the same for both cases. Remarkably, the performance of the test $T^{ad}_{\gamma,\beta,n}$ is better than $\mathrm{MR}_{\gamma,n}$ when $1/2 - 1/a < \gamma < 1/2$ and vice versa when $0 < \gamma < 1/2 - 1/a$.



(a) Case: $0 < \gamma < 1/2 - 1/a$ (b) Case: $1/2 - 1/a < \gamma < 1/2$

Figure 3: Empirical power of the tests under the similar settings with $a = 2.5, n = 300, \alpha = 0.05$ from Log-gamma distribution

Table 2 given in the Appendix summarizes the performance of tests $\mathrm{MR}_{\gamma,n}$ and $T^{ad}_{\gamma,\beta,n}$ under the Log-Gamma and Pareto distributions with regular variation index $a = 5$, sample size of $n = 200, 600, 1000$, amplitude of the change $\mu_n^* = 0.3$, length ℓ^* of approximately of $\frac{1}{7}, \frac{1}{4}$ the sample size n. From the results, it can be seen that increasing sample size, length and amplitude of the change also increase the empirical power.

In conclusion, this paper introduces a new $T^{ad}_{\gamma,\beta,n}$ test that is compared with $\mathrm{MR}_{\gamma,n}$ test. Simulation study shows the power of the test $T^{ad}_{\gamma,\beta,n}$ is better than $\mathrm{MR}_{\gamma,n}$ when $1/2 - 1/a < \gamma < 1/2$ and vice versa when $0 < \gamma < 1/2 - 1/a$.

Acknowledgements: I convey my sincere gratitude to my Academic Supervisor Prof. Habil. dr. Alfredas Račkauskas. The completion of this undertaking could not have been possible without his participation and assistance.

Bibliography

[1] Gudan, J. and Račkauskas, A. and Suquet, Ch. (2021). Testing mean changes by maximal ratio statistics. *Extremes* (accepted for publication).

[2] Račkauskas, A. and Suquet, Ch. (2001). Invariance principles for adaptive self-normalized partial sums processes. *Stochastic Processes and their Applications* 95(1), 63–81.

Appendix

α/γ	**0.01**	**0.09**	**0.11**	**0.2**	**0.29**	**0.31**	**0.49**
0.025	2.6195	2.5592	2.4319	2.6152	2.7078	2.8187	3.7913
0.05	2.2485	2.1404	2.1894	2.4627	2.5157	2.6608	3.6081
0.1	1.9382	1.9404	1.9406	2.1870	2.3034	2.4437	3.3937

Table 1: Asymptotic critical values of $T^{ad}_{\gamma,\beta,n}$ for $\beta = 0$

			$n = 200$		$n = 600$		$n = 1000$	
			$\ell^* = 35$	$\ell^* = 45$	$\ell^* = 80$	$\ell^* = 140$	$\ell^* = 165$	$\ell^* = 235$
Log-Gamma	$\mathrm{MR}_{\gamma,n}$	$\gamma = 0.01$	0.4216	0.4420	0.4850	0.6581	0.6873	0.8006
		$\gamma = 0.29$	0.7026	0.7130	0.7403	0.8567	0.8549	0.9415
		$\gamma = 0.31$	0.4712	0.5310	0.5405	0.7194	0.7108	0.8165
	$T^{ad}_{\gamma,\beta,n}$	$\gamma = 0.01$	0.1387	0.1075	0.2237	0.4815	0.44534	0.7131
		$\gamma = 0.29$	0.1814	0.1329	0.3181	0.6254	0.6297	0.8731
		$\gamma = 0.31$	0.1273	0.1084	0.2803	0.5812	0.5939	0.8227
Pareto	$\mathrm{MR}_{\gamma,n}$	$\gamma = 0.01$	0.8522	0.9290	0.9472	0.9970	0.9951	1
		$\gamma = 0.29$	0.9514	0.9780	0.9833	1	1	1
		$\gamma = 0.31$	0.8899	0.9250	0.9419	0.9990	0.9951	1
	$T^{ad}_{\gamma,\beta,n}$	$\gamma = 0.01$	0.5888	0.8094	0.9039	0.9984	0.9991	1
		$\gamma = 0.29$	0.8006	0.9349	0.9946	1	1	1
		$\gamma = 0.31$	0.7664	0.9198	0.9865	1	1	1

Table 2: Empirical power of $T^{ad}_{\gamma,\beta,n}$ and $\mathrm{MR}_{\gamma,n}$ statistical tests under constant change in mean, $\mu^*_n = 0.3$, $a = 5$, $\alpha = 0.05$ and $\beta = 0$

COVID-19: Study of the spread of the pandemic in Bulgaria

Silvi-Maria Gurova[1*]

[1]*Institute of Information and Communication Technologies - Bulgarian Academy of Sciences Acad. G. Bonchev str., block 25A, Sofia, Bulgaria*

Abstract: The aim of this work is to predict the future spread of COVID-19 in Bulgaria based on the real epidemiological data for a year ago. The data are taken from the unified information portal for Bulgaria where they are published daily. To study the spread of the epidemic in the country and to find some dependencies, we divide the country into two regions: (i) the Sofia-city and (ii) the province. To predict the spread of the disease in Bulgaria for a few months ahead we use a statistical forecasting package Prophet in R software. Our results show that the spread of the disease in the province depends on its spread in the capital.

Keywords: data, COVID-19, time series, statistics, predictions
AMS subject classification: 62P10, 65C20

1 Introduction

At the end of 2019 in the city Wuhan, China, COVID-19 was identified, which manifested itself as the new pandemic of the 21st century, affecting the world for over a year. To study the behavior of the COVID-19 pandemic, the scientists usually use two approaches: (i) creating mathematical epidemiological models describing the stages and disease development in the population and (ii) analyzing statistics data to predict the development of the disease. In the first approach, researchers apply epidemiological models such as SI, SIR, SEIR, SEIRS, describing the spread of viral diseases [1],[2]. We note that these epidemiological models can be applied not only for description of pandemics among humans, but also in the spread of viral diseases to animal species [4]. In the second approach, researchers analyze COVID-19 data using statistical and machine-inspired time series methods as [5]. With the advent of COVID-19, a lot of scientists apply the two approaches mention above to describe the spread of pandemic in a given geographical area [3], [5].

In this work we use the real COVID-19 data for Bulgaria, provided by the unified information portal for COVID-19 for Bulgaria [6]. To analyze the spread of the pandemic in the country, we consider how the data are distributed in the two regions: (i) in the capital and (ii) in the province. The Prophet package in R

*Corresponding author: smgurova@parallel.bas.bg

software is used to predict the spread of the disease in Bulgaria for a few months ahead. The paper is organized as following: In the second section the spread of the COVID-19 pandemic is studied for two regions in Bulgaria. Using the coronavirus data for one year period in Bulgaria, in the third section we apply an Automatic Forecasting Procedure described in the Prophet package in order to predict the spread of the pandemic for two months ahead. The last section summaries the results and outline the future work.

2 Analysis of COVID-19 data for Bulgaria

Let us introduce a "daily infection rate" in percentages, which is defined as the ratio between "daily positive cases" and "daily tests performed" multiplied by 100. Let us introduce also "7-daily infection rate" and "14-daily infection rate" by averaging the values of the "daily infection rate" per seven and fourteen days, respectively. To analyse the spread of the infection in Bulgaria for the period from 1^{st} of June 2020 to 3^{rd} of May 2021, we consider the data from the official governmental COVID-19 portal [6], where they are aggregated daily. We note the restrictions imposed by the Bulgarian government to reduce the spread of infection were too weak even when the waves reach their peaks, while many European countries, including Bulgaria's neighbors went into full lockdown.

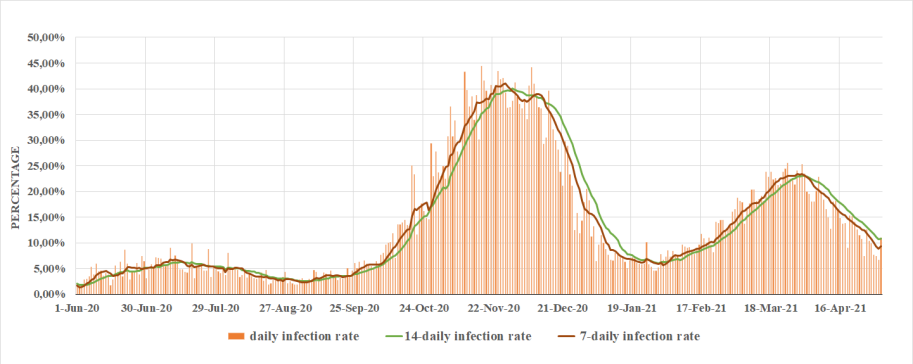
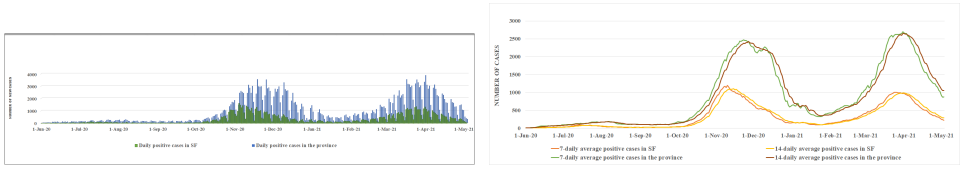


Figure 1: Percentage of positive cases with respect to the tests performed on a daily, 7-daily and 14-daily basis

The trend of the spread of the coronavirus infection by a daily infection rate, 7-daily infection rate and 14-daily infection rate is presented in Figure 1 in the country for the period of 11 months while in Figure 2 we compare the daily, 7-daily and 14-daily positive cases in Sofia-city and the province.

The presented pictures clearly demonstrate three waves of the spread of the infection with 3 peaks around the end of July 2020, the end of November 2020 and the end of March 2021. We see that the third wave is twice as weak as the second wave, see Figure 1. This is explained by twice as many tests performed and that more than 1 million people have already met the virus and built up immunity,



(a) (b)

Figure 2: Comparison of (a) daily positive cases in the Sofia-city and the province and (b) 7-daily and 14-daily average positive cases in the Sofia-city and the province

although the government lifted almost all restrictions during the third wave. In Figure 2 we observe that in the second wave the peak of the spread of the infection in the province is delayed by two weeks, while in the third wave the two peaks of the spread of the infection for the Sofia-city and the province are reached at one and the same time. This can be explained by the fact that in the second wave there is a diffuse spread of the infection from the Sofia-city toward the province while in the third wave the change of the reproductive number is with one and the same rate at weakened measures for the whole country.

3 Forecasting the coronavirus pandemic by using Prophet

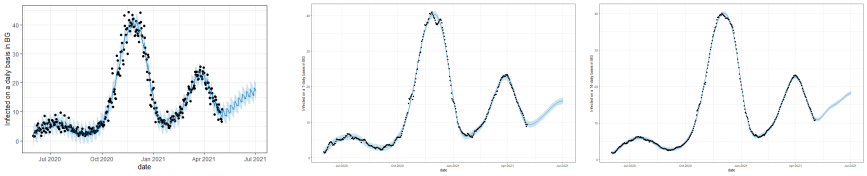
The analysis of time series is focused on the study of past observations of a random variable $\phi(x) = \phi(x_1, x_2, \ldots, x_n)$ in order to find a good statistical model that fits the data. When that model is found, it can be used to predict the future values of the random variable. In this section we consider a Prophet package from the software R with which we can predict the spread of the infection in Bulgaria.

Using the object predict from the Prophet package in the software R, we can predict the future values of n-daily infected rate (n=1, 7, 14). In our tests with Prophet, we use a time series for the given period from 11 months and the future values for forecasting are 60 days ahead, i.e. up to 2^{nd} of July 2021. In Table 1 the three latest estimated future values of the n-daily infected rates in Bulgaria are presented, where the supposed estimated value is $\hat{y}$ which belongs to the confidence interval $[\hat{y}_{lower}, \hat{y}_{upper}]$.

The results presented in Figure 3 and Figure 4 show that 4^{th} wave is born in the next 2 months for the both regions–the Sofia-city and the province. We believe that the peak of this 4^{th} wave will be small and depends on the successfully applying the vaccination plan in Bulgaria.

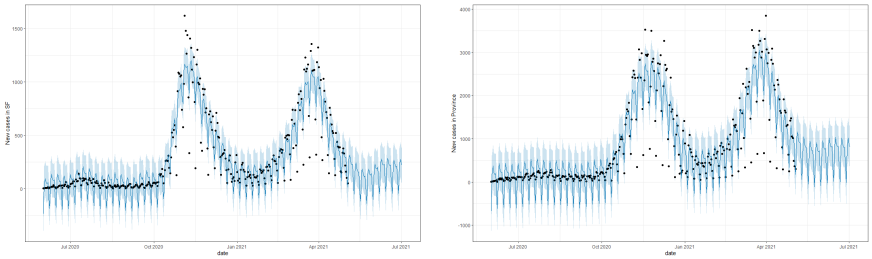
date		$\hat{y}$	$\hat{y}_{lower}$	$\hat{y}_{upper}$
2021-06-30	daily infected rate	19.97	15.45	20.56
2021-07-01		17.56	14.84	20.10
2021-07-02		16.97	14.28	19.63
2021-06-30	7-daily infected rate	16.00	15.18	16.94
2021-07-01		16.05	15.21	16.90
2021-07-02		16.09	15.19	16.95
2021-06-30	14-daily infected rate	18.16	17.65	18.69
2021-07-01		18.24	17.76	18.76
2021-07-02		18.31	17.76	18.83

Table 1: Predictive values in the last three days of 60-daily period for the n-daily infected rates



(a) (b) (c)

Figure 3: Forecast graphs for (a) daily infection rate, (b) 7-daily infection rate and (c) 14-daily infection rate for Bulgaria



(a) (b)

Figure 4: Forecast graphs for predicting the new positive cases in the two regions: (a) the Sofia–city and (b) the province

4 Conclusion

In this study we present the spread of the coronavirus pandemic in the country, the Sofia–city and the province taking the real data from the Bulgarian information portal [6] for almost year ago. Some dependence was observed in the peaks of the second and third wave of the regions under considerations. The Prophet package was used to predict the spread of the disease in Bulgaria for two months ahead.

date		$\hat{y}$	$\hat{y}_{lower}$	$\hat{y}_{upper}$
2021-06-30	New positive cases in Sofia-city	277	117.42	443.21
2021-07-01		234	62.78	393.64
2021-07-02		221	46.37	388.29
2021-06-30	New positive cases in the province	1018	527.23	1517.14
2021-07-01		861	393.80	11329.62
2021-07-02		814	347.11	1319.68

Table 2: Forecasting values of the new positive cases in the two regions for the last three days of 60-daily period

The results show that this statistical package can be successfully used to predict new wave. As a future work we plan to use other time series forecasting methods from the software R such as ARIMA, HWAAS and others to predict the spread of the pandemic in Bulgaria.

Acknowledgements: This work has been accomplished with the partial support by the Grant No.BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014-2020) and co-financed by the European Union through the European structural and Investment funds and it has been also supported by the financial funds allocated to the Sofia University "St. Kl. Ohridski", grant No.80-10-87/2021.

Bibliography

[1] Abouelkheir I., et al., "Optimal Impulse Vaccination Approach for an SIR Control Model with Short-Term Immunity", Mathematics, MDPI, 7(5):420, (2019) https://doi.org/10.3390/math7050420.

[2] Kamrujjaman M., et al., "Pandemic and the Dynamics of SEIR Model: Case COVID-19". Preprints (2020), 2020040378 (doi: 10.20944/preprints202004.0378.v1).

[3] Margenov S., et al., "Mathematical and computer modeling of COVID-19 transmission dynamics in Bulgaria by time-depended inverse SEIR model", AIP Conference Proceedings 2333, 090024 (2021) https://doi.org/10.1063/5.0041868.

[4] Gurova S.-M., "A predator-prey model with SEIR and SEIRS epidemic in the prey", AIP Conference Proceedings 2164, 080003 (2019) https://doi.org/10.1063/1.5130826.

[5] Papastefanopoulos V, et al., "COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population", Applied Sciences, 2020; 10(11):3880. https://doi.org/10.3390/app10113880.

[6] Official Bulgarian information portal for combating the spread of COVID-19, https://coronavirus.bg/bg/statistika(Availableat08.07.2021).

Reversible genetically modified mode jumping MCMC

Aliaksandr Hubin,[1*] **Florian Frommlet**[2] **and Geir Storvik**[1]

[1]*University of Oslo*
[2]*Medical University of Vienna*

Abstract: In this paper, we introduce a reversible version of a genetically modified mode jumping Markov chain Monte Carlo algorithm (GMJMCMC) for inference on posterior model probabilities in complex model spaces, where the number of explanatory variables is prohibitively large for classical Markov Chain Monte Carlo methods. Unlike the earlier proposed GMJMCMC algorithm, the introduced algorithm is a proper MCMC and its limiting distribution corresponds to the posterior marginal model probabilities in the explored model space under reasonable regularity conditions.

Keywords: Markov chain Monte Carlo; Mode jumping in MCMC; Genetic algorithms; Bayesian Model selection; Bayesian Model averaging

AMS subject classification: 62-02, 62-09, 62F07, 62F15, 62J12, 62J05, 62J99, 62M05, 05A16, 60J22, 92D20, 90C27, 90C59

1 Introduction

A genetically modified Markov chain Monte Carlo algorithm (GMJMCMC) was introduced [3, 6, 5] for Bayesian model selection/averaging problems when the total number of covariates (including functions of covariates) is prohibitively large. Applications include GWAS studies with Bayesian generalized linear models [3] as well as Bayesian logic regressions (BLR) [5] and Bayesian generalized nonlinear models (BGNLM) [6]. If certain regularity conditions are met, the GMJMCMC algorithm will asymptotically explore all models in the defined model spaces. However, GMJMCMC is not a proper MCMC algorithm in the sense that its limiting distribution does not correspond to the marginal posterior model probabilities and thus only renormalized estimates of these probabilities [2, 4] can be obtained. In this paper, we introduce a reversible genetically modified Markov chain Monte Carlo algorithm (RGMJMCMC), which modifies GMJMCMC to become a proper MCMC method providing marginal posterior probabilities directly as Monte Carlo estimates.

*Corresponding author: aliaksah@math.uio.no

2 The algorithm

Genetically Modified MJMCMC

Consider the case of a fixed predefined set of correlated features (covariates). Then the general model space $\mathcal{M}$ is of size 2^q and standard MCMC algorithms tend to get stuck in local maxima if there are correlations between covariates [2, 4]. The basic idea of MJMCMC [4] is to make a large jump (changing many model components) followed by local optimization within the discrete model space to obtain a proposal. The MJMCMC algorithm requires that all q covariates defining the model space are potentially considered at each iteration of the algorithm. But if q is large, it becomes impossible to specify and store all 2^q models in $\mathcal{M}$. The idea behind GMJMCMC is to apply the MJMCMC algorithm iteratively to smaller sets of model components of size $s \ll q$. Here, s is specified to be larger or equal to the maximal possible size Q of the assumed true model in the defined model space, which is a necessary condition to be able to use GMJMCMC [5]. This constraint also reduces the number of models in the model space $\mathcal{M}$ to $\sum_{k=1}^{Q} \binom{q}{k}$. As shown in Theorem 1 in [5], GMJMCMC is irreducible in the defined model space of models of size up to s under some easy to satisfy regularity conditions. Yet, GMJMCMC is not a proper MCMC and one cannot use the frequencies of different models in the Markov chain to estimate their posteriors. Instead, we use the renormalized estimates [2, 4, 5, 6] from a subspace $\mathcal{M}^* \subset \mathcal{M}$:

$$\widehat{p}(\mathfrak{m}|\mathbf{y}) = \frac{p(\mathfrak{m})p(\mathbf{y}|\mathfrak{m})}{\sum_{\mathfrak{m}' \in \mathcal{M}^*} p(\mathfrak{m}')p(\mathbf{y}|\mathfrak{m}')} \, \mathrm{I}(\mathfrak{m} \in \mathcal{M}^*) \, , \tag{1}$$

which asymptotically converge to $p(\mathfrak{m}|\mathbf{y})$ as the number of iterations grows.

In GMJMCMC, we let $\mathcal{F}_0$ be all q input features and $\mathcal{S}_0 \subseteq \mathcal{F}_0$ be some subset of them. Then, throughout our search we generate a sequence of so called *populations* $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_{T_{max}}$. Each $\mathcal{S}_t$ is a set of s features and forms a separate *search space* for exploration through MJMCMC iterations. Populations dynamically evolve allowing GMJMCMC to explore different parts of the total model space. Algorithm 2 in [5] summarizes this procedure. The generation of $\mathcal{S}_{t+1}$ given $\mathcal{S}_t$ works as follows: Members of the new population $\mathcal{S}_{t+1}$ are generated by applying certain transformations to components of $\mathcal{S}_t$. First, some components with low frequency from search space $\mathcal{S}_t$ are removed using a *filtration* operator. The removed components are then replaced, where each replacement is generated randomly by a *mutation* operator with probability P_m, by a *crossover* operator with probability P_c, by a *modification* operator with probability P_t or by a *projection* operator with probability P_p, where $P_c + P_m + P_t + P_p = 1$. The operators to generate potential features of $\mathcal{S}_{t+1}$ are formally defined in [3, 5, 6].

Reversible Genetically Modified MJMCMC

The GMJMCMC algorithm described above cannot guarantee that the ergodic distribution of its Markov chain corresponds to the target distribution of interest [5]. An easy modification based on performing both forward and backward swaps between populations can provide a proper MCMC algorithm in the model space of

interest. Consider a transition $\mathfrak{m} \to \mathcal{S}' \to \mathfrak{m}'_0 \to ... \to \mathfrak{m}'_k \to \mathfrak{m}'$ with a given probability kernel. Here, $q_S(\mathcal{S}'|\mathfrak{m})$ is the proposal for a new population, transitions $\mathfrak{m}'_0 \to ... \to \mathfrak{m}'_k$ are generated by local MJMCMC within the model space induced by $\mathcal{S}'$, and the transition $\mathfrak{m}'_k \to \mathfrak{m}'$ is some randomization at the end of the procedure as described in the next paragraph. The following theorem shows the detailed balance equation for the suggested swaps between models.

Theorem 1. *Assume* $\mathfrak{m} \sim p(\cdot|\mathbf{y})$ *and* $(\mathcal{S}', \mathfrak{m}'_k, \mathfrak{m}')$ *are generated according to the proposal distribution* $q_S(\mathcal{S}'|\mathfrak{m})q_o(\mathfrak{m}'_k|\mathcal{S}', \mathfrak{m})q_r(\mathfrak{m}'|\mathcal{S}, \mathfrak{m}'_k)$. *Assume further* $(\mathcal{S}, \mathfrak{m}_k)$ *are generated according to* $\tilde{q}_S(\mathcal{S}|\mathfrak{m}', \mathcal{S}, \mathfrak{m})q_o(\mathfrak{m}_k|\mathcal{S}, \mathfrak{m}')$. *Let*

$$\mathfrak{m}^* = \begin{cases} \mathfrak{m}' & \textit{with probability } \min\{1, a_{mh}\}; \\ \mathfrak{m} & \textit{otherwise.} \end{cases}$$

where

$$a_{mh} = \frac{p(\mathfrak{m}'|\mathbf{y})q_r(\mathfrak{m}|\mathcal{S}, \mathfrak{m}_k)}{p(\mathfrak{m}|\mathbf{y})q_r(\mathfrak{m}'|\mathcal{S}', \mathfrak{m}'_k)}. \tag{2}$$

Then $\mathfrak{m}^* \sim p(\cdot|\mathbf{y})$.

Proof. Define $\bar{p}(\mathfrak{m}, \mathcal{S}', \mathfrak{m}'_k) \equiv p(\mathfrak{m}|\mathbf{y})q_{\mathcal{S}}(\mathcal{S}'|\mathfrak{m})q_o(\mathfrak{m}'_k|\mathcal{S}', \mathfrak{m})$. Then by construction $(\mathfrak{m}, \mathcal{S}', \mathfrak{m}'_k) \sim \bar{p}(\mathfrak{m}, \mathcal{S}, \mathfrak{m}'_k)$. Define $(\mathfrak{m}', \mathcal{S}, \mathfrak{m}_k)$ to be a proposal from the distribution $q_r(\mathfrak{m}'|\mathcal{S}, \mathfrak{m}'_k)q_S(\mathcal{S}|\mathfrak{m}')q_o(\mathfrak{m}_k|\mathcal{S}, \mathfrak{m}')$. Then the Metropolis-Hastings acceptance ratio becomes

$$\frac{\bar{p}(\mathfrak{m}', \mathcal{S}, \mathfrak{m}_k)q_r(\mathfrak{m}|\mathcal{S}, \mathfrak{m}_k)q_S(\mathcal{S}'|\mathfrak{m})q_o(\mathfrak{m}'_k|\mathcal{S}', \mathfrak{m})}{\bar{p}(\mathfrak{m}, \mathcal{S}', \mathfrak{m}'_k)q_r(\mathfrak{m}'|\mathcal{S}', \mathfrak{m}'_k)q_S(\mathcal{S}|\mathfrak{m}')q_o(\mathfrak{m}_k|\mathcal{S}, \mathfrak{m}')}$$

which reduces to a_{mh}. □

From Theorem 1, it follows that if the Markov chain is irreducible in the model space then it is ergodic and converges to the right posterior distribution. The described procedure marginally generates samples from the target distribution, i.e. the posterior model probabilities $p(\mathfrak{m}|\mathbf{y})$. Instead of using the approximation (1) one can get frequency-based estimates of the model posteriors $p(\mathfrak{m}|\mathbf{y})$. For a sequence of simulated models $\mathfrak{m}^1, ..., \mathfrak{m}^W$ from an ergodic MCMC algorithm with a stationary distribution $p(\mathfrak{m}|\mathbf{y})$ it holds that

$$\widetilde{p}(\mathfrak{m}|\mathbf{y}) = W^{-1} \sum_{i=1}^{W} \mathrm{I}(\mathfrak{m}^{(i)} = \mathfrak{m}) \xrightarrow[W\to\infty]{d} p(\mathfrak{m}|\mathbf{y}), \tag{3}$$

and similar results are valid for estimates of the posterior marginal inclusion probabilities or any other parameters of interest [2, 4].

In practice, proposals $q_{\mathcal{S}}(\mathcal{S}'|\mathfrak{m})$ are obtained as follows: First, all members of $\mathfrak{m}$ are included. Then additional features are added by the same operators as described in Section 2 but with $\mathcal{S}_t$ replaced by the population including all components in $\mathfrak{m}$. The randomization $\mathfrak{m}' \sim q_r(\mathfrak{m}|\mathcal{S}', \mathfrak{m}'_k)$ is performed by potential swapping of the features within $\mathcal{S}'$, each with a small probability ρ_r. Note that this might give a reverse probability $q_r(\mathfrak{m}|\mathcal{S}, \mathfrak{m}_k)$ being zero if $\mathcal{S}$ does not include all components in $\mathfrak{m}$. In that case, the proposed model is not accepted. Otherwise, the ratio of the proposal probabilities can be written as $\frac{q_r(\mathfrak{m}|\mathcal{S}, \mathfrak{m}_k)}{q_r(\mathfrak{m}'|\mathcal{S}', \mathfrak{m}'_k)} = \rho_r^{d(\mathfrak{m}, \mathfrak{m}_k) - d(\mathfrak{m}', \mathfrak{m}'_k)}$, where $d(\cdot, \cdot)$ is the Hamming distance (the number of components differing).

Delayed rejection

The computationally most demanding parts of the RGMJMCMC algorithm are the forward and backward MCMC searches. Often, the proposals generated by forward search have a very small probability $\pi(\mathfrak{m}')$ resulting in a low acceptance probability regardless of the way the backward auxiliary variables are generated. In such cases, one would like to reject directly without performing the backward search. This is achieved by the delayed acceptance procedure [1] which can be applied in our case due to the following result:

Theorem 2. *Assume* $\mathfrak{m} \sim p(\cdot|\mathbf{y})$ *and* $\mathfrak{m}'$ *is generated according to the RGMJMCMC algorithm. Accept* $\mathfrak{m}'$ *if both*

1. $\mathfrak{m}'$ *is preliminarily accepted with a probability* $\min\{1, \frac{p(\mathfrak{m}'|\mathbf{y})}{p(\mathfrak{m}|\mathbf{y})}\}$,
2. $\mathfrak{m}'$ *is finally accepted with a probability* $\min\{1, \frac{q_r(\mathfrak{m}|\mathcal{S},\mathfrak{m}'_k)}{q_r(\mathfrak{m}'|\mathcal{S}',\mathfrak{m}_k)}\}$.

Then also $\mathfrak{m} \sim p(\cdot|\mathbf{y})$.

Proof. It holds for a_{mh} given by (2) that

$$a_{mh}(\mathfrak{m}, \mathcal{S}', \mathfrak{m}'_k; \mathfrak{m}', \mathcal{S}, \mathfrak{m}_k) = a^1_{mh}(\mathfrak{m}, \mathcal{S}', \mathfrak{m}'_k; \mathfrak{m}', \mathcal{S}, \mathfrak{m}_k) \times a^2_{mh}(\mathfrak{m}, \mathcal{S}', \mathfrak{m}'_k; \mathfrak{m}', \mathcal{S}, \mathfrak{m}_k)$$

where

$$a^1_{mh}(\mathfrak{m}, \mathcal{S}', \mathfrak{m}'_k; \mathfrak{m}', \mathcal{S}, \mathfrak{m}_k) = \frac{p(\mathfrak{m}'|\mathbf{y})}{p(\mathfrak{m}|\mathbf{y})}, \quad a^2_{mh}(\mathfrak{m}, \mathcal{S}', \mathfrak{m}'_k; \mathfrak{m}', \mathcal{S}, \mathfrak{m}_k) = \frac{q_r(\mathfrak{m}|\mathcal{S}, \mathfrak{m}'_k)}{q_r(\mathfrak{m}'|\mathcal{S}', \mathfrak{m}_k)}$$

Since $a^j_{mh}(\mathfrak{m}, \mathcal{S}', \mathfrak{m}'_k; \mathfrak{m}', \mathcal{S}, \mathfrak{m}_k) = [a^j_{mh}(\mathfrak{m}', \mathcal{S}, \mathfrak{m}_k; \mathfrak{m}, \mathcal{S}, \mathfrak{m}'_k)]^{-1}$ for $j = 1, 2$, by the general results of [1] we obtain an invariant kernel for the target. □

3 Applications

We repeat the experiments from [6] concerned with recovering the planetary mass law (I), the 3rd Kepler's law (II), and a logic regression example (III). We follow the original experimental design and refer the reader to [6] for full detail. The parallelization strategy is described in [5, 6]. The number of threads is denoted by T in the following tables. Table 1 reports results for datasets (I) and (II). Power, FDR, and the expected number of FP are estimated based on 100 runs of RGMJMCMC and GMJMCMC for each number of threads, respectively. The set of nonlinearities $\mathcal{G}_2$ from [6] is used in these experiments. Similarly, Table 2 gives results for the logic regression example (III), where the set of nonlinearities $\mathcal{G}_1$ from [6] is used for BGNLM.

Table 1: Overall Power, average number of false positives (FP) and FDR for detecting the planetary mass law (planetary mass as response) and 3rd Kepler's law (semi-major axis as response). For the former, only $R_p^3 \times \rho_p$ is considered as a TP discovery. For the latter, $F_1 = (P \times P \times M_h)^{\frac{1}{3}}$, $F_2 = (P \times P \times R_h)^{\frac{1}{3}}$, and $F_3 = (P \times P \times T_h)^{\frac{1}{3}}$ are counted as TPs. Results for GMJMCMC are given in parentheses.

Planetary mass law				3rd Kepler's law			
T	Power	FP	FDR	T	Power	FP	FDR
16	0.94 (0.93)	0.29 (0.36)	0.18 (0.22)	64	1.00 (0.99)	0.04 (0.04)	0.02 (0.02)
4	0.63 (0.69)	0.64 (0.49)	0.38 (0.34)	16	0.65 (0.83)	0.88 (0.55)	0.39 (0.22)
1	0.29 (0.42)	1.54 (1.25)	0.71 (0.58)	1	0.06 (0.14)	2.14 (1.81)	0.94 (0.86)

Table 2: Power for individual trees, overall power, expected number of FP, and FDR are compared between RGMJMCMC (R), GMJMCMC (G) and Bayesian Logic regression (L). Logic expressions from the data generating model are $L_1 = X_7$, $L_2 = X_8$, $L_3 = X_2 * X_9$, $L_4 = X_{18} * X_{21}$, $L_5 = X_1 * X_3 * X_{27}$, $L_6 = X_{12} * X_{20} * X_{37}$, $L_7 = X_4 * X_{10} * X_{17} * X_{30}$, $L_8 = X_{11} * X_{13} * X_{19} * X_{50}$.

	T	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	Power	FP	FDR
R	32	1.00	1.00	0.96	1.00	1.00	1.00	0.92	0.89	0.97	1.14	0.13
G	32	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	1.00	0.51	0.06
L	32	0.99	1.00	1.00	0.96	1.00	0.99	0.91	0.38	0.90	1.09	0.13

Tables 1 and 2 illustrate that in all three applications the RGMJMCMC algorithm is capable of estimating the posterior marginal probabilities of different features, and thus is able to recover the true data generative processes with reasonable Power and FDR. The performance is on par with GMJMCMC for datasets (I) and (II). For dataset (III), it is a bit worse than GMJMCMC for BGNLM but slightly better than BLR (see logic regression example in [6]).

4 Discussion

In this paper, we have introduced RGMJMCMC and proved its theoretical properties. We have also repeated some experiments described in [6] to assess RGMJMCMC in terms of model identification. In an extended publication, it would be of interest to additionally evaluate the accuracy of posterior estimates and the performance in terms of out-of-sample prediction accuracy.

Bibliography

[1] M. Banterle, C. Grazian, A. Lee, and C. Robert. Accelerating Metropolis-Hastings algorithms by Delayed Acceptance. *Found. Data Science*, 2019.

[2] M. Clyde, J. Ghosh, and M. Littman. Bayesian adaptive sampling for variable selection and model averaging. *J. Comp. Graph. Stat.*, 2011.

[3] A. Hubin. *Bayesian model configuration, selection and averaging in complex regression contexts*. PhD thesis, University of Oslo, 2018.
[4] A. Hubin and G. Storvik. Mode jumping MCMC for Bayesian variable selection in GLMM. *Comp. Stat. Data Anal.*, 127:281–297, 2018.
[5] A. Hubin, G. Storvik, and F. Frommlet. A Novel Algorithmic Approach to Bayesian Logic Regression (with Discussion). *Bayes. Anal.,* 15(1):263–333, 2020.
[6] A. Hubin, G. Storvik, and F. Frommlet. Flexible Bayesian Nonlinear Model Configuration. To appear in the *J. Artif. Intell. Res.*, 2021.

A lasso-type estimation for the Lorenz regression

Alexandre Jacquemain,[1*] **Cédric Heuchenne**[2] **and Eugen Pircalabelu**[1]

[1] *Université catholique de Louvain, ISBA, B-1348 Louvain-la-Neuve, Belgium*
[2] *University of Liège, Rue Louvrex 14, B-4000 Liège, Belgium*

Abstract: The Lorenz regression procedure introduced by [4] aims to estimate the explained Gini coefficient, a quantity with a natural application in the field of inequality of opportunity. In this paper, we introduce a lasso-type estimator for the explained Gini coefficient and discuss the selection of the regularization parameter. The performance of the procedure is compared to an oracle estimator on simulated data. Finally, an illustration on real-data is provided.

Keywords: Lorenz curve, Inequality of opportunity, Single-index models, LASSO, FABS algorithm

AMS subject classification: 62P20

1 Introduction

The concept of inequality of opportunity (IOP) has been defined by [6] in order to quantify unjust socioeconomic inequalities. More precisely, IOP measures the extent of inequality in an economic advantage which can be attributed to circumstances, i.e. variables over which individuals have no control. In this literature, one difficulty arises from balancing the normative measurement of inequality with the statistical modelling of the advantage. Indeed, the first aspect tends to call for simple linear or log-linear regression models.

Before introducing the Lorenz regression procedure, we give the definition of the Gini coefficient. Denote the expected value by $E[\cdot]$ and let Y be a continuous random variable such that $0 < E[Y] < \infty$. The Gini coefficient is an index ranging from 0 (perfect equality) to 1 (perfect inequality), formally defined as

$$\mathrm{Gi}_Y := \frac{2C[Y, F_Y(Y)]}{E[Y]},$$

where F_Y is the CDF of Y and $C[\cdot, \cdot]$ is the covariance operator.

Let $(Y, X) \in \mathbb{R} \times \mathbb{R}^p$ where Y is a response such that $0 < E[Y] < \infty$ and X is a vector of covariates. In our example, Y is the economic advantage and X gathers the circumstances. We start from the following single-index model

$$E[Y|X = x] = H(x^\intercal \theta_0), \tag{1}$$

*Corresponding author: alexandre.jacquemain@uclouvain.be

where H is an increasing link function and θ_0 is a vector of unknown parameters, with parameter space Θ_0 constrained in order to ensure identifiability. Introduced by [4], the Lorenz regression estimates the *explained* Gini coefficient, defined as

$$\mathrm{Gi}_{Y,X} := \max_{\theta} \frac{2C[Y, F_\theta(X^\intercal\theta)]}{E[Y]} = \frac{2C[Y, F_{\theta_0}(X^\intercal\theta_0)]}{E[Y]} = \mathrm{Gi}_{H(X^\intercal\theta_0)},$$

where the last two equalities arise by using (1) and Theorem 3.1 from [3]. $E[Y|X = x]$ predicts the advantage of an individual with circumstances x. The extent of inequality in the distribution of $E[Y|X]$ thus provides information on IOP. The more inequality, the more the circumstances bear an impact on the expected income and hence, the greater IOP is. Since $\mathrm{Gi}_{Y,X}$ corresponds to the Gini coefficient of $E[Y|X]$ assuming the single-index model, we judge it a natural measure of IOP. Furthermore, the relationship between Y and X is semiparametric and, hence, more flexible than usual parametric methods. Given an iid sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ of size n with the same distribution as (Y, X), the vector θ_0 and $\mathrm{Gi}_{Y,X}$ are consistently estimated with

$$\overline{\theta} := \arg\max_{\theta} \frac{1}{n(n-1)} \sum_{i \neq j} Y_i \mathbf{1}\{X_i^\intercal\theta > X_j^\intercal\theta\}, \tag{2}$$

$$\overline{\mathrm{Gi}}_{Y,X} := \frac{2}{n^2} \sum_{i \neq j} \frac{Y_i}{\overline{Y}} \mathbf{1}\{X_i^\intercal\overline{\theta} > X_j^\intercal\overline{\theta}\} - \frac{n-1}{n}, \tag{3}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function and $\overline{Y}$ is the sample average. The estimator $\overline{\theta}$ is a special case of the monotone rank estimator proposed by [2], which derived its asymptotic properties.

Similarly to the R^2 in linear regression, it is easy to see that $\overline{\mathrm{Gi}}_{Y,X}$ will never decrease as we introduce new covariates. In a sparse setup, where some of the covariates have no influence on the response, an estimation on the full model will lead to an overfit of the explained Gini coefficient. In this paper, we tackle this issue by using an L_1 penalized procedure. We argue that it suits perfectly the setup explained above. First, it makes it possible for an economist to include many covariates without inducing an overestimation of the explained Gini coefficient. Second, it also provides a selection method for the relevant features.

The rest of this paper is structured as follows. In Section 2, we introduce a lasso-type estimator for the Lorenz regression and discuss the choice of the regularization parameter. Section 3 compares the performance of the method with an oracle estimator. Finally, a real-data example is presented in Section 4.

2 The penalized Lorenz regression

In this section, we introduce the penalized Lorenz regression. We borrow the idea developped by [5] for the maximum rank correlation estimator. It consists in replacing the original discrete objective function by a differentiable approximation and adding a penalty part to it. This procedure has the double advantage of introducing penalization and facilitating the numerical computation.

A lasso-type estimator for θ_0 is given by

$$\hat{\theta}(\lambda) := \arg\max_{\theta} \left\{ G(\theta) - \lambda \sum_{k=1}^{p} |\theta_k| \right\},$$

where $\lambda > 0$ is a regularization parameter and $G(\theta)$ is a smooth approximation of the indicator function displayed in (2). Formally,

$$G(\theta) := \frac{1}{n(n-1)} \sum_{i \neq j} Y_i S_\sigma(X_i^\intercal \theta - X_j^\intercal \theta),$$

where $S_\sigma(t) = 1/(1 + \exp(-t/\sigma))$, and σ is a tuning parameter determining the accuracy of the approximation. Following [7], we allow σ to depend on n and set it to $\sigma = 1/\sqrt{n}$. An estimator for the explained Gini coefficient is then obtained by plugging $\hat{\theta}(\lambda)$ in (3).

The numerical implementation can be carried out using the FABS algorithm developped by [7] for nonconvex loss functions and the adaptive Lasso penalty. The method starts with a very large value for λ, imposing full sparsity, and then allows the penalty parameter to decrease. Hence, it produces a whole solution path ranging from high to low sparsity. For each λ, the optimization problem is solved using a coordinate-descent algorithm.

In practice, an optimal regularization parameter can be determined via cross-validation. Another possibility lies in the use of an information criterion, as proposed by [5]. In this respect, we set λ_n as the one which maximizes

$$\mathrm{IC}_\lambda := \log(G(\hat{\theta}(\lambda))) - p_{\mathrm{IC}}(n) k_\lambda,$$

where k_λ is the number of covariates selected using λ and $p_{IC}(n)$ is the penalty associated to the information criterion. For a BIC-like criterion, $p_{IC}(n) = \log(n)/(2n)$. For an AIC-like criterion, $p_{IC}(n) = 1/n$. We evaluate the performance of these alternative methods in Section 3.

3 Monte-Carlo Simulations

In this section, we evaluate the performance of the procedure presented in Section 2 by means of Monte-Carlo simulations. We use the following data generating process (DGP)

$$Y_i = H(X_i^\intercal \theta)\epsilon_i,$$

where $i = 1 \ldots, n = 100$ and $\sum_k |\theta_k| = 1$. X is a multivariate normal with mean 0, unit variance and a correlation matrix following an AR(1) process with correlation parameter $\rho = 0.3$. The variable ϵ_i is a lognormal noise with mean 1 and a variance set to ensure a signal-to-noise ratio of 3. Finally, we use the following link function

$$H(t) = 1000 \exp\left(1 + \frac{1}{2}\left(\frac{t}{3} - 1\right)^3\right).$$

We consider a low-dimensional scenario with $p = 20$ and a high-dimensional one with $p = 120$. In both cases, 5 covariates are active. We simulate from the DGP 2000 different datasets and focus on selecting λ by BIC and AIC-like scores, CV and benchmark the performance of the penalized procedure against an oracle estimator that has knowledge about which covariates are active.

The performance of the procedure is assessed first by the square-root of the MSE of the explained Gini coefficient (RMSE.Gini). Moreover, we judge the quality of the model selection by the true positive rate (TPR), giving the percentage of selected covariates which are truly active, and by the false positive rate (FPR), which represents the percentage of selected variables which are not active. The results are displayed in Table 1. In terms of model selection, we observe a clear tradeoff between TPR and FPR. Cross-validation yields the best performance in terms of TPR but the worst in terms of FPR. The converse story holds for the BIC. Finally, the AIC selection seems to offer the best balance between these two aspects. Concerning the estimation of the explained Gini, the AIC yields once again the best performance, and even slightly outperforms the oracle.

Table 1: Comparison with an oracle estimator

	$p = 20$			$p = 120$		
	RMSE.Gini	TPR	FPR	RMSE.Gini	TPR	FPR
BIC	2.25	82.4	2.0	2.27	75.3	0.6
AIC	2.23	90.6	9.0	2.22	82.5	2.2
CV	2.28	92.5	23.8	2.37	83.6	5.0
Oracle	2.25	/	/	2.26	/	/

Table 2: Estimation of θ_0 in the penalized and unpenalized cases

	Unpenalized	Penalized
Years of education	.027	.028
Sex (reference is male)	-.100	-.101
Married (reference is not married)	.051	.046
Ethnicity (reference is Caucasian)		
Hispanic	-.066	-.065
Black	-.057	-.055
Region (reference is South)	-.044	-.043
Union (reference is nonunionized)	.146	.146
Occupation (reference is tradesperson or assembly line worker)		
Technical or professional	.139	.142
Services	-.028	-.029
Office and clerical	.041	.041
Sales	-.001	
Management and administration	.195	.203
Sector (reference is other)		
manufacturing and mining	.053	.053
construction	.051	.048

4 Real data example

This illustration is based on a random sample of wages data from the 1985 Current Population Survey constructed in [1]. Besides hourly wages, the data provide information on 14 covariates for 534 individuals. A description of the covariates is given in Table 2. The Gini coefficient of wages is of 29.5%. Table 2 displays the estimated vector of coefficients for the penalized and unpenalized Lorenz regression. Taking into account the results of the last section, the regularization parameter is chosen by AIC. As we can observe, only one covariate is not included in the model selection. The estimated explained Gini coefficient is of 18.02% in the unpenalized case and of 17.99% in the penalized regression.

Bibliography

[1] E. R. Berndt. *The Practice of Econometrics: Classical and Contemporary.* Addison-Wesley, Boston, 1991.

[2] C. Cavanagh and R. P. Sherman. Rank estimators for monotonic index models. *Journal of Econometrics*, 84(2):351-381, 1998.

[3] S. Das Gupta. Gini association and pseudo Lorenz curve. *Commun. Stat. Theory Methods*, 28(9):2181-2199, 1999.

[4] C. Heuchenne and A. Jacquemain. Inference for monotone single-index conditional means: a Lorenz regression approach. (Submitted).

[5] H. Lin and H. Peng. Smoothed rank correlation of the linear transformation regression model. *Comput Stat Data Anal*, 57:615-630, 2013.

[6] J. E. Roemer. *Equality of Opportunity.* Harvard University Press, Cambridge, 1998.

[7] X. Shi, Y. Huang, J. Huang and S. Ma. A Forward an Backward Stagewise Algorithm for Nonconvex Loss Functions with Adaptive Lasso. *Comput Stat Data Anal*, 124:235-251, 2018.

On some aspects of discrete-time semi-Markov switching models

Emmanouil-Nektarios Kalligeris,[1*] Alex Karagrigoriou[1], Andreas Makrides[1], Christina Parpoula[2] and Vlad Stefan Barbu[3]

[1]*Lab of Statistics and Data Analysis, Department of Stati stics and Actuarial-Financial Mathematics, University of the Aegean, Karlovasi, Samos 83200, Greece*

[2] *Department of Psychology, Panteion University of Social and Political Sciences, 17671 Athens, Greece*

[3]*Université de Rouen, Laboratoire de Mathématiques Raphaël Salem, UMR 6085, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France*

Abstract: Discrete time semi-Markov switching models constitute a useful modeling technique in various scientific areas. In this work, we establish the basic aspects of such models and present the semi-Markov switching model of conditional mean with covariates along with parameter inference.

Keywords: Hidden semi-Markov chains, Regime switching, Sojourn times,Time-series.

AMS subject classification: 62M10, 62F10, 60K15 & 60J05.

1 Introduction

Markov switching models (MSM) [11], are capable of characterizing non-linear behaviors in different regimes by permitting switching between multiple structures. The switching is memoryless, due to the Markov property, and the waiting time till the next switch follows, in the discrete case, typically a geometric distribution. As a result, the regime switching can be modeled by a finite-state Markov chain. Nowadays, MSM constitute a powerful modeling tool that is being used in several scientific fields such as Financial [6], Epidemiology [3], etc. The concept of MSM though, comes with some serious drawbacks since (a) the memoryless property that governs Markov chains is often inadequate for real-life problems [12], and (b) there are cases where the sojourn time and the associated distribution at a certain state plays a crucial role for modeling data such as those of time-series [4].

A solution to the aforementioned problems, could be provided by semi-Markov switching models (semi-MSM). As stated by Hunt and Devolder [8], Markov chains constitute a subclass of semi-Markov chains [9] and consequently semi-MSM should

*Corresponding author: ekalligeris@aegean.gr

perform at least as good, if not better, as the Markov switching models. One could argue, that this feature alone constitutes a powerful and motivating factor to consider semi-MSM instead of MSM.

In this work, we discuss the concept of discrete time semi-MSM and we establish the basic aspects of such models. The rest of the paper is organized as follows. In Section 2, the discrete time semi-MSM of conditional mean with covariates is presented. In Section 3, we make parameter inference and discuss the problem of noncensored likelihood. Finally, we conclude by summarizing the work done.

2 Discrete Time semi-Markov Switching Model

Let us assume a random system that has a finite state space $E = \{1, ..., N\}$, $N < \infty$, for which the time evolution is governed by a stochastic process $Z = (Z_t)_{t\in\mathbb{N}}$. In addition, let us denote by $S = (S_k)_{k\in\mathbb{N}}$ the successive time-points when state changes in $(Z_t)_{t\in\mathbb{N}}$ and by $J = (J_k)_{k\in\mathbb{N}}$ the associated visited states at these time-points.

Definition 2.1 **(Markov renewal & semi-Markov chain)**

If $(J, S) = (J_k, S_k)_{k\in\mathbb{N}}$ satisfies the relation

$$\begin{aligned} &P(J_{k+1} = j, S_{k+1} - S_k = t|J_0, J_1, ..., J_k; S_1, ..., S_k) \\ &= P(J_{k+1} = j, S_{k+1} - S_k = t|J_k),\ j \in E,\ t \in \mathbb{N}, \end{aligned} \tag{1}$$

then $Z = (Z_t)_{t\in\mathbb{N}}$ is a semi-Markov chain associated to the Markov renewal chain $(J, S) = (J_k, S_k)_{k\in\mathbb{N}}$, where

$$Z_t = J_{N(t)} \Leftrightarrow J_k = Z_{S_k},$$

with $N(t) = max\{k \in \mathbb{N}|S_{k+1} \leq t\}, t \in \mathbb{N}$, being the counting process of the number of jumps in the time interval $(0, t]$. As a result, Z_t represents the state of the system at time t [10]. The fact that $(J_k, S_k)_{k\in\mathbb{N}}$ is a Markov renewal chain, implies that $(J_k)_{k\in\mathbb{N}}$ is an embedded Markov chain. Note that throughout this paper, we consider (J, S) to be homogeneous that is equation (1) is independent of k.

In order to provide some basic definitions, we introduce now the proper notation. Consider $l, k \in \mathbb{N}$, $l \leq k$, two nonnegative integers and let $y_l, ..., y_k \in A = \{1, ..., s\}, s < \infty$. We will denote by Y_l^k the vector $Y_l^k = (Y_l, ..., Y_k)$ and we will write $\{Y_l^k = y_l^k\}$ for the event $\{Y_l = y_l, ..., Y_k = y_k\}$. In the case of a single state within the state space, i.e., $y_l, ..., y_k \equiv y \in A$, we denote by $\{Y_l^k = y\}$ the event $\{Y_l = y, ..., Y_k = y\}$. Finally, the notation $\{Y_l^k = \cdot\}$ refers to the event $\{Y_l = \cdot, ..., Y_k = \cdot\}$.

It is obvious that all the above notations in terms of the chain Y can be easily expressed in terms of the chain Z.

Definition 2.2 **(Hidden semi-Markov chain of order k)**

Let $Y = (Y_t)_{t\in\mathbb{N}}$ be a homogeneous Markov chain of order $k,\ k \geq 1$, conditioned on the semi-Markov chain Z which means that $\forall y_0, ..., y_k \in A,\ i \in E, t \in \mathbb{N}^*$:

$$\begin{aligned} P(Y_{t+1} = y_k | Y_{t-k+1}^t = y_0^{k-1}, Y_0^{t-k} = \cdot, Z_{t+1} = i, Z_0^t = \cdot) \\ = P(Y_{t+1} = y_k | Y_{t-k+1}^t = y_0^{k-1}, Z_{t+1} = i). \end{aligned} \tag{2}$$

The chain $(Z, Y) = (Z_t, Y_t)_{t\in\mathbb{N}}$ is called a hidden semi-Markov chain of order k and the probability in (2) is known as the *emission probability matrix* of the conditional Markov chain Y.

If in (2) the observation process is characterized by the conditional independence property, then $\forall y \in A,\ i \in E, t \in \mathbb{N}^*$:

$$\begin{aligned} P(Y_t = y | Y_0^{t-1} = \cdot, Z_t = i, Z_0^{t-1} = \cdot) \\ = P(Y_t = y | Z_t = i), \end{aligned}$$

where $\sum_{y_k} P(Y_{t+1} = y_k | Z_{t+1} = i) = 1$.

For more information on the topic of hidden semi-Markov chains the interested reader may refer to [2].

Based on the Definition 2.2, we now define the semi-Markov switching model of conditional mean with covariates.

Let us suppose a series of observations $\left\{y_0^{T-1}\right\}$ and $\left\{z_0^{T-1}\right\}$ a hidden state variable which follows a first order semi-Markov chain which is characterized by the following semi-Markov kernel $\boldsymbol{q}$:

$$q_{ij}(t) = P(J_k = j, S_{k+1} - S_k = t | J_{k-1} = i).$$

Definition 2.3 **(Discrete Time semi-Markov Switching Model of Conditional Mean with Covariates)**

A semi-Markov switching model of conditional mean with covariates $\Omega_1, ..., \Omega_q$ for $y_t,\ t \in \mathbb{N}^*$, is defined by:

$$y_t = c_{z_t} + \sum_{i=1}^{p} \phi_{iz_t} y_{t-i} + \sum_{d=1}^{q} \gamma_{dz_t} \Omega_d + \epsilon_t,\ t = 0, 1, ..., T-1, \tag{3}$$

where c_{z_t} is a switching intercept, ϕ_{iz_t}, $i = 1, ..., p$, are autoregressive (AR) switching coefficients, γ_{dz_t} the coefficient associated with the Ω_d covariate, $d = 1, ..., q$, and ϵ_t are *i.i.d* zero-mean normally distributed random variables with variance $\sigma_{z_t}^2$.

Under the model in (3) and for a N-state setting, one could consider various underlying (discrete) distributions for the waiting (sojourn) times between states.

3 Parameter Inference

Consider a series of observations $\left\{y_0^{T-1}\right\}$ and $\left\{z_0^{T-1}\right\}$ a hidden state variable as in Section 2. Moreover, suppose that the number of sojourn (waiting) times, denoted by $v_0, v_1, ..., v_R$, fulfills the obvious equality:

$$v_0 + v_1 + ... + v_R = T.$$

The relationship between the sojourn times and the state sequence can be simplified by reducing the entire sequence of states $z_0, z_1, ..., z_{T-1}$ to the sequence of states $j_0, j_1, ..., j_R$ which have been visited:

$$\begin{aligned} j_0 &:= \{z_0, z_1, ..., z_{v_0-1}\} \\ j_1 &:= \{z_{v_0}, z_{v_0+1}, ..., z_{v_0+v_1-1}\} \\ &\vdots \\ j_R &:= \{z_{v_0+v_1...+v_{R-1}}, z_{v_0+v_1...+v_{R-1}+1}, ..., z_{T-1}\}. \end{aligned}$$

Ferguson in [5] introduced the classical form of the complete (noncensored) data likelihood which allows only for sequences in which the last observation coincides with an exit from the hidden state. This form though, comes with some limitations since the summation includes all the possible paths considered in the complete-data likelihood and as a result the probability of obtaining an analytical solution is negligible. Furthermore, it assumes that the exit from a state coincides with the end of the sequence of observations Y_0^{T-1} since the sojourn times $v_r, r = 0, ..., R$ sum up to T. This results in the forbiddance of the consideration of semi-Markov chains with absorbing states which is unrealistic for most applications. Considering the aforementioned, Guédon in [7] proposed the implementation of the survivor function into (3):

$$\begin{aligned} &L'_{complete}\left(z_0^{T-1}, Y_0^{T-1}|\theta\right) \\ &= \sum_{t=0}^{T-1} log \sum_{r=1}^{R-1} f(y_t|Z_t, Y_0^t; \theta) P_{j_0} w_{j_0}(v_0) P_{j_r|j_{r-1}} w_{j_{r-1}}(v_{r-1}) P_{j_R|j_{R-1}} W_{j_R}(v_R), \end{aligned} \tag{4}$$

where

$$W_{j_r}(v_r) = \sum_{u_r \geq v_r} w_{j_r}(u_r),$$

is the survivor function for the sojourn time in state j_r and $\theta \in \mathbb{R}^m, m \in \mathbb{N}$, is the a parameter vector.

The estimator resulting through $L'_{complete}\left(z_0^{T-1}, Y_0^{T-1}|\theta\right)$ is known as *partial likelihood estimator*. Estimating the likelihood of semi-Markov switching model constitutes an incomplete (censored) data problem since the only accessible quantity is the observations. This fact makes the Expectation-Maximization (EM) algorithm the most suitable ML estimation technique for such models. For more on the estimation of censored semi-Markov switching models the interested reader may refer to [2, 7].

4 Conclusion

In this work, we discussed the concept of semi-Markov switching models under the discrete time framework. The fundamental aspects of such models were presented. To that end, the proper notations as well as the formulation of discrete time semi-Markov switching models of conditional mean with covariates were provided together with the associated parameter inference.

Acknowledgements: The authors wish to express their appreciation to the Editor and anonymous Referee for their comments, suggestions, and recommendations which helped in improving both the quality and the presentation of the manuscript. Finally, note that this work was carried out at the Lab of Statistics and Data Analysis of the University of the Aegean.

Bibliography

[1] V.S. Barbu, A. Karagrigoriou and A. Makrides. *Semi-Markov modelling for multi-state systems.* Methodol. Comput. Appl. Probab., 19:1011–1028, 2017.

[2] V.S. Barbu and N. Limnios. *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications: Their Use in Reliability and DNA Analysis.* Lecture Notes in Statistics, Springer-Verlag New York, 2008.

[3] M. Dahlquist and S.F. Gray. *Regime Switching and Interest rates in the European Monetary System.* J. Int. Econ., 50:399–419, 2000.

[4] F. Diebold and G. Rudebusch. *A Nonparametric Investigation of Duration Dependence in the American Business Cycle.* J. Polit. Econ., 98:596–616, 1990.

[5] J.D. Ferguson. *Variable Duration Models for Speech.* Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech, Princeton, New Jersey, 143–179, 1980.

[6] S.M. Goldfeld and R.E. Quandt. *A Markov Model for Switching Regressions.* J. Econom., 1(1):3–15, 1973.

[7] Y. Guédon. *Estimating Hidden semi-Markov Chains from Discrete Sequences.* J. Comput. Graph. Stat., 12(3):604–639, 2003.

[8] J. Hunt and P. Devolder. *Semi-Markov Regime Switching Interest Rate Models and Minimal Entropy Measure.* Physica A, 390(21-22):3767–3781, 2011.

[9] J.S. Hunt. *A Short Note on Continuous-Time Markov and Semi-Markov Processes.* Discussion Paper, DP-1003, ISBA, UCL, 2010.

[10] N. Limnios and G. Oprişan. *Semi-Markov Processes and Reliability.* Birkhäuser, Boston, 2001.

[11] G. Lindgren. *Markov Regime Models for Mixed Distributions and Switching Regressions.* Scand. J. Stat., 5(2):81–91, 1978.

[12] D. Silvestrov and F. Stenberg. *A Pricing Process with Stochastic Volatility Controlled by a Semi-Markov Process.* Commun. Stat. - Theory Methods, 33(3):591–608, 2004.

A note on parameter estimation of thinned random intersection graphs

Joona Karjalainen[1]*

[1]*Department of Mathematics and Systems Analysis, Aalto University*

Abstract: We study a network model with overlapping communities on n nodes parameterized by the number of communities, the community size distribution, and the within-community edge probability. Moment-based parameter estimators were given in [6] for the model with binomial community sizes. We extend this approach for more general distributions, and give sufficient conditions for their consistency as n tends to infinity.

Keywords: complex networks, random graphs, parameter estimation
AMS subject classification: 05C80, 05C07, 62F12, 91D30

1 Introduction

Many types of data in different fields of science can be naturally represented and modeled as undirected graphs. In real-world networks it is common to observe community structure, i.e., some groups of nodes are more densely connected to each other than to the rest of the nodes. This can be naturally explained in the context of social networks, e.g., with common hobbies, location, or occupation. Since communities can be formed by many different types of mechanisms, it is often desirable that a network model allows the communities to overlap. A number of models with this property have been presented in the literature – we only mention the *active* and *passive random intersection graphs* [4], which have some features of real-world networks, such as non-trivial clustering, power-law degree distributions, and degree-clustering correlations [1]. These models have also motivated the model studied in this work.

We now define the thinned random intersection graph. Let n be the number of nodes, m the number of communities, π the community size distribution, and q the edge probability within the communities. Each community C_k, $k = 1 \dots m$, is generated independently of others as follows:

1. Generate the number of nodes, $|V(C_k)|$ from the distribution π.
2. Choose the node set $V(C_k)$ uniformly at random from the subsets of $[n] = \{1, \dots, n\}$ of size $|V(C_k)|$.

*Corresponding author: joona.karjalainen@aalto.fi

3. Generate the edges between the node pairs of $V(C_k)$ independently with probability q.

The resulting graph G is defined as the superposition of the communities:

$$V(G) = [n], \qquad E(G) = \bigcup_{k=1}^{m} E(C_k).$$

Setting $q = 1$ and letting π be arbitrary gives the passive random intersection graph. The motivation for our model is that the assumption $q = 1$ would imply that each community forms a clique. The parameter q may be viewed as a relaxation constant, or as a measure of how closely the communities resemble cliques.

When $\pi = \mathrm{Bin}(n, p)$ (and $q \in (0, 1]$), the indicators $\mathbf{I}(i \in V(C_k))$ are mutually independent, $i = 1 \dots n$, $k = 1 \dots m$. It was noted in [6] that in this case, we have that asymptotically

$$q \;\approx\; \tau(1 + \frac{\lambda^2}{\sigma^2 - \lambda}),$$

where τ is the clustering coefficient $3\frac{\#\text{triangles}}{\#\text{2-stars}}$, λ is the average degree, and σ^2 is the degree variance. This observation directly suggests a moment-based estimator. Namely, since the λ, σ^2 and τ can be easily computed from data (i.e., the graph G), the above expression can be evaluated, and the resulting estimator $\hat{q}$ is also consistent if these empirical quantities are sufficiently well-behaved. In this paper we show that the parameter q can be consistently estimated with more general, non-binomial community size distributions. The second result of this paper gives a parameter estimator for π.

A generalization of the model has been studied in [2, 3], where q is allowed to be random and depend on the size of the community. We only note here that under certain assumptions, the asymptotic degree distribution is known to be a compound Poisson distribution, and that an explicit formula for the limiting assortativity coefficient is known.

2 Large-scale assumptions and notation

A large network is modeled by letting $n \to \infty$, which results in a sequence of independent graphs $(G_1, G_2, \dots)$, indexed by the number of nodes. The number of communities is assumed to grow linearly with the number of nodes, $m/n \to \mu$ for some number $\mu \in (0, \infty)$, which results in non-trivial average degrees and clustering coefficients. The dependence of m on n is omitted in the notation. Denoting the communities by $C_{n,k}$, the graph G_n is defined by $V(G_n) = [n]$, $E(G_n) = \bigcup_{k=1}^{m} E(C_{n,k})$.

The community size distribution is allowed to vary with n, and is assumed to converge weakly to a limit, $\pi_n \xrightarrow{w} \pi$. For simplicity, we assume that q does not depend on n, although we expect the results to remain largely unchanged if we assume convergence to a constant in $(0, 1]$.

We denote $(n)_k = n(n-1)\dots(n-k+1)$ for a constant n. For a distribution π and a random variable $X \sim \pi$ we denote $(\pi)_k = \mathbb{E}\left[X(X-1)\dots(X-k+1)\right]$. The notation $X = o_{\mathbb{P}}(1)$ is used to mean "$X \to 0$ in probability as $n \to \infty$".

3 Moment-based parameter estimators

We present estimators for the thinning parameter q and the community size distribution π. The estimators can be evaluated based on the numbers of edges, 2-stars and triangles.

Denote by $X_{i,j}$ the edge indicator $\mathbf{I}(\{i,j\} \in E(G_n))$, and by $d_i = \sum_{j=1}^{n} X_{i,j}$ the degree of node i in G_n. The subgraph counts for the edges, 2-stars, and triangles are given by

$$N_{K_2} = \sum_{1 \le i < j \le n} X_{i,j}, \quad N_{S_2} = \sum_{i=1}^{n} \sum_{j<k} X_{i,j} X_{i,k}, \quad N_{K_3} = \sum_{i<j<k} X_{i,j} X_{i,k} X_{j,k}.$$

Natural estimators of the mean degree, degree variance, and clustering coefficient are given by

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} d_i = \frac{2}{n} N_{K_2}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (d_i - \hat{\lambda})^2, \quad \hat{\tau} = 3 \frac{N_{K_3}}{N_{S_2}}.$$

Theorem 1. *Let $m \to \infty$, $n \to \infty$, and $m/n \to \mu \in (0, \infty)$. Assume that π_n converges weakly to a distribution π on $\mathbb{N}$ with $(\pi_n)_3 \to (\pi)_3 \in (0, \infty)$. Then*

$$\hat{q} := \hat{\tau}\Big(1 + \frac{\hat{\lambda}^2}{\hat{\sigma}^2 - \hat{\lambda}}\Big) = q + o_{\mathbb{P}}(1).$$

In the following we consider a family of single-parameter distributions. Let $P \subset \mathbb{R}$, and let $\mathcal{D} = \{\nu^p\}_{p \in P}$ be a family of distributions on $\mathbb{N}$ indexed by the parameter p. We require an identifiability condition: there must exist a function f such that

$$f\left(\frac{(\nu^p)_3}{(\nu^p)_2}\right) = p, \quad \forall p \in P. \tag{1}$$

Example 1. Assume that π is a Pareto-mixed Poisson distribution with shape parameter $\alpha > 3$,

$$\pi(k) = \int_1^{\infty} \alpha x^{-(\alpha+1)} \frac{x^k e^{-x}}{k!} dx, \quad k \in \{0, 1, 2, \dots\}.$$

With the auxiliary random variables $X \sim \mathrm{Poi}(Y)$ and $Y \sim \mathrm{Par}(\alpha, 1)$ we obtain

$$(\pi)_2 = \mathbb{E}\left[\mathbb{E}\left[X(X-1) \mid Y\right]\right] = \mathbb{E}\left[Y^2\right] = \frac{\alpha}{\alpha - 2}, \quad (\pi)_3 = \mathbb{E}\left[Y^3\right] = \frac{\alpha}{\alpha - 3}.$$

Hence, $\frac{(\pi)_3}{(\pi)_2} = \frac{\alpha-2}{\alpha-3}$, which gives $\alpha = \frac{3\frac{(\pi)_3}{(\pi)_2}-2}{\frac{(\pi)_3}{(\pi)_2}-1}$. Thus, for the family of Pareto-mixed Poisson distributions with shape parameter $\alpha > 3$, the function $f(x) = \mathbf{I}(x \neq 1)\frac{3x-2}{x-1}$ satisfies (1).

Example 2. Consider a Zipf's law of the form

$$\pi(k) = \frac{k^{-s}}{\sum_{n=1}^{\infty} n^{-s}}, \quad k \in \{1, 2, \ldots\},$$

parameterized by $s > 4$. Denoting the Riemann zeta function by $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$, we may then define $f(x)$ as the solution to the equation

$$x = \frac{\zeta(s-3) - 3\zeta(s-2) + 2\zeta(s-1)}{\zeta(s-2) - \zeta(s-1)},$$

and solve it numerically.

The following result gives an estimator for the parameter p of the limiting community size distribution π. For large n, this estimate may then be used to describe the community size distribution in an observed graph.

Theorem 2. *Let $m \to \infty$, $n \to \infty$, and $m/n \to \mu \in (0, \infty)$. Assume that π_n converges weakly to a distribution $\pi \in \mathcal{D}$ with $(\pi_n)_3 \to (\pi)_3 \in (0, \infty)$. Let $f : \mathbb{R} \to \mathbb{R}$ be a function that satisfies (1). If f is continuous at $(\pi)_3/(\pi)_2$, then $f\left(\frac{\hat{\sigma}^2 - \hat{\lambda}}{\hat{\lambda}\hat{q}}\right) = p + o_{\mathbb{P}}(1)$.*

4 Proofs

Proof of Theorem 1. A simple calculation shows that

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(d_i - \hat{\lambda})^2 = \frac{2}{n}N_{K_2} + \frac{2}{n}N_{S_2} - 4\frac{N_{K_2}^2}{n^2}. \tag{2}$$

Hence,

$$\hat{q} = \hat{\tau}\left(1 + \frac{\hat{\lambda}^2}{\hat{\sigma}^2 - \hat{\lambda}}\right) = 3\frac{N_{K_3}}{N_{S_2}}\left(1 + \frac{4N_{K_2}^2}{2nN_{S_2} - 4N_{K_2}^2}\right). \tag{3}$$

Since $\pi_n \to \pi$ weakly and $(\pi_n)_3 \to (\pi)_3 \in (0, \infty)$, it follows from [5] (Theorem 4.2) that

$$N_R = (1 + o_{\mathbb{P}}(1))\mathbb{E}\, N_R, \quad R = K_2,\, S_2, K_3, \tag{4}$$

and by [6] (Propositions 1 and 2)

$$\mathbb{E}\, N_{K_2} = \binom{n}{2} m \frac{(\pi)_2}{(n)_2} q(1 + O(n^{-1})) = \frac{1}{2} m(\pi)_2 q(1 + O(n^{-1})), \tag{5}$$

$$\mathbb{E}\, N_{S_2} = \frac{1}{2} q^2 \left(m(\pi)_3 + \frac{(m)_2}{(n)_2}(n-2)(\pi)_2^2 \right) (1 + O(n^{-1})), \tag{6}$$

$$\mathbb{E}\, N_{K_3} = \frac{1}{6} q^3 m(\pi)_3 (1 + O(n^{-1})). \tag{7}$$

The claim follows (after simple but tedious calculations) by the continuous mapping theorem and (3), (4) and (5)–(7). By using the same equations, one may verify that the denominator $\hat{\sigma}^2 - \hat{\lambda}$ in the expression of $\hat{q}$ converges in probability to a nonzero constant. □

Proof of Theorem 2. Consider the expression $\frac{\hat{\sigma}^2 - \hat{\lambda}}{\hat{\lambda}\hat{q}}$. Inserting the definitions of $\hat{\sigma}^2$ and $\hat{\lambda}$, and recalling that $\hat{q} = q + o_{\mathbb{P}}(1)$ we obtain that

$$\frac{\hat{\sigma}^2 - \hat{\lambda}}{\hat{\lambda}\hat{q}} = \frac{\frac{2}{n} N_{S_2} - 4\frac{N_{K_2}^2}{n^2}}{\frac{2N_{K_2}}{n} q(1 + o_{\mathbb{P}}(1))}.$$

It now follows from (4)–(6), together with the continuous mapping theorem, that $\frac{\hat{\sigma}^2 - \hat{\lambda}}{\hat{\lambda}\hat{q}} = \frac{(\pi)_3}{(\pi)_2} + o_{\mathbb{P}}(1)$. Since by assumption f is continuous at $(\pi)_3/(\pi)_2$ with $f((\pi)_3/(\pi)_2) = p$, the claim follows by applying the continuous mapping theorem again. □

Acknowledgements: The author was supported by the Magnus Ehrnrooth Foundation.

Bibliography

[1] M. Bloznelis. Degree and clustering coefficient in sparse random intersection graphs. Ann. Appl. Probab. 23(3):1254–1289, 2013.

[2] M. Bloznelis, J. Karjalainen and L. Leskelä. Assortativity and bidegree distributions on Bernoulli random graph superpositions. 17th Workshop on Algorithms and Models for the Web Graph (WAW). Lecture Notes in Computer Science 12091, pp. 68–81. Springer, 2020.

[3] M. Bloznelis and L. Leskelä. Clustering and percolation on superpositions of Bernoulli random graphs. arXiv preprint, 2020.

[4] E. Godehardt and J. Jaworski. Two models of random intersection graphs for classification. In Exploratory Data Analysis in Empirical Research 67–81. Springer, 2003.

[5] T. Gröhn, J. Karjalainen and L. Leskelä. Clique and cycle frequencies in a sparse random graph model with overlapping communities. arXiv preprint, 2021.

[6] J. Karjalainen, J.S.H. van Leeuwaarden and L. Leskelä. Parameter estimators of sparse random intersection graphs with thinned communities. 15th Workshop on Algorithms and Models for the Web Graph (WAW). Lecture Notes in Computer Science 10836, pp. 44–58, Springer 2018.

Techniques from functional data analysis adaptable for spatial point patterns

Kateřina Koňasová,[1*] **and Jiří Dvořák**[1]

[1]*Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic*

Abstract: Spatial point pattern is a collection of points observed in a bounded region of d-dimensional Euclidean space, $d \geq 2$. Usually, $d = 2$ or $d = 3$. Individual points then represent e.g. observed locations of cell nuclei in human or animal tissue, nests of a specific bird species etc. Popular tools for point pattern analysis are functional summary characteristics describing different features of the complex point pattern structure via functions of one or more arguments. Thus, we can benefit from the link between a point pattern and the corresponding empirical value of a chosen functional characteristic. This paper presents a brief overview of classical techniques from functional data analysis that can be effortlessly adapted to the context of point pattern data.

Keywords: Point processes, dissimilarity measures, multidimensional scaling, supervised classification, data depth.

AMS subject classification: 60G55, 62H30

1 Introduction

Spatial point processes, studied within the scope of spatial statistics, have been recently given more and more attention in a broad range of scientific disciplines, including biology, statistical physics, or material science [5]. They are used to model locations of objects or events randomly occurring in the d-dimensional Euclidean space, $d \geq 2$. We distinguish between the random model, referred to as point process, and its realization called point pattern.

Functional summary characteristics play a fundamental role in the whole process of investigating point patterns, from exploratory analysis to parameter estimation and hypothesis testing. With the increasing availability of point pattern data, many functional characteristics have been developed to capture various features of the data that can be relevant to the problem at hand.

We focus on the link between a point pattern and the corresponding empirical value of a selected functional characteristic. Thanks to this connection, we can analyze functional data instead of the original patterns. This brings new perspectives to the field of spatial statistics; well-developed methods from functional data analysis can be applied, including supervised classification [3] or ranking techniques connected with functional depth [8].

*Corresponding author: konasova@karlin.mff.cuni.cz

2 Spatial point patterns

This section gives the necessary definitions concerning spatial point processes. We closely follow the book [10]; foundations of the point process theory are explained to a greater extent in [2]. Fix $d \geq 2$.

We define a simple point process X as a random locally finite subset of $\mathbb{R}^d$, where each point $x \in X$ corresponds to a specific object or event occurring at the location $x \in \mathbb{R}^d$. As an illustration with $d = 2$, one can think about modeling random locations of centers of cell nuclei in human tissue. In what follows, we distinguish the random element X (point process) and its realization $\mathcal{X}$ (point pattern).

To describe various features of X, all kinds of functional summary characteristics have been developed. It is far beyond the scope of this paper to list them all; we focus on the pair correlation function g mainly because of its widespread use in practical applications. Before introducing g itself, we define some moment properties of X.

The *intensity function* $\lambda(\cdot)$ is a non-negative measurable function on $\mathbb{R}^d$ such that $\lambda(x)\,\mathrm{d}x$ corresponds to the probability of observing a point of X in a neighborhood of x with infinitesimally small area $\mathrm{d}x$. If X is stationary (distribution invariant w.r.t translations in $\mathbb{R}^d$), then $\lambda(\cdot) = \lambda$ is a constant function and the constant λ is called the *intensity*. In this case, λ is interpreted as the expected number of points of X occurring in a set with unit d-dimensional volume. Similarly, the *second-order product density* $\lambda^{(2)}(\cdot,\cdot)$ is a non-negative measurable function on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\lambda^{(2)}(x,y)\,\mathrm{d}x\,\mathrm{d}y$ corresponds to the probability of observing two points of X jointly occurring at neighborhoods of x and y with infinitesimally small areas $\mathrm{d}x$ resp. $\mathrm{d}y$.

Assuming λ and $\lambda^{(2)}$ exist for the process X, the *pair correlation function* $g(x,y)$ is defined as $\lambda^{(2)}(x,y)/(\lambda(x)\lambda(y))$, whenever $\lambda(x)\lambda(y) > 0$. If $\lambda(x) = 0$ or $\lambda(y) = 0$, we set $g(x,y) = 0$. We write $g(x,y) = g(x-y)$ whenever g is translation invariant. For the Poisson point process, a theoretical model for points randomly placed in space without any interactions among the points, $\lambda^{(2)}(x,y) = \lambda(x)\lambda(y)$ and $g \equiv 1$. Thus, $g(x,y)$ quantifies how likely it is in the process X to observe two points jointly occurring in infinitesimally small neighbourhoods of x and y, compared to the "no interactions" benchmark.

Other popular characteristics are based on g, e.g. its cumulative counterpart traditionally called the K-function. Others are based on inter-point distances, such as the nearest neighbour distance distribution function G and the spherical contact distribution function F [10].

While analyzing point patterns, empirical estimators of the theoretical characteristics are needed. A comprehensive list is given in [5, 10]. Estimators of g, K, G and F are implemented in the `R` package `spatstat` [1].

3 Multidimensional scaling, supervised classification

This section presents a short illustration of two well-established methods in the functional data context that can be easily modified to the point pattern setting. They

have a point in common: a mapping quantifying how dissimilar two point patterns (or functions) are is required.

For real-valued functions f_1, f_2 defined on $B \subset \mathbb{R}$, a commonly used dissimilarity measure is $\delta(f_1, f_2) = \int_B |f_1(x) - f_2(x)|^2 \, \mathrm{d}x$, fulfilling all the properties of a metric except from $\delta(f_1, f_2) = 0 \iff f_1 = f_2$ [3]. If f_1 and f_2 are two cumulative distribution functions, then δ is the Cramér-von Mises test statistic.

Let $\mathcal{X}$ and $\mathcal{Y}$ be two point patterns observed on (not necessarily identical) bounded regions $W_{\mathcal{X}}, W_{\mathcal{Y}} \subset \mathbb{R}^d$. If we select a functional characteristic f and denote by $\widehat{f}(\mathcal{X}, r)$ the value of $f(r)$, $r \in \mathbb{R}$, estimated from $\mathcal{X}$ on $W_{\mathcal{X}}$, we can measure dissimilarity between $\widehat{f}(\mathcal{X}, \cdot)$ and $\widehat{f}(\mathcal{Y}, \cdot)$ instead of comparing $\mathcal{X}$ and $\mathcal{Y}$ directly. The dissimilarity measure based on f is then defined as $\delta[f](\mathcal{X}, \mathcal{Y}) = \delta\left(\widehat{f}(\mathcal{X}, \cdot), \widehat{f}(\mathcal{Y}, \cdot)\right) = \int_{r_{min}}^{r_{max}} |\widehat{f}(\mathcal{X}, r) - \widehat{f}(\mathcal{Y}, r)|^2 \, \mathrm{d}r$, where $r_{min} < r_{max}$ are given constants depending on the size and shape of $W_{\mathcal{X}}$, $W_{\mathcal{Y}}$.

Dissimilarities $\delta[f]$ can be visualized using the multidimensional scaling (MDS). This technique takes elements of a high-dimensional space and visualizes them as points in $\mathbb{R}^m$ ($m \geq 1$ small), in such a way that the Euclidean distances in $\mathbb{R}^m$ are approximately proportional to the original dissimilarities. For a collection of point patterns $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$, let D be a symmetric, hollow matrix composed of non-negative elements $\delta_{i,j} = \delta[f](\mathcal{X}_i, \mathcal{X}_j)$. We want to find an n-tuple of points $\mathbf{z} = (z_1, \ldots, z_n) \in (\mathbb{R}^m)^n$ so that $\mathbf{z}$ minimizes a predefined loss function called stress [4]. To solve this optimization problem, a strategy known as SMACOF (Scaling by Majorizing a Complicated Function [4]) is used. Note that D is not required to be an Euclidean distance matrix (this is required for the classical MDS [12]). In the software `R`, MDS via SMACOF is implemented in the package `smacof` [7].

For illustration, set $d = 2$ and consider a collection Γ of 40 patterns observed on a unit square containing 20 realizations of stationary Poisson point process Π and 20 realization of stationary Thomas process X, the latter exhibiting attractive interactions among pairs of points [10]. Sample realizations can be seen in Fig. 1. For ease of presentation, we fix the intensity of Π and X to be 120, and we consider X to be a model with one parameter σ with small values of σ indicating strong, short-range attractive interactions.

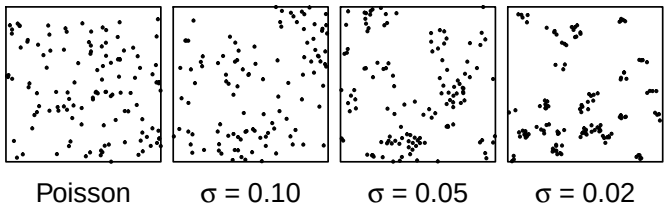


Figure 1: Realization of stationary Poisson point process and stationary Thomas process with parameter $\sigma = 0.02, 0.05, 0.1$.

Fig. 2 (left) shows dissimilarities based on $\delta[g]$ among patterns in Γ visualized in $\mathbb{R}^2$ with the help of MDS. Parameter σ is set to be 0.05. Realizations of Π exhibits low within-group dissimilarities and are well-separated from the group of realizations of X. On the other hand, the presence of attractive interactions among pairs of points causes larger variability of the within-group dissimilarities for the

group of realizations of X.

Fig. 2 (right) shows average misclassification rates (MCRs) corresponding to supervised classification of realizations of Π and X. For each σ between 0.02 and 0.2, dissimilarities among realizations of Π and $X(\sigma)$ are computed, and class membership (Poisson vs Thomas) is predicted based on the kernel regression algorithm for functional data [3]. To train the algorithm, a training set of patterns with known labels is needed. Individual MCRs are computed from another set of labeled patterns, called testing set, by comparing the true and the estimated labels. We perform 100 independent repetitions of the experiment and take the empirical average of the individual MCRs. For $\sigma < 0.1$ (beyond 0.1, realizations of Π and $X(\sigma)$ are indistinguishable on the unit square), satisfactory results are obtain even for a small number of points (120 on average) per pattern. Also, the algorithm's behaviour is in coherence with the MDS visualization; large within-group dissimilarities make the incoming realization of $X(\sigma)$ harder to label correctly (compared to the realization of Π), see [6] for more details.

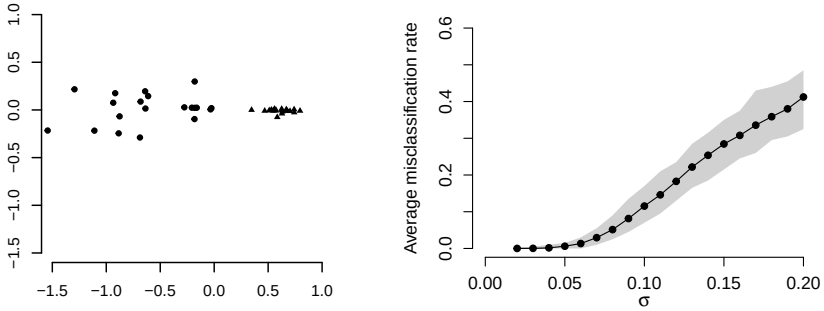


Figure 2: Left: Coordinates of 40 points in $\mathbb{R}^2$ are plotted, each point representing one element from Γ (realizations of Π = triangles, realizations of X = circles). These coordinates are obtained by applying MDS on a matrix of dissimilarities $\delta[g]$ among the elements of Γ. The Euclidean distances among the 40 points are approximately proportional to the dissimilarities $\delta[g]$. Right: Average misclassification rates (together with 90%-pointwise envelope) for binary classification based on $\delta[g]$. Realizations of Π and $X(\sigma)$ are classified using the kernel regression algorithm for functional data [3].

4 Further techniques exploiting the functional representation

When point patterns are represented by functions, dissimilarities between patterns can be defined, as discussed above. This enables performing different exploratory and inferential tasks for replicated point pattern data.

To name a few, apart from supervised classification and visualization using MDS discussed in Sect. 3, one can perform unsupervised classification (clustering) using standard techniques or determine a prototype (median pattern) by finding $\operatorname{argmin}_i \sum_{i \neq j} \delta[f](\mathcal{X}_i, \mathcal{X}_j)$ [9]. Furthermore, a ranking of patterns (represented by

functions) can be defined using functional depth [8]. The ranking provides the possibility of outlier detection, Monte Carlo goodness-of-of fit testing (in spatial statistics now predominantly performed using the approach of [11]) or two-sample tests in the spirit of the Wilcoxon rank-sum test.

Acknowledgements: This work has been supported by The Charles University Grant Agency, project no. 1198120, and The Czech Science Foundation, project no. 19-04412S.

Bibliography

[1] A. Baddeley, E. Rubak and R. Turner. *Spatial Point Patterns: Methodology and Applications with R.* Chapman & Hall/CRC Press, Boca Raton, 2015.

[2] D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes. Vol II., 2nd edn.* Springer-Verlag, New York, 2008.

[3] F. Ferraty and P. Vieu. *Nonparametric functional data analysis. Theory and practice.* Springer-Verlag, New York, 2006.

[4] P. Groenen and M. van de Velden. Multidimensional Scaling by Majorization: A Review. *J. Stat. Softw.*, 73(8);1-26, 2016.

[5] J. Illian, A. Penttinen, H. Stoyan and D. Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns.* Wiley, Chichester, 2004.

[6] K. Koňasová and J. Dvořák. Supervised nonparametric classificationin the context of replicated point patterns. *Submitted*, 2021(+).)

[7] J. de Leeuw and P. Mair. Multidimensional Scaling Using Majorization: SMACOF in R. *J. Stat. Softw.*, 31(3);1-30, 2009.

[8] S. López Pintado and J. Romo. On the Concept of Depth for Functional Data. *J. Am. Stat. Assoc.*, 104(486):718-734, 2009.

[9] J. Mateu, F. Schoenberg, D. Diez, J. González and W. Lu. On measures of dissimilarity between point patterns: classification based on prototypes and multidimensional scaling. *Biom. J.*, 57(2);340-358, 2015.

[10] J. Møller and R. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall/CRC, Boca Raton, 2004.

[11] M. Myllymäki, T. Mrkvička, P. Grabarnik, H. Seijo and U. Hahn. Global envelope tests for spatial processes. *J R Stat Soc B*, 79(2);381-404, 2017.

[12] W. Torgerson. Multidimensional Scaling: I. Theory and Method. *Psychometrika*, 17; 401419, 1952.

Look-ahead screening rules for the Lasso

Johan Larsson[1]*

[1]*The Department of Statistics, Lund University*

Abstract: The lasso is a popular method to induce shrinkage and sparsity in the solution vector (coefficients) of regression problems, particularly when there are many predictors relative to the number of observations. Solving the lasso in this high-dimensional setting can, however, be computationally demanding. Fortunately, this demand can be alleviated via the use of *screening rules* that discard predictors prior to fitting the model, leading to a reduced problem to be solved. In this paper, we present a new screening strategy: *look-ahead screening*. Our method uses safe screening rules to find a range of penalty values for which a given predictor cannot enter the model, thereby screening predictors along the remainder of the path. In experiments we show that these look-ahead screening rules outperform the active warm-start version of the Gap Safe rules.

Keywords: lasso, sparse regression, screening rules, safe screening rules
AMS subject classification: 62J07

1 Introduction

The lasso [6] is a staple among regression models for high-dimensional data. It induces shrinkage and sparsity in the solution vector (regression coefficients) through penalization by the ℓ_1-norm. The optimal level of penalization is, however, usually unknown, which means we typically need to estimate it through model tuning across a grid of candidate values: the regularization path. This leads to a heavy computational load.

Thankfully, the advent of so-called *screening rules* have lead to remarkable advances in tackling this problem. Screening rules discard a subset of the predictors *before* fitting the model, leading to, often considerable, reductions in problem size. There are two types of screening rules: heuristic and safe rules. The latter kind provides a certificate that discarded predictors cannot be active at the optimum—that is, have a non-zero corresponding coefficients—whereas heuristic rules do not. In this paper, we will focus entirely on safe rules.

A prominent type of safe rules are the Gap Safe rules [5, 1], which use the duality gap in a problem to provide effective screening rules. There currently exists sequential versions of the Gap Safe rules, that discard predictors for the next step

*Corresponding author: johan.larsson@stat.lu.se

on the regularization path, as well as dynamic rules, which discard predictors during optimization at the current penalization value.

The objective of this paper is to introduce a new screening strategy based on Gap Safe screening: *look-ahead screening*, which screens predictors for a range of penalization parameters. We show that this method can be used to screen predictors for the entire stretch of the regularization path, leading to substantial improvements in the time to fit the entire lasso path.

2 Look-Ahead Screening

Let $X \in \mathbb{R}^{n \times p}$ be the design matrix with n observations and p predictors and $y \in \mathbb{R}^n$ the response vector. The lasso is represented by the following convex optimization problem:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \left\{ P(\beta; \lambda) = \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{1}$$

where $P(\beta; \lambda)$ is the *primal* objective. We let $\hat{\beta}_\lambda$ be the solution to (1) for a given λ. Moreover, the dual problem of (1) is

$$\underset{\theta \in \mathbb{R}^n}{\text{maximize}} \left\{ D(\theta; \lambda) = \frac{1}{2} y^T y - \frac{\lambda^2}{2} \left\| \theta - \frac{y}{\lambda} \right\|_2^2 \right\} \tag{2}$$

where $D(\theta; \lambda)$ is the *dual* objective. The relationship between the primal and dual problems is given by $y = X\hat{\beta}_\lambda + \lambda \hat{\theta}_\lambda$.

Next, we let G be the so-called *duality gap*, defined as

$$G(\beta, \theta; \lambda) = P(\beta; \lambda) - D(\theta; \lambda) = \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 - \lambda \theta^T y + \frac{\lambda^2}{2} \theta^T \theta. \tag{3}$$

In the case of the lasso, strong duality holds, which means that $G(\hat{\beta}_\lambda, \hat{\theta}_\lambda; \lambda) = 0$ for any choice of λ.

Suppose, now, that we have solved the lasso for λ; then for any given $\lambda^* \leq \lambda$, the Gap Safe rule [5] discards the jth predictor if

$$|X^T \theta_\lambda|_j + \|x_j\|_2 \sqrt{\frac{1}{\lambda_*^2} G(\beta_\lambda, \theta_\lambda; \lambda^*)} < 1 \tag{4}$$

where

$$\theta_\lambda = \frac{y - X\beta_\lambda}{\max\left(|X^T(y - X\beta_\lambda)|, \lambda\right)}$$

is a dual-feasible point [5] obtained through dual scaling.

Observe that (4) is a quadratic inequality with respect to λ_*, which means that it is trivial to discover the boundary points via the quadratic formula:

$$\lambda_* = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad \text{where} \quad \begin{aligned} a &= \left(1 - |x_j^T \theta_\lambda|\right)^2 - \frac{1}{2} \theta_\lambda^T \theta_\lambda \|x_j\|_2^2, \\ b &= \left(\theta_\lambda^T y - \|\beta_\lambda\|_1\right) \|x_j\|_2^2, \\ c &= -\frac{1}{2} \|y - X\beta_\lambda\|_2^2 \|x_j\|_2^2. \end{aligned}$$

By restricting ourselves to an index j corresponding to a predictor that is inactive at λ and recalling that we have $\lambda_* \leq \lambda$ by construction, we can inspect the signs of a, b, and c and find a range λ values for which predictor j must be inactive. Using this idea for the lasso path—a grid of λ values starting from the null (intercept-only) model, which corresponds to $\lambda_{\max} = \max_i |x_i^T y|$, and finishing at fraction of this (see section 3 for specifics)—we can screen predictor j for all upcoming λs, possibly discarding it for multiple steps on the path rather than just the next step. We call this idea *look-ahead screening*.

To illustrate the effectiveness of this screening method, we consider an instance of employing look-ahead screening for fitting a full lasso path to the *leukemia* data set [3]. At the first step of the path, the screening method discards 99.6% of the predictors for the steps up to and including step 5. The respective figures for steps 10 and 15 are 99.3% and 57%. At step 20, however, the rule does not discard a single predictor. In Figure 1, we have visualized the screening performance of look-ahead screening for a random sample of 25 predictors from this data set.

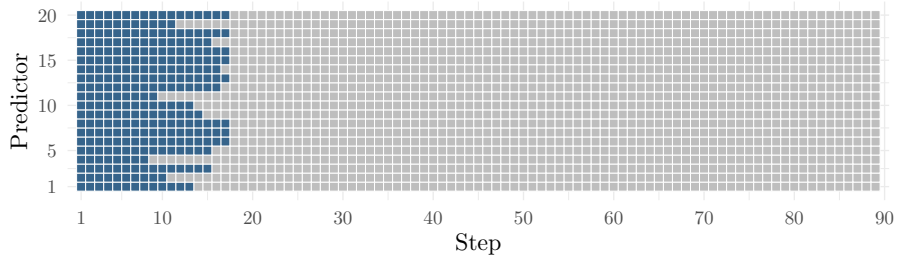


Figure 1: This figure shows the predictors screened at the first step of the lasso path via look-ahead screening for a random sample of 20 predictors from the *leukemia* data set. A blue square indicates that the corresponding predictor can be discarded at the respective step.

As is typical for all screening methods, the effectiveness of look-ahead screening is greatest at the start of the path and diminishes as the strength of penalization decreases later on in the path. Note, however, that all of the quantities involved in the rule are available as a by-product of solving the problem at the previous step, which means that the costs of look-ahead screening are diminutive.

3 Simulations

In this section, we study the effectiveness of the look-ahead screening rules by comparing them against the active warm start version of the Gap Safe rules [1, 5]. We follow the recommendations in [5] and run the screening procedure every tenth pass of the solver. Throughout the experiments, we center the response vector by its mean, as well as center and scale the predictors by their means and uncorrected sample standard deviations respectively.

To construct the regularization path, we employ the standard settings from glmnet, using a log-spaced path of 100 λ values from $\lambda_{\max}$ to $\varepsilon\lambda_{\max}$, where $\varepsilon = 10^{-2}$

if $p > n$ and 10^{-4} otherwise. We also use the default path stopping criteria from glmnet, that is, stop the path whenever the deviance ratio, $1 - \text{dev}/\text{dev}_{\text{null}}$, is greater than or equal to 0.999, the fractional increase in deviance explained is lower than 10^{-5}, or, if $p \geq n$, when the number of active predictors exceeds or is equal to n.

To fit the lasso, we use cyclical coordinate descent [2]. We consider the solver to have converged whenever the duality gap as a fraction of the primal value for the null model is less than or equal to 10^{-6} and the amount of *infeasibility*, which we define as $\max_j \left(|x_j^T(y - X\beta_\lambda)| - \lambda\right)$, as a fraction of $\lambda_{\max}$ is lower than or equal to 10^{-5}.

Source code for the experiments, including a container to facilitate reproducibility, can be found at https://github.com/jolars/LookAheadScreening/. An HPC cluster node with two Intel Xeon E5-2650 v3 processors (Haswell, 20 compute cores per node) and 64 GB of RAM was used to run the experiments.

We run experiments on a design with $n = 100$ and $p = 50\,000$, drawing the rows of X i.i.d. from $\mathcal{N}(0, \Sigma)$ and y from $\mathcal{N}(X\beta, \sigma^2 I)$ with $\sigma^2 = \beta^T \Sigma \beta / \text{SNR}$, where SNR is the signal-to-noise ratio. We set 5 coefficients, equally spaced throughout the coefficient vector, to 1 and the rest to zero. Taking inspiration from **(author?)** [4], we consider SNR values of 0.1, 1, and 6.

Judging by the results (Figure 2), the addition of look-ahead screening results in sizable reductions in the solving time of the lasso path, particularly in the high signal-to-noise context.

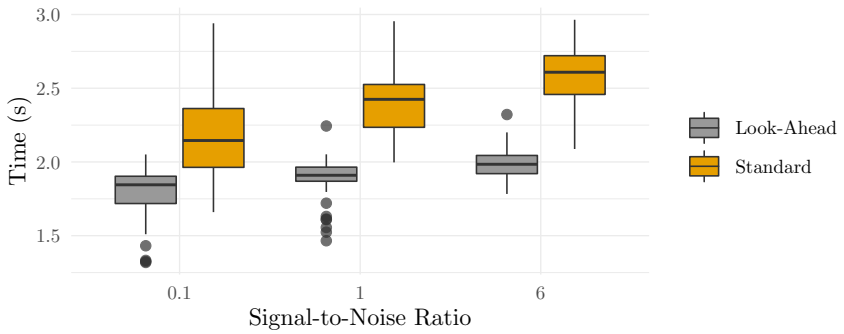


Figure 2: Standard box plots of timings to fit a full lasso path to a simulated data set with $n = 100$, $p = 50\,000$, and five true signals.

4 Discussion

In this paper, we have presented *look-ahead screening*, which is a novel method to screen predictors for a range of penalization values along the lasso regularization path using Gap Safe screening. Our results show that this type of screening can yield considerable improvements in performance for the standard lasso. For other loss functions, (4) may no longer reduce to a quadratic inequality and will hence

require more computation. Nevertheless, we believe that applying these rules in these cases is feasible and likely to result in comparable results.

Moreover, the idea is general and can therefore be extended to any type of safe screening rule and also used in tandem with heuristic screening rules in order to avoid expensive KKT computations. Finally, although we only cover one type of cyclical coordinate descent in our experiments, note that our screening method is agnostic to the solver used and that we expect the results hold for any solver that benefits from predictor screening.

Acknowledgements: I would like to thank my supervisor, Jonas Wallin, for valuable feedback on this work. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at LUNARC partially funded by the Swedish Research Council through grant agreement no. 2017-05973.

Bibliography

[1] O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: Safer rules for the lasso. In F. Bach and D. Blei, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 333–342, Lille, France, July 2015. PMLR.

[2] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

[3] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct. 1999. ISSN 0036-8075. doi: 10.1126/science.286.5439.531.

[4] T. Hastie, R. Tibshirani, and R. Tibshirani. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, Nov. 2020. ISSN 0883-4237. doi: 10. 1214/19-STS733.

[5] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(128):1–33, 2017.

[6] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246.

Adversarial attacks against Bayesian forecasting dynamic models

Roi Naveiro[1*]

[1]*Institute of Mathematical Sciences (ICMAT-CSIC), Madrid, Spain*

Abstract: The last decade has seen the rise of Adversarial Machine Learning (AML). This discipline studies how to manipulate data to fool inference engines, and how to protect those systems against such manipulation attacks. Extensive work on attacks against regression and classification systems is available, while little attention has been paid to attacks against time series forecasting systems. In this paper, we propose a decision analysis based attacking strategy that could be utilized against Bayesian forecasting dynamic models.

Keywords: Bayesian forecasting, Adversarial machine learning, Bayesian model monitoring.

AMS subject classification: 62H30, 91A35.

1 Introduction

Machine learning (ML) applications have experienced an impressive growth over the last decade. However, the ever increasing adoption of ML methodologies has revealed important security issues. Among these, vulnerabilities to adversarial examples [3], intentionally manipulated data instances targeted at fooling ML algorithms, are especially important. In contexts in which ML systems are susceptible of being attacked, algorithms should acknowledge the presence of possible adversaries and be trained in such a way that they are robust against their potential data manipulations. This is the main goal of AML. As recently pointed out in [4], having relevant models of how an adversary might modify input data to a learning system is key to guarantee protection against adversarial attacks. In the last few years, several attacking models to classification and regression systems have been released. However, much less attention has been paid to attacks to time series forecasting models. In this paper, we put ourselves in the shoes of an adversary, willing to manipulate input data to a time series forecasting system in order to drive predictions to a target of his interest. This is a crucial step prior to developing robust defence mechanisms.

*Corresponding author: roi.naveiro@icmat.es

2 Attacks against Bayesian forecasting dynamic models

Previously proposed attacks against time series forecasting models [1] focused on attacking simple auto-regressive (AR) models. Their idea is straightforward: an adversary is interested in driving the predictions made by an AR model during a fixed time window towards a target of his interest. To do so, he deliberately modifies the observations received by the AR system prior to the onset of the time window. The adversary is assumed to have complete knowledge about the parameters governing the AR model and the original data that would be fed into the system. In this case, given a candidate data manipulation, the attacker could compute the vector of predictions that the model would yield. A data manipulation is then selected to minimize the distance between the predictions produced and the target predictions, subject to some constraints about the size of the data manipulation reflecting a limited budget or the fact that the adversary wants to avoid being detected.

In this paper, our main purpose is to propose attacking strategies that target Bayesian forecasting dynamic models [6]. Rather than extending previous attacks to target these models, we build a novel, more realistic, attacking strategy based on decision analysis ideas. Let us start with the setting. We consider two agents: a Defender (D, she), which is implementing a Bayesian forecasting dynamic model to aid her in making some decision, and an attacker (A, he), manipulating the data that D receives, in order to modify such a decision. As a running example, consider that D is an ad company monitoring traffic flow into one of its webpages, using a conditionally Poisson dynamic model [2]. At some point, based on predictions for the next few days, D must decide whether to place an ad in the monitored node or not. In turn, A modifies the input data of the model by producing fake connections to the website in order to induce D to make a wrong decision. For instance, A could be interested in making D spend resources on placing the ad, when, from a decision analysis point of view, was not convenient.

To fix ideas, say D is using a Bayesian dynamic model to forecast a quantity y_t of interest. Inferences about this quantity are updated sequentially as new data are observed, with $\mathcal{D}_t$ representing all information available at time t.[1] When $t = \alpha$, the Defender must make a decision based on her forecasts from time $\alpha + 1$ until time β. From a decision analysis perspective, her optimal decision should maximize her posterior predictive utility

$$d^* = \arg\max_d \Psi(d|\mathcal{D}_\alpha) = \arg\max_d \int u(d, y_{\alpha+1:\beta}) p(y_{\alpha+1:\beta}|\mathcal{D}_\alpha)\, \mathrm{d}y_{\alpha+1:\beta},$$

where $u(d, y_{\alpha+1:\beta})$ is the utility perceived from making decision d when future data is $y_{\alpha+1:\beta}$ and $p(y_{\alpha+1:\beta}|\mathcal{D}_\alpha)$ is the posterior predictive distribution over the next $\beta - \alpha$ periods at time α.

Consider now A's problem. We study the case of a very powerful attacker: A has knowledge about D's utility, her probability model, and also the observations that the defender receives from time 0 until time α. For instance, A could be

[1] $\mathcal{D}_t$ is recursively defined through $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{y_t\}$.

an insider within D's company. Assume that A is interested in modifying the observations to be received by D from time $t = \alpha - h$ until time $t = \alpha$, so that, D will update her model with contaminated data $\tilde{y}_{\alpha-h}, \ldots, \tilde{y}_{\alpha}$ rather than with clean data $y_{\alpha-h}, \ldots, y_{\alpha}$. We denote the attacked data until time α as $\tilde{\mathcal{D}}_{\alpha}$. Through these modified data, A's goal is to make the Defender decide d_A instead of d^*. A main difference between our approach and previous ones is that earlier work on attacks to time series forecasting systems considered that the attacker modifies data to drive the predictions produced by the defender towards a certain target. Instead, we consider that a more encompassing objective to the attacker is to make the defender decide something inconvenient.

Clearly, not every data manipulation that induces decision d_A is equally interesting for A. He might have limited resources and, more importantly, he would typically want to avoid being detected. In previous work, this later goal was formalized limiting the size of the perturbation $\|\tilde{y}_{\alpha-h:\alpha} - y_{\alpha-h:\alpha}\|$, under certain norm. This would yield the following optimization problem to be solved by the attacker when looking for optimal manipulations

$$\min_{\tilde{y}_{\alpha-h:\alpha}} \quad \|\tilde{y}_{\alpha-h:\alpha} - y_{\alpha-h:\alpha}\| \quad \text{s.t.} \quad \Psi(d_A|\tilde{\mathcal{D}}_{\alpha}) > \Psi(d|\tilde{\mathcal{D}}_{\alpha}) \quad \forall d. \tag{1}$$

Resource limitations could be incorporated as additional constraints.

We argue that this way of formalizing the goal of wanting to avoid detection, could be inconvenient in several scenarios: even slight changes in the data, if they happen always in the same direction, could induce structural changes in time series that can be easily detected by an appropriate monitoring strategy. Imagine that D is indeed monitoring the predictive performance of her Bayesian forecasting model using the strategy described in [5]. The main idea is to compare, at each time, the predictive performance of the current model, with that of an alternative model using the *local Bayes factor* $H_t = p(y_t|\mathcal{D}_{t-1})/p_A(y_t|\mathcal{D}_{t-1})$. A small H_t indicates low predictive performance at time t, and, thus, y_t is considered discrepant. However, in order to be able to detect not only single discrepant points but also structural changes, [5] proposes looking for the most discrepant group of recent, consecutive observations, which entails calculating at each time the *minimum cumulative Bayes factor* $V_t = \min_{1\leq k\leq t} H_t H_{t-1} \ldots H_{t-k+1}$. This can be sequentially computed as $V_t = H_t \min[1, V_{t-1}]$. The basic diagnostic mode of operation of D would be to accept the current model as satisfactory unless V_t falls below some threshold value γ. This monitoring strategy has been proved to be useful to detect model failures happening due to outliers or structural changes.

Thus, to avoid detection, data manipulations should not produce significant decrease in the the minimum of the cumulative Bayes factors. Otherwise, they could trigger the monitor. We therefore propose looking for attacks, solving the problem

$$\max_{\tilde{y}_{\alpha-h:\alpha}} \quad \min \tilde{V}_{\alpha-h:\alpha} \quad \text{s.t.} \quad \Psi(d_A|\tilde{\mathcal{D}}_{\alpha}) > \Psi(d|\tilde{\mathcal{D}}_{\alpha}) \quad \forall d, \tag{2}$$

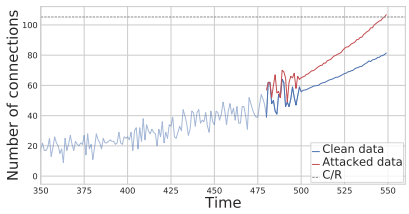
where $\tilde{V}_t$ is the minimum of the cumulative Bayes factors at time t, under attacked data. We believe that this formalization of the concept of *avoiding detection* is more coherent than that in (1), as it modifies data trying to mimic the predictive

behaviour of the model being used by D, and thus produces tainted data points that are aligned with D's beliefs.

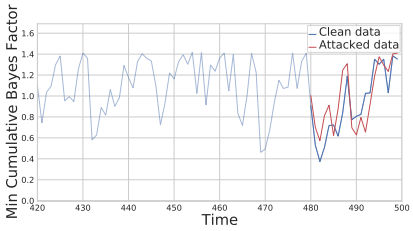
Of course, solving problems (1) and (2) exactly is unfeasible, and we must use heuristic methods to approximate the optimal solution. In the next Section, problem (2) is approximately solved using simulated annealing. Algorithmic details are not shown due to lack of space.

3 Case study

Continuing with our running example[2], imagine that at time $\alpha = 500$ the ad company (D) must decide about placing an ad on the monitored node, that will appear at a certain time $\beta = 550$ in the future, in which the company is interested on. The cost of placing the ad is $C = 100$. The reward perceived by the company per user watching the ad is $R = 0.95$. Assuming that D is risk neutral, it is straightforward to see that her posterior predictive utility is $R \cdot \mathbb{E}[y_\beta|\mathcal{D}_\alpha] - C$ if she decides to place the ad and 0 otherwise. $\mathbb{E}[y_\beta|\mathcal{D}_\alpha]$ is the posterior predictive mean for the number of connections at time β. Thus, D should place the ad if $\mathbb{E}[y_\beta|\mathcal{D}_\alpha] > \frac{C}{R}$. Having a good estimate of $\mathbb{E}[y_\beta|\mathcal{D}_\alpha]$ is crucial to inform D's decision. To that end, D fits a conditionally Poisson model with local linear growth in the latent process [2]. To fit this model, D uses data of previous traffic flow from time 0 until time $\alpha = 500$. The blue line in Figure ?? shows the original data and the Monte Carlo estimate of the predictive mean from $t = 500$ to $t = 550$. As can be seen, the predictive mean of the number of connections at time $t = 550$ is around 80, far below $C/R \simeq 105$. Thus, the company should decide not to spend resources on placing the ad.



(a) Attacked data and forecast for predictive mean



(b) Min cumulative Bayes factor, V_t

Figure 1: Attacked and original data, forecast for predictive mean and V_t

Now imagine that A, an insider in the company, has resources to create fake connections to the monitored node from time $t = \alpha - h$ (with $h = 20$) until time $t = \alpha$. His goal is to make D waste resources in placing the ad (when it is not recommended as we have seen). He can just create fake connections that will be added to the normal traffic flow. D uses a monitoring system as described in

[2]The code to reproduce this experiment is available at https://github.com/roinaveiro/attacksSSMs

Section 2. A knows this and, to avoid detection, chooses his attack approximately solving (2). The red line in Figure 1a shows that, adding few connections from time $t = 480$ to $t = 500$, will produce a flip of D's optimal strategy, as the tainted predictive mean at time $t = 550$ surpasses C/R. In addition, these connections are added in a very subtle way, and they do not trigger an alarm as can be seen in Figure 1b. We see that the minimum cumulative Bayes factor for the attacking period is not substantially different from that under original data.

4 Discussion

We have presented a decision analysis-based approach to generate data manipulation attacks against Bayesian forecasting dynamic models. Our framework has two central differences with previous approaches: (1) the attacker's goal is to change the decision made by the defender, rather than drive the predictions towards an specific target; (2) the formalization of "subtle" attack is based on cumulative Bayes factors rather than the size of the manipulations.

Several further directions of research could be explored. First of all, we have assumed an attacker that has full knowledge about the defender. Considering limited knowledge cases is an interesting way to go. Secondly, we have focused on static attacks: the attacker makes a single data manipulation decision in view of D's model and the data that D will be receiving. A more realistic case would be the dynamic attack: the attacker decides the attack for the next time period, the defender updates the manipulated data point and updates her model, then the attacker decides the next attack, an so on.

Acknowledgements: I would like to thank SAMSI, AXA, the FBBVA, the Trustonomy project and professor Mike West for insightful discussions.

Bibliography

[1] S. Alfeld, X. Zhu, and P. Barford. Data poisoning attacks against autoregressive models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[2] X. Chen, K. Irie, D. Banks, R. Haslinger, J. Thomas, and M. West. Scalable bayesian modeling, monitoring, and analysis of dynamic network flow data. *Journal of the American Statistical Association*, 113(522):519–533, 2018.

[3] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[4] R. Naveiro, A. Redondo, D.R. Insua, and F. Ruggeri. Adversarial classification: An adversarial risk analysis approach. *International Journal of Approximate Reasoning*, 113:133 – 148, 2019.

[5] M. West. Bayesian model monitoring. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(1):70–78, 1986.

[6] M. West and J. Harrison. Bayesian forecasting and dynamic models. Springer Science Business Media, 2006.

Yield curve modelling in insurance

Matúš Padyšák[1*]

[1]*Comenius University; Faculty of Mathematics, Physics and Informatics; Department of Applied Mathematics and Statistics*

Abstract: The paper is focused on the application of yield curves in the insurance and to regulations of Solvency II (see [2]). Legislative programme Solvency II requires for a yield curve to be fitted on liquid instruments until the last liquid point, from which the curve is extrapolated to the Ultimate Forward Rate for a long pre-determined horizon. The forward curve has to converge to the Ultimate Forward Rate with a pre-determined accuracy of one basis point. The aim is to calibrate Solvency II consistent parametric models and compare them to the spline-based Smith-Wilson model, which is suggested by the European Insurance and Occupational Pensions Authority.

Keywords: yield curve, Ultimate Forward Rate, calibration, testing
AMS subject classification: 91B30, 91G30

1 Introduction

The primary purpose of yield curves is, based on yields observed in the market, to interpolate yields that cannot be observed. In insurance, yield curves are crucial to discount future cash-flows using market-consistent rates. Moreover, as a requirement of the European Insurance and Occupational Pensions Authority (EIOPA), yield curves must comply with the Solvency II framework. Forward curves must converge at the pre-determined horizon to Ultimate Forward Rate (UFR).

EIOPA suggests using the spline-based Smith-Wilson model. On the one hand, the model is easy to calibrate, and the optimization which is necessary for the regulatory requirements is time-efficient and straightforward. The reason is that the UFR is the model's parameter, and we only need to adjust the speed of convergence. On the other hand, the model could perform worse in terms of accuracy of out-of-sample yields compared to other models.

Natural counterparts to the Smith-Wilson model are parametric models such as Nelson-Siegel, Svensson model or the extension of the Svensson model, Five-factor model. These parametric models can be viewed as non-linear regression models, where the convergence to the UFR is highly uncertain. The key question of this paper is whether it is possible to calibrate UFR consistent parametric yield curves. Additionally, the aim is to compare the Smith-Wilson model with parametric models since the Smith-Wilson model has a perfect fit for the known yields. Still, it does not mean that it does not have greater error for yields of unobserved maturities.

*Corresponding author: matus.padysak@fmph.uniba.sk

2 Models

Since the parametric models are the main focus of the paper, for the Smith-Wilson model, we refer to the Technical documentation of EIOPA (see [2]).

Definition 1. Let us define by z the time to maturity of zero-coupon bond measured in months. We define the general parametric (spot) yield function at time t as:

$$y_t(z) = \beta_{0t} + \beta_{1t}\Big(\frac{1-exp(-\frac{z}{\lambda_{1t}})}{\frac{z}{\lambda_{1t}}}\Big) + \beta_{2t}\Big(\frac{1-exp(-\frac{z}{\lambda_{1t}})}{\frac{z}{\lambda_{1t}}} - exp(-\frac{z}{\lambda_{1t}})\Big) +$$
$$+\beta_{3t}\Big(\frac{1-exp(-\frac{z}{\lambda_{2t}})}{\frac{z}{\lambda_{2t}}} - exp(-\frac{z}{\lambda_{2t}})\Big) + \beta_{4t}\Big(\frac{1-exp(-\frac{z}{\lambda_{2t}})}{\frac{z}{\lambda_{2t}}}\Big),$$

and the general instantaneous parametric forward yield function as:

$$f_t(z) = \beta_{0t} + \beta_{1t}exp(-\tfrac{z}{\lambda_{1t}}) + \beta_{2t}\tfrac{z}{\lambda_{1t}}exp(-\tfrac{z}{\lambda_{1t}}) + \beta_{3t}\tfrac{z}{\lambda_{2t}}exp(-\tfrac{z}{\lambda_{2t}}) + \beta_{4t}exp(-\tfrac{z}{\lambda_{2t}}).$$

If $\beta_{3t} = \beta_{4t} = 0$, the general model is labelled as the Nelson-Siegel model. If $\beta_{4t} = 0$ and $\lambda_{1t} \neq \lambda_{2t}$, the general model is labelled as the Svensson model. Lastly, if $\lambda_{1t} \neq \lambda_{2t}$, the general model is labeled as the Five-Factor model.

3 Calibration

Market consistent curves

In general, the aim is to find model parameters based on the observed bond yields such that the difference between fitted and observed yields would be minimal. In other words, to find the optimal parameters based on a given objective function that measures the error between observed and fitted yields. Throughout the paper, we use the mean-squared error as the objective function:

$$MSE_t = \textstyle\sum_{n=1}^{N}(y_{t,n} - \hat{y}_{t,n})^2,$$

where N is the number of observed bonds.

One strand of the literature suggests a linearization by fixing the λ parameters to a constant [1]. However, the parameter is not time-dependent which could be detrimental for the long-term accuracy.

We suggest finding the optimal time-dependent λ parameters iteratively (for example, for the Svensson model on a grid). At time t, we fix λ parameters and find estimates using OLS. For each iteration, we compute the MSE and find λ parameters such that the MSE is minimal.

Therefore, we define a set $\Omega = \{(\lambda_{1t}, \lambda_{2t}); \lambda_{1t} \in T_1, \lambda_{2t} \in T_2\}$, where $T_1 = \{a_1 + bc\}_{c=0}^{d_1}$ and $T_2 = \{a_2 + bc\}_{c=0}^{d_2}$, a_1, a_2, b, d_1 and d_2 are appropriate constants. The estimated λ coefficients are equal to the:

$$(\hat{\lambda_{1t}}, \hat{\lambda_{2t}}) = \arg\max_{(\lambda_{1t},\lambda_{2t})\in\Omega} MSE_t.$$

Another strand of the literature considers calibration as a non-linear optimization problem. [3] compared direct-search (Nelder-Mead, Powell), stochastic (Simulated annealing) and gradient-based algorithms. Unsurprisingly, given the numerous local optima of yield curves, direct-search and stochastic algorithms performed with significantly lower errors.

Throughout the paper, we opt for the direct search Nelder-Mead algorithm (NM), linearization (LM) and Differential Evolution (DE) algorithm (see Table 1). The latter could be potentially superior to Simulated annealing since it utilizes a whole population of possible solutions, and apart from the stochastic element that allows jumping out of local optima, the algorithm also utilizes mutation, recombination and selection.

UFR consistent curves

Firstly, the curves must be fitted based on yields with maturity up to the last liquid point. The last liquid point (LLP) depends on the currency, while for pound sterling, the last liquid point is equal to 50 years, for Polish zloty, the point is 10 years. For the eurozone, the LLP is 20 years. Additionally, the forward curve has to converge to the Ultimate Forward Rate for a maximum of $20 + LLP$ and 60 years with accuracy of one basis point (0.01%). Each year the UFR is published by the EIOPA and it is a sum of the expected inflation and real interest rate. Each year, the UFR can change by the maximum of 15 basis points (see [2]). To sum it up, forward curves have to be extrapolated from the last liquid point to the UFR and have to converge with a pre-defined accuracy.

In practice, this is problematic for parametric models since parametric models do not have UFR as a model parameter to which they converge to. For $z \to \infty$ they converge to β_0. However, even if we would fix β_0 to UFR, the maturity of the maximum of $20 + LLP$ and 60 years might not be long enough. Our proposed heuristic could solve these problems:

1. Calibration of spot yield curve based on observed bond yields.

2. Calculation of fitted forward rates for the observed maturities using estimated parameters from the first step.

3. Addition of the UFR rate as observed forward rate for the maturity of $\max(20 + LLP, 60)$ years.

4. Calibration of the forward yield curve where the observed forward rates are the fitted rates for known maturities and UFR rate. From this point, any calibration method mentioned above could be used, where instead of spot yields, forward yields are used.

5. As the final step, we need to check whether the fitted UFR rate does not differ from the published rate by more than 1 basis point. If it does, β_0 is modified so that the condition holds true - the curve is shifted up or down to satisfy the condition. The final parameters can be used to calculate fitted spot or forward rates.

4 Testing

Traditionally, yield curves are compared based on average errors (e.g. MSE or RMSE). Residuals of parametric models do not tend to be normally distributed, and for each curve, there are only a few observations. Moreover, these metrics make it impossible to compare parametric models with the Smith-Wilson model since it has a perfect fit for maturities of observed yields.

For Solvency II consistent curves, we propose the leave-k-out approach where we randomly drop some observed yields from the sample and calibrate the curves. This allows for the evaluation of the curves out-of-sample since we can compare the fitted yields for dropped maturities with real observed yields. Additionally, we can measure the errors using MSE and using a long history of curves, we can test the differences (mean or median value) across models with statistical tests.

Method	Average MSE	Standard deviation
NM	0.0808344	0.06223664
LM	0.06005777	0.04971987
DE	0.0765261	0.05414692

Table 1: Comparison of optimization methods for Five-Factor model, France

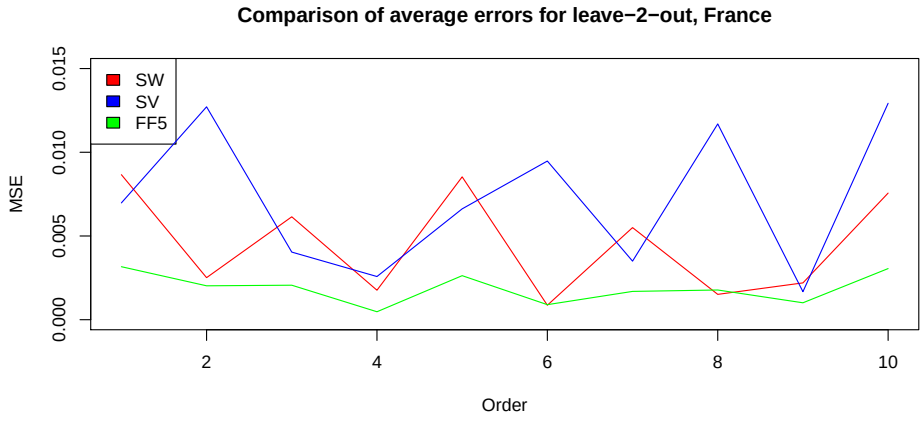


Figure 1: Tenfold leave-2-out validation, France, DE optimization for parametric models

5 Results and Discussion

Data includes sixteen monthly observations of zero-coupon government bonds of France for maturities ranging from one month to approximately 20 years. Dataset

spans from 31.9.2009 to 31.8.2018 and is merged from Thomson Reuters Eikon and Investing.com. A subset of this dataset is also used for the Solvency II consistent curves. This set consists of monthly observation for years 2017-2019 (given by the availability of published UFRs; **2017: 4.2%, 2018: 4.05%, 2019: 3.9%**).

Lastly, we provide results for the comparison of parametric (Svensson, Five-factor) and Smith-Wilson models (see Fig. 1). We repeat the leave-2-out approach ten times during the whole UFR sample to obtain mean-squared errors for each month in the sample. Based on the Sign test (non-symmetrical distribution) and 5% significance level, these differences are significant in eight cases for Smith-Wilson and Five-factor comparison. For both Five-Factor and Smith-Wilson models compared to the Svensson model, differences are significant in all cases. Lastly, we cannot conclude that the Five-factor model is superior to the Smith-Wilson model for every timeframe or country. Perhaps, the decision which model to use should be individual for each country and market situation.

Bibliography

[1] F. X. Diebold and C. Li (2006). *Forecasting the Term Structure of Government Bond Yields.*. Journal of Econometrics, 130, 337-364.

[2] EIOPA (2019). *Technical documentation of the methodology to derive EIOPA's risk-free interest rate term structures*, Published on-line https://www.eiopa.europa.eu/sites/default/files/risk_free_interest_rate/12092019-technical_documentation.pdf

[3] P. Manousopoulos and M. Michalopoulos (2009). *Comparison of non-linear optimization algorithms for yield curve estimation.* European Journal of Operational Research, 192, 594-602.

[4] R Core Team (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[5] R. B. de Rezende (2009). *Modeling and Forecasting the Yield Curve by an Extended Nelson-Siegel Class of Models: A Quantile Autoregression Approach.* https://ssrn.com/abstract=1290741.

[6] Y. S. Stander (2005). *Yield Curve Modeling*, first edition, PALGRAVE MACMILLAN New York. ISBN 978–1–4039–4726–0.

[7] R. Storn and K. Price (1997). *Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces.* Journal of Global Optimization, 11, 341-359.

An application of a geometric process model for debugging and testing costs

Mustafa Hilmi Pekalp,[1*] **and Halil Aydoğdu**[2]

[1] *Ankara University, Faculty of Applied Sciences, Department of Actuarial Sciences*
[2] *Ankara University, Faculty of Science, Department of Statistics*

Abstract: This paper models debugging and testing costs for a software product using geometric process (GP) model. Since these costs are based on the mean value and second moment functions of a GP model, it is necessary to compute these functions to calculate the costs. However, the computation procedure depends on both distributional assumption of the first interarrival time of the GP and the estimations of the model and distribution parameters. In this study, we consider the software system failure data which agrees with a GP model. Since it has been shown in previous studies that this data set can be modeled by a particular GP with gamma distribution, the model and distribution parameters are estimated under this distributional assumption. Then, the mean value and second moment functions of the GP model are calculated with the help of these estimators. Finally, by using the values of mean value and second moment functions, the cost functions are obtained for the data set.

Keywords: Geometric process, geometric function, second moment function, cost function, software reliability

AMS subject classification: 60K99, 62M99, 65Z05

1 Introduction

The GP is a stochastic monotone model which is widely used in various fields of statistics and applied probability since its introduction. Many researchers and authors made a significance effort on GP by publishing a considerable amount of research papers. For example, the GP is used as a model in reliability analysis, maintenance, warranty analysis, electricity prices and epidemic disease. This process has following definition.

Definition 1. Let $\{N(t), t \geq 0\}$ be a counting process (CP) and X_k be the interarrival time between $(k-1)$th and kth event of this process for $k = 1, 2, \ldots$. The CP $\{N(t), t \geq 0\}$ is said to be a GP with the ratio a if there exists a real number $a > 0$ such that $a^{k-1}X_k$ for $k = 1, 2, \ldots$ generate a renewal process (RP) with a common distribution function F_1 where F_1 is the distribution function of the first interarrival time X_1.

*Corresponding author: mustafa.hilmi.pekalp@ankara.edu.tr

The GP is also called quasi-renewal process with ratio parameter $\alpha = 1/a$ by [4].

Consider a GP model $\{N(t), t \geq 0\}$ with the ratio parameter a and let F_k be the distribution function of $X_k, k = 1, 2, \ldots$. It is clear from the definition of the GP that F_k can be uniquely determined by the distribution function of the first interarrival time X_1 of a GP, that is, $F_k(x) = F_1(a^{k-1}x)$ for $k = 1, 2, \ldots$. The sequence $\{X_k, k = 1, 2, \ldots\}$ is stochastically increasing if $a < 1$ and stochastically decreasing if $a > 1$, respectively. If $a = 1$, then the GP reduces to a RP.

Let $\{N(t), t \geq 0\}$ be a GP with ratio parameter a and let $*$ denote the Stieltjes convolution. The mean value function of a GP, which is also called the geometric function, is given by

$$M_1(t) = E(N(t)) = \sum_{k=0}^{\infty} F_1 * F_2 * \ldots * F_k(t), t \geq 0. \tag{1}$$

$M_1(t)$ satisfies the following integral equation [12].

$$M_1(t) = F_1(t) + \int_0^t M_1(a(t-x))dF_1(x), t \geq 0. \tag{2}$$

The second moment function of a GP is given by

$$M_2(t) = E(N^2(t)) = 2\sum_{k=0}^{\infty} k(F_1 * \ldots * F_k(t)) - \sum_{k=0}^{\infty} F_1 * \ldots * F_k(t), t \geq 0. \tag{3}$$

[9] showed that $M_2(t)$ satisfies the following integral equation.

$$M_2(t) = 2M_1(t) - F_1(t) + \int_0^t M_2(a(t-x))dF_1(x), t \geq 0. \tag{4}$$

For $a \leq 1$, $M_1(t)$ and $M_2(t)$ are finite for all $t \geq 0$. Further, if F_1 is continuous, then the integral equations (2) and (4) can be solved uniquely although $M_1(t)$ and $M_2(t)$ cannot be obtained in closed forms. If $a > 1, M_1(t)$ and $M_2(t)$ are infinite for all $t \geq 0$. The proof of these results can be found, for example, in [12].

It can be easily seen from the definition of the functions $M_1(t)$ and $M_2(t)$ that the analytical forms of these functions do not exist. In the literature, many authors make a significant effort on the computations of these functions. Known studies on the functions $M_1(t)$ and $M_2(t)$ of the GP can be given as follows: [2], [3], [8]-[10] and [13].

Let $\{N(t), t \geq 0\}$ be a GP with the ratio parameter a and assume that $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$. Thus, it can be obtained that $E(X_k) = \mu/a^{(k-1)}$ and $Var(X_k) = \sigma^2/a^{(2(k-1))}, k = 1, 2, \ldots$. Obviously, the parameters a, μ and σ^2 are the very crucial in the context of GP since these parameters completely determine the mean and variance of X_k. The estimation problem of these parameters is extensively studied in the literature. see [1], [6], [11]. In all studies cited, authors propose maximum likelihood (ML) estimators for the distribution considered. For this reason, in this study, we consider only ML estimators for the model parameters of the GP.

As one can easily notice from the studies above, the computation and, further, estimation procedures depend on distributional assumption of the first interarrival time. For this reason, it is an important problem to discriminate the distribution of the first interarrival time in GP model. It is well known that gamma, lognormal or Weibull distributions can be used quite effectively to analyze the data from a series of events. [7] deal with the problem of selecting one of these distributions for a given data set which is consistent with the GP model according to T-statistic based on the ratio of the maximized likelihood (RML). [7] conclude that T-statistic based on RML performs better than Kolmogorov-Simirnov, mean square error and maximum percentage error criteria according to their extensive simulation study. After validating the distribution for the data set which is consistent with the GP model, [7], further, calculate the estimates of the parameters by using the suitable method given in [6] for gamma, [1] for Weibull or [11] for lognormal.

2 Problem Description

The software testing is a powerful tool to remove bugs (faults) from the software products. However, an extensive testing procedure in a large program may not be reasonable. Debugging and testing decrease the error contents but this also increases the improvement costs. Actually, after reaching a certain level of software refinement, more studies on increasing reliability will cause much increase in cost and debugging time. Thus, it is essential to determine when to stop testing or when to release the product [4]. [4] propose a GP $\{N(t), t \geq 0\}$ to model the detection of software bugs where $N(t)$ denotes the number of software bugs in the interval $(0, t]$ for each fixed $t \geq 0$. In this model, suppose that the cost of fixing the ith software bug is a random variable, W_i, and we consider two parts for this random variable, a deterministic part c_0 and an incremental random part $(i-1)U$, that is, $W_i = c_0 + (i-1)U, i = 1, 2, \ldots$ where c_0 is a constant and U is a random variable with mean c_1. It is worth to note that this assumption is plausible because the fixing cost of a software bug is increasing as the number of bugs removed is increasing. Then, it is obvious that the expected total debugging cost in the interval $(0, t]$ is $E(\sum_{i=1}^{N(t)}(c_0 + (i-1)U)$. By conditioning on $N(t)$, this cost function is obtained depending on the functions $M_1(t)$ and $M_2(t)$ as

$$C_1(t) = \frac{(2c_0 - c_1)}{2} M_1(t) + \frac{c_1}{2} M_2(t), t \geq 0. \tag{5}$$

Further, if we assume that the cost of testing per unit time is a random variable with mean c_2, then the expected testing cost up to time t is tc_2 and if we consider this cost in the above cost model, we obtain the total expected testing and debugging cost up to time t.

$$C_2(t) = tc_2 + C_1(t), t \geq 0. \tag{6}$$

To compute the cost functions $C_1(t)$ and $C_2(t)$, it is clear that we need to calculate the first and second moments of the number of software bugs up to time t, i.e., $M_1(t)$ and $M_2(t)$. Furthermore, the values of $M_1(t)$ and $M_2(t)$ can be obtained

by determining the distribution of the first interarrival time of the GP model. The calculation procedure will be explained in the next section.

3 Calculation Procedure

Assume that the data set $\{X_1, X_2, \ldots, X_n\}$ comes from a GP with ratio parameter a. To calculate the cost functions $C_1(t)$ and $C_2(t)$, we will apply the following steps: Step 1. Calculate the estimators of the ratio parameter a and distribution parameters for each distribution considered, see [6] for gamma, [1] for Weibull and [11] for lognormal. Step 2. Compute their likelihood functions and use T-statistic based on RML to discriminate the distributions, see [7]. Step 3. Obtain the estimates of the model parameters associated with the validated distribution of the first interarrival time of the GP model. Step 4. Calculate the functions $M_1(t)$ and $M_2(t)$ with one of the suitable computation methods mentioned in the first section. Step 5. For a given t, c_0, c_1 and c_2, calculate the cost functions $C_1(t)$ and $C_2(t)$.

4 Real Data Application

Let us consider the software system failure data given in [5]. This data contains 136 failure times (in CPU seconds, measured in terms of execution time) of a real time command and control software system. As this data set contains three cases that the consecutive failure times are identical, interarrival times of these consecutive failure times are adjusted from 0 to 0.5. [12] prove that the data is consistent with a GP model. Moreover, [12] also show that the ratio parameter a of this GP model is less than 1. When we consider the results given in [7] for this data set, it can be concluded that the data set $\{X_1, X_2, \ldots, X_n\}$ comes from a particular GP with gamma distribution with shape parameter α and scale parameter β. The ML estimates of the model parameters are $\hat{a} = 0.9771, \hat{\alpha} = 0.7604$ and $\hat{\beta} = 123.7110$. According to the calculation procedure proposed by [13] and [9], $M_1(t)$ and $M_2(t)$ can be calculated iteratively by using the trapezoidal integration rule. When we apply this method, we calculate that $M_1(t) = 0.3005$ and $M_2(t) = 0.4068$ for $t = 20$. For a given $c_0 = 13.5, c_1 = 1$ and $c_2 = 1.2$, the cost functions can be computed as $C_1(t) = 4.1099$ and $C_2(t) = 28.1099$, respectively.

Bibliography

[1] H. Aydoğdu, B. Şenoğlu, M. Kara. Parameter Estimation in Geometric Process with Weibull Distribution. App. Math. and Comp., 217-6, 2657-2665, 2010.

[2] H. Aydoğdu, İ. Karabulut, E. Şen. On the Exact Distribution and Mean Value Function of a Geometric Process with Exponential Interarrival Times, Stat. and Prob. Let., 83, 2577-2582, 2013.

[3] H. Aydoğdu, İ. Karabulut. Power Series Expansions for the Distribution and Mean Value Function of a Geometric Process with Weibull Interarrival Times, Naval Res. Log., 61, 599-603, 2014.

[4] H. Pham, H. Wang. A Quasi-Renewal Process for Software Reliability and Testing Costs, IEEE Trans Syst Man Cybern.-Part A:Syst. and Hum., 31:6:623-631, 2001.
[5] J.D. Musa, A. Iannino, K. Okumoto. Software Reliability: Measurement, Prediction, Application, McGraw-Hill, New York, 1987.
[6] J.S.K. Chan, Y. Lam, D.Y. Leung. Statistical Inference for Geometric Processes with Gamma Distributions. Comp. Stat. and Data Ana., 47-3, 565-581, 2004.
[7] M.H. Pekalp, H. Aydoğdu, K.F. Türkman. Discriminating Between Some Lifetime Distributions in Geometric Counting Processes. Comm. in Stat.-Sim. and Comp., 10.1080/03610918.2019.1657452, 2019.
[8] M.H. Pekalp, H. Aydoğdu. An Asymptotic Solution of the Integral Equation for the Second Moment Function in Geometric Processes. J. of Comp. and App. Math., 353:179-190, 2019.
[9] M.H. Pekalp, H. Aydoğdu. An Integral Equation for the Second Moment Function of a Geometric Process and Its Numerical Solution. Naval Research Logistics, 65(2):176-184, 2018.
[10] M.H. Pekalp, H. Aydoğdu. Power Series Expansions for The Probability Distribution, Mean Value and Variance Functions of a Geometric Process with Gamma Interarrival Times, J. of Comp. and App. Math., 10.1016/j.cam.2020.113287, 2021.
[11] Y. Lam, J.S.K. Chan. Statistical Inference for Geometric Processes with Lognormal Distribution. Computational Statistics and Data Analysis, 27-1, 99-112, 1998.
[12] Y. Lam. The Geometric Process and Its Applications. World Scientific, Singapore, 2007.
[13] Y. Tang, Y. Lam, Numerical Solution to an Integral Equation in Geometric Process, Journal of Statistical Computation and Simulation 77, 549-560, 2007.

A regime switching on Covid-19 analysis and prediction in Romania

Marian Petrica,[1]* Radu D. Stochitoiu[2], Marius Leordeanu[3] and Ionel Popescu[4]

[1]*Faculty of Mathematics and Computer Science, University of Bucharest*
[2]*Faculty of Automatic Control and Computers, Polytechnic University of Bucharest*
[3]*Institute of Mathematics of the Romanian Academy and Polytechnic University of Bucharest*
[4]*Faculty of Mathematics and Computer Science, University of Bucharest and Institute of Mathematics of the Romanian Academy*

Abstract: Many problems in nature are modeled via a system of (partial) differential equations. For example the growth of a population, the spread of a virus, the evolution of weather, the heat distribution in a certain body, the flow of a fluid in a certain environment, models of dynamical systems, the evolution of the price of financial instruments and many other. Let's summarize this as a system of the form $X_t' = f(\alpha, X)$ where α is a set of parameters our model depends on. This equation can be also interpreted as a stochastic differential equation, but for the moment we will only consider the deterministic version of it.

In this framework there are two problems we want to treat. One is the determination of the parameters from a limited number of observations of the system at some given times. The other is the forecast on the system beyond the already observed data and eventually also prediction on what happened at intermediate times between the observations.

Keywords: statistical learning, Covid19, regimes, data
AMS subject classification: 92B20, 62P35

1 Introduction

In the fight with the COVID-19, quarantine was one of the main measures, at least when the hospitals were overwhelmed with patients and the virus propagation and its inside body working was not well understood.

A basic tool in analyzing the spread of the virus is the mathematical modeling. There is a growing body of mathematical models used at the moment as for a small sample by no means exhaustive see [1], [2], [3], [4], [5]. This turned out to

*Corresponding author: marianpetrica11@gmail.com

be a valuable tool that can be used in the assessment, prediction and control of infectious diseases, as it is the COVID-19 pandemic, which significantly impacted almost all countries, with important social and economical implications worldwide. The main purpose of this work is to develop a predictive model that can accurately assess the transmission dynamic of COVID-19.

In this paper we propose a regime separation for the analysis of Covid19 on Romania combined with mathematical models of SIR and SIRD. The main regimes we study are the free spread of the virus, the quarantine and partial relaxation and the last one is the relaxation regime. The main model we use is SIR which is a classical model, but because we can not fully trust the numbers of infected or recovered people we base our analysis on the number of deceased people which is more reliable. To actually deal with this we introduce a simple modification of the SIR model to account for the deceased separately. This in turn will be our base for fitting the parameters. We actually use the classical SIR model to detect the regime switching and in fact prove a proposition which shows that we can recover the parameters in a unique way from the daily observation of the number of infected and susceptible. This is the basis for guessing the main parameters in the model.

The actual estimation of the parameters in our SIRD model is done in two steps. The first one consists in training a neural network based on SIR models to detect the regime changes. Once this is done, we fit the main parameters of the SIRD model using a grid search near the values suggested by the neural network. At the end, we make some predictions on what the evolution will be in a timeframe of a month with the fitted parameters.

2 SIR Model, neural network and the parameters regime

We analyze the following SIR model: at time t, we consider $\bar{S}(t)$ as the number of susceptible individuals, $\bar{I}(t)$ as the number of infected individuals, and $\bar{R}(t)$ as the number of removed/recovered individuals. The equations of the SIR model are the following:

$$\begin{cases} \frac{d\bar{S}}{dt} = -\frac{\bar{\beta}\bar{S}\bar{I}}{N} \\ \frac{d\bar{I}}{dt} = \frac{\bar{\beta}\bar{S}\bar{I}}{N} - \gamma\bar{I} \\ \frac{d\bar{R}}{dt} = \gamma\bar{I} \end{cases} \tag{1}$$

Because there is no canonical choice of N, we will transform the system (1) by dividing it by N and considering $S(t) = \bar{S}(t)/N$, $I(t) = \bar{I}/N$ and $R(t) = \bar{R}(t)/N$. It is customary to choose $N = 10^6$ for convenience but this is just an arbitrary choice. For instance, analysis on smaller communities, or cities involves less than 10^6, however 10^6 is a common choice because countries number their populations

in multiples of 10^6. With these notations we translate (1) into

$$\begin{cases} \frac{dS}{dt} = -\beta SI \\ \frac{dI}{dt} = -\beta SI - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \tag{2}$$

where $\beta = \bar{\beta}/N$ and γ is the same as in (1).

Notice that now we actually have that $S(t) + I(t) + R(t) = S_0 + I_0 + R_0 = 1$ for all $t \geq 0$. Since we are interested in the reverse problem, namely determining the parameters β, γ from the observations, we put this as a formal mathematical result as follows.

Proposition 1. *Referring to the system (2), if we know I_0, S_0 and the values $I(t_1), S(t_1)$ for some $t_1 > 0$, these determine uniquely the parameters β and γ of the system.*

Notice there the main assumption, that the parameters β, γ do not change in time.

Our next goal is to get estimates on the parameters β, γ of the SIR model. There are two basic ideas here. The first one is to train a neural network using a typical inverse problem. The second one is to use this neural network combined with the data to estimate the regimes of the parameters. In a real world the parameters do not stay constant, they change slightly and we would like to catch part of this behavior. We combine the neural network with this day by day estimate to get an indication of the regimes for β and γ. In fact, what we do is we try to detect the regimes where the parameters stay more or less constant. As we will see the regime change is confirmed by the quarantine imposed as a fighting measure against the virus.

The neural network

To deal with the parameter estimates, we do the following. First we discretize β by considering 200 points equally spaced in the interval $[0.1; 1.5]$ and for γ we consider 100 points equally spaced in the interval $[0.05; 0.67]$. These intervals were chosen based on apriori analysis and much experimentation. Next, we solve the system of differential equations for each pair (β_i, γ_j), for 50 days, for a population of 10^6 individuals, and we store the results in a dataset.

We train a neural network on the resulting dataset so that the input is of the form:

$$XTrain = (Day,\ \ \#Susceptible,\ \ \#Infected,\ \ \#Recovered)$$

or in the terminology of the previous paragraph, we have

$$XTrain = (t, S(t), I(t), R(t)) \text{ where } t = 0, 1, \ldots, 50.$$

and the output is exactly the pair

$$YTrain = (\beta, \gamma),$$

which generated the solution above. We fixed the initial conditions $S_0 = 1 - I_0$, $I_0 = 2/N$ and $R_0 = 0$. We started with 2 infected people because there were two initial individuals who traveled in outside the country in exposed regions and were first spotted as the original spreaders.

The day by day fit of the parameters

Before we move on with the results of the day by day estimates, we point out that the result of Proposition 1 guarantees that the parameters estimated should be well-determined by the network as each triple $(t, S(t), I(t))$ determines in a unique way the parameters (β, γ).

Once the model had been trained, we use it to predict the day by day β and γ for Romania Covid-19 reported numbers. What this means is that we try to predict a set of parameters such that for a given day t, what we observe is exactly the number of suspected, infected and recovered reported on that day by the officials. Therefore, we assume and try to predict a single set of parameters for the time period $[0, t]$, t here being the corresponding day. The results are represented in the chart below.

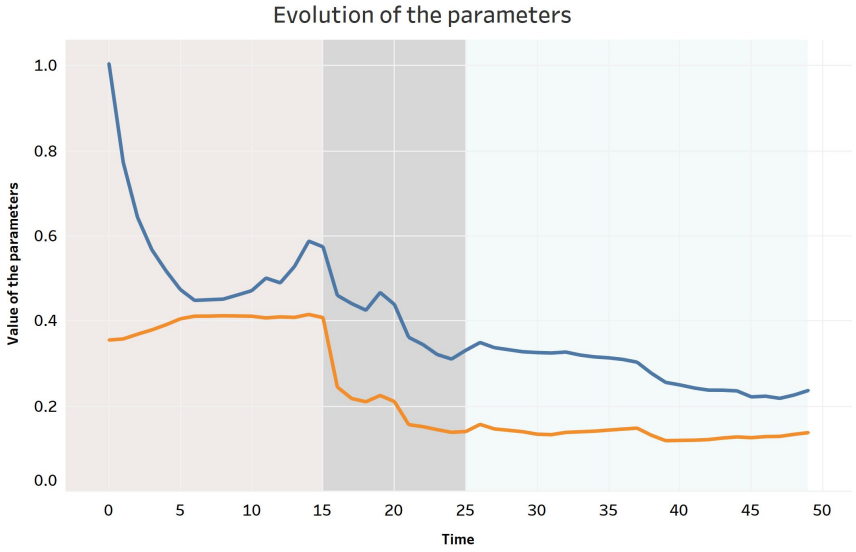


Figure 1: The prediction of day by day neural network trained on SIR models.

What this suggests is that we can identify three regimes. The first one is characterized by uncertainty, with a high infection spread and big variation of the two parameters values from day to day. This is approximately for the first 15 days or so. This may be due to the fact that, even if there were not many cases reported yet and the restrictions have not already been imposed, people were starting to be aware of the severity of the situation. On the other hand, the last 25 days have a lower volatility for both parameters, which can be a consequence of the measures

taken by the authorities. The intermediate regime can be considered a transition between the first one and the last one. This is roughly centered on the 30th day with a period of ± 10 days of regime switching.

We should also comment on the fact that the data that is available shows the number of individuals that have been tested positive, but it is very likely that the real number of people infected is in fact much higher, as there are also asymptomatic individuals, people that are not being tested although they present the specific symptoms, so they are not part of the official reports. Another aspect that should be taken into consideration is that the long incubation period characteristic to this virus determines a delay between the moment when a person has been infected and the moment when that person has been tested positive.

3 Conclusions

Now we summarize what we did here. The main idea is that within the models we used we split the problem according to various regimes. In this paper we take three regimes. One is the regime before any measures were taken. The second regime is the one in which the quarantine was imposed on the population. We also model the transition from one regime to another. The third regime we consider is the one following the relaxation. The transition is also modeled with the help of logistic function.

The fit is done using the number of deaths. The search of the parameters is done around the values of β provided by the neural network constructed based on the simpler SIR model.

We believe that this methodology is a general one and can be extended to any country provided that we have data, in particular some information about the regime switch for each of the regimes.

As a disclaimer, there are several assumptions made here. One of them is that people build up immunity to this virus and the reinfections are negligible.

Bibliography

[1] Tridip Sardar, Sk Shahid Nadim, and Joydev Chattopadhyay, *Assessment of 21 days lockdown effect in some states and overall india: a predictive mathematical study on covid-19 outbreak*, arXiv preprint arXiv:2004.03487 (2020).

[2] Janik Schüttler, Reinhard Schlickeiser, Frank Schlickeiser, and Martin Kröger, *Covid-19 predictions using a gauss model, based on data from april 2*, Physics **2** (2020), no. 2, 197–212.

[3] Calvin Tsay, Fernando Lejarza, Mark A Stadtherr, and Michael Baldea, *Modeling, state estimation, and optimal control for the us covid-19 outbreak*, arXiv preprint arXiv:2004.06291 (2020).

[4] Jon Wakefield, Tracy Qi Dong, and Vladimir N Minin, *Spatio-temporal analysis of surveillance data*, Handbook of Infectious Disease Data Analysis (2019), 455–476.

[5] Nicholas C Grassly and Christophe Fraser, *Mathematical models of infectious disease transmission*, Nature Reviews Microbiology **6** (2008), no. 6, 477–487.

An analogue of the Feynman-Kac formula for higher order evolution equations

Mariia Platonova[1]*

[1]*St. Petersburg Department of V.A. Steklov Institute; Chebyshev Laboratory, St. Petersburg State University, St. Petersburg, Russian Federation*

Abstract: We construct a probabilistic approximation of the Cauchy problem solution for higher order evolution equation. The approximating operators take the form of expectations of functionals of a certain random point field. This approximation can be considered as a generalization of the Feynman-Kac formula for the case of a higher order differential equation.

Keywords: Evolution equations, Poisson random measures, Feynman-Kac formula.

AMS subject classification: 28C20, 35K25, 60G55.

1 Introduction

It is well known that a solution to the Cauchy problem for the heat equation

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2} + V(x)u,\ u(0,x) = \varphi(x)$$

can be represented as the expectation of a functional of a Wiener process. Namely,

$$u(t,x) = \mathbf{E}\,\varphi(x - w(t)) \exp\Big(\int\limits_0^t V(x - w(s))ds\Big), \tag{1}$$

where $w(t)$ is a standard Wiener process. The formula (1) is called the Feynman-Kac formula (see [9], p. 308).

If a differential operator in an evolution equation is of order $2m > 2$ and has the form

$$\frac{\partial u}{\partial t} = \frac{(-1)^{m+1}}{(2m)!}\frac{\partial^{2m} u}{\partial x^{2m}} + V(x)u, \tag{2}$$

then any representation of the Cauchy problem solution analogous to (1) with $w(t)$ replaced by some random process is impossible, since the fundamental solution of (2)

*Corresponding author: mariyaplat@gmail.com

is not a probability measure. However, some analogous representations were found in papers ([1], [2], [5], [6], [7]). There are two main approaches to constructing an analogue of formula (1). One of them is based on using a pseudo-process instead of the Wiener process. The other one is based on the use of complex-valued process.

Finally, in [3], a probabilistic method was proposed for constructing an approximation of the Cauchy problem solution for the Schrödinger equation with a real bounded potential by mean values of functionals of stochastic processes. The approximating operators took the form of the expectations of functionals of a certain random point field. Using ideas and methods of [3], we are constructing an approximation of the Cauchy problem solution for the evolution equation (2).

A probabilistic approximation of the solution in the case $V = 0$ was constructed in [8]. The methods used in [8] were different for $m = 2k+1$ and $m = 2k$. In the first case only real-valued processes were used, while in the second case, complex-valued processes were used. In this paper we also have two different cases.

2 The case $m = 2k + 1$

Let $\nu(dt, dx)$ be a Poisson random measure on $(0, \infty)\times(0, \infty)$ with intensity measure $\frac{dt\,dx}{x^{1+2m}}$. For $\varepsilon > 0$ we denote a random process

$$\xi_\varepsilon(t) = \int_0^t \int_\varepsilon^{e\varepsilon} x\,\nu(ds, dx), \tag{3}$$

where e is the base of the natural logarithm. Note that for each $\varepsilon > 0$ the process $\xi_\varepsilon(t)$ is a compound Poisson process.

For any positive $\varepsilon > 0$ and $t > 0$ we define an operator $R^t_\varepsilon : L_2(\mathbf{R}) \to L_2(\mathbf{R})$, by setting for $\varphi \in L_2(\mathbf{R})$

$$R^t_\varepsilon\varphi(y) = \varphi * \omega^t_\varepsilon(-y) = \frac{1}{2\pi}\int_{\mathbf{R}} e^{ipy}\,\widehat{\varphi}(p)\,\widehat{\omega}^t_\varepsilon(p)\,dp,$$

where $\omega^t_\varepsilon(x)$ is defined by its Fourier transform

$$\widehat{\omega}^t_\varepsilon(p) = \exp\Big(-t\int_\varepsilon^{e\varepsilon}\big(ipy + \frac{(ipy)^2}{2!} + \ldots + \frac{(ipy)^{2m-1}}{(2m-1)!}\big)\frac{dy}{y^{1+2m}}\Big).$$

Throughout the paper, T_a denotes the shift operator: $T_a\varphi(x) = \varphi(x + a)$.

Next, we construct an analogue of the Feynman-Kac formula for the operator $A_\varepsilon + V$, where A_ε is a pseudo-differential operator with symbol

$$g_\varepsilon(p) = \int_\varepsilon^{e\varepsilon}\Big(e^{ipy} - 1 - \ldots - \frac{(ipy)^{2m-1}}{(2m-1)!}\Big)\frac{dy}{y^{1+2m}}.$$

It is easy to check that the symbol $g_\varepsilon(p)$ approximates the function $-\frac{p^{2m}}{(2m)!}$ as $\varepsilon \to 0$. It means that the operator A_ε approximates the operator $\frac{1}{(2m)!}\frac{d^{2m}}{dx^{2m}}$ as $\varepsilon \to 0$. In

this case the analogue of the Feynman-Kac formula cannot be written in terms of elementary functions, therefore it is defined through an integral representation, as was done in [3]. For this we should build a functional from the trajectory of the process under the sign of expectation. This functional is constructed as an integral over the positive semi-axis with respect to a Poisson random measure with Lebesgue intensity.

Let us denote $U(x) = V(x) + 1$ and define a family of operators

$$N_k(\tau_1, \tau_2, \dots, \tau_{k+1}, y_1, \dots, y_{k+1}),\ k = 0, 1, \dots,$$

by setting

$$\Big[N_k(\tau_1, \tau_2, \dots, \tau_{k+1}, y_1, \dots, y_{k+1})\varphi\Big](x) \\ = R^{\tau_1}_{\varepsilon, y_1} T^{y_1}_x U(y_1) R^{\tau_2}_{\varepsilon, y_2} T^{y_2}_{y_1} U(y_2) \dots T^{y_{k-1}}_{y_k} U(y_k) R^{\tau_{k+1}}_{\varepsilon, y_{k+1}} T^{y_{k+1}}_{y_k} \varphi(y_{k+1}).$$

Let $\mathcal{X} = \mathcal{X}(\mathbf{R}_+)$ be the space of configurations on $\mathbf{R}_+$. Each point X of the space $\mathcal{X}$ is a strictly increasing locally finite sequence of positive numbers $0 < t_1 < \dots < t_n < \dots$. Further, let $\mathbf{P}_0$ be a Poisson measure on $\mathcal{X}$ whose intensity measure is the Lebesgue measure (see [4]).

Let $\gamma(\cdot)$ be a locally bounded measurable function on $[0, \infty)$. For every s, t, such that $0 \leqslant s \leqslant t$, we define an operator $H_{s,t}(\gamma, X)$, acting on a function φ as

$$\Big[H_{s,t}(\gamma, X)\varphi\Big](x) = \Big[N_k\big(t_{l+1} - s, t_{l+2} - t_{l+1}, \dots, t_{l+k} - t_{l+k-1}, t - t_{l+k}, \\ m(s, t_{l+1}), m(t_{l+1}, t_{l+2}), \dots, m(t_{l+k-1}, t_{l+k}), m(t_{l+k}, t)\big)\varphi\Big](x), \quad (4)$$

where $m(t, s) = \gamma(s) - \gamma(t)$, $\quad l = \operatorname{card}\big(X \cap (0, s)\big)$, $\quad k = \operatorname{card}\big(X \cap (s, t)\big)$.

Now, we define a new family of operators $\Phi_{s,t}(\gamma)$, depending on γ and ε by setting

$$\Phi_{s,t}(\gamma) = \int_{\mathcal{X}} \mathbf{P}_0(dX) H_{s,t}(\gamma, X) = \int_{\mathcal{X} \cap [s,t]} \mathbf{P}_0(dX) H_{s,t}(\gamma, X).$$

Next, we define a semigroup Q^t_ε setting

$$\Big[Q^t_\varepsilon\, \varphi\Big](x) = \mathbf{E}\Big[\Phi_{0,t}\big(\xi_\varepsilon(t)\big)\varphi\Big](x) \quad (5)$$

for $\varphi \in L_2(\mathbf{R})$ and a semigroup Q^t setting

$$Q^t = \exp\Big(t\big(\frac{(-1)^{m+1}}{(2m)!}\frac{d^{2m}}{dx^{2m}} + V\big)\Big).$$

By definition, the operator Q^t takes the initial function φ into a solution of the Cauchy problem for the evolution equation (2). We additionally assume that the potential V has $2m + 1$ bounded derivatives and set

$$L = \max(\|V\|_\infty, \|V^{(1)}\|_\infty, \dots, \|V^{(2m+1)}\|_\infty).$$

Theorem 1. *Let $V \in \mathbf{C}^{(2m+1)}(\mathbf{R})$. Then there exists a positive constant C such that for every $\varphi \in W_2^{2m+1}(\mathbf{R})$ and $t \geqslant 0$*

$$\|Q_\varepsilon^t \varphi - Q^t \varphi\|_{L_2} \leqslant C\varepsilon\, t\, e^{t\|V\|_\infty} \Big(1 + \frac{t^{2m+1} L^{2m+1}}{2m+2}\Big) \|\varphi\|_{W_2^{2m+1}}.$$

3 The case $m = 2k$

In this case we use complex-valued processes $\sigma\xi_\varepsilon(t)$ where σ is a complex constant and $\xi_\varepsilon(t)$ is defined by (3). Take two complex numbers $\sigma_+ = \exp(\frac{i\pi}{2m})$ and $\sigma_- = \exp(-\frac{i\pi}{2m})$. Let us represent the initial function φ as

$$\varphi(x) = P_+\varphi(x) + P_-\varphi(x) = \varphi_+(x) + \varphi_-(x),$$

where P_+, P_- are Riesz projectors, defined on $L_2(\mathbf{R}) \cap L_1(\mathbf{R})$ by

$$P_+\varphi(x) = \frac{1}{2\pi} \int\limits_{-\infty}^{0} e^{-ipx}\, \widehat{\varphi}(p)\, dp, \quad P_-\varphi(x) = \frac{1}{2\pi} \int\limits_{0}^{\infty} e^{-ipx}\, \widehat{\varphi}(p)\, dp.$$

For any positive $\varepsilon > 0$ and $t > 0$ we define an operator $R_\varepsilon^t : L_2(\mathbf{R}) \to L_2(\mathbf{R})$, by setting for $\varphi \in L_2(\mathbf{R})$

$$R_\varepsilon^t \varphi(y) = \varphi * \omega_\varepsilon^t(-y) = \frac{1}{2\pi} \int_{\mathbf{R}} e^{ipy}\, \widehat{\varphi}(p)\, \widehat{\omega}_\varepsilon^t(p)\, dp,$$

where $\omega_\varepsilon^t(x)$ is defined by its Fourier transform

$$\widehat{\omega}_\varepsilon^t(p) = \begin{cases} \exp\Big(-t \int\limits_{\varepsilon}^{e\varepsilon} \Big(\sum\limits_{j=1}^{2m-1} \frac{(ip\sigma_+ y)^j}{j!} \Big) \frac{dy}{y^{2m+1}} \Big),\ p \geqslant 0, \\ \exp\Big(-t \int\limits_{\varepsilon}^{e\varepsilon} \Big(\sum\limits_{j=1}^{2m-1} \frac{(ip\sigma_- y)^j}{j!} \Big) \frac{dy}{y^{2m+1}} \Big),\ p < 0. \end{cases}$$

Let S be an operator acting on a function $\varphi \in L_2(\mathbf{R})$ as

$$S\varphi(x) = \frac{1}{2\pi} \int\limits_{-\infty}^{0} e^{ip\sigma_- x}\, \widehat{\varphi}(p)\, dp + \frac{1}{2\pi} \int\limits_{0}^{\infty} e^{ip\sigma_+ x}\, \widehat{\varphi}(p)\, dp = \varphi_+(-\sigma_- x) + \varphi_-(-\sigma_+ x).$$

As before, we define an operator $H_{s,t}(\gamma, X)$ by (4) using a new family of operators $\widetilde{N}_k$ instead of N_k, where

$$\Big[\widetilde{N}_k(\tau_1, \tau_2, \dots, \tau_{k+1}, y_1, \dots, y_{k+1})\varphi\Big](x) \\ = R_{\varepsilon, y_1}^{\tau_1} T_x^{y_1} S_{y_1} U(y_1) R_{\varepsilon, y_2}^{\tau_2} T_{y_1}^{y_2} S_{y_2} U(y_2) \dots R_{\varepsilon, y_{k+1}}^{\tau_{k+1}} T_{y_k}^{y_{k+1}} S_{y_{k+1}} \varphi(y_{k+1}).$$

Then we define a semigroup Q_ε^t by (5). The following statement holds.

Theorem 2. *Let $V \in \mathbf{C}^{(2m+1)}(\mathbf{R})$. Then there exists a positive constant C such that for every $\varphi \in W_2^{2m+1}(\mathbf{R})$ and $t \geqslant 0$*

$$\|Q_\varepsilon^t \varphi - Q^t \varphi\|_{L_2} \leqslant C\varepsilon\, t\, e^{t\|V\|_\infty} \Big(1 + \frac{t^{2m+1} L^{2m+1}}{2m+2}\Big) \|\varphi\|_{W_2^{2m+1}}.$$

The main result of the present paper follows from theorems 1, 2 and the Banach–Steinhaus theorem.

Theorem 3. *Let V be a bounded real potential. Then for any function $\varphi \in L_2(\mathbf{R})$*

$$\lim_{\varepsilon \to 0} \|Q_\varepsilon^t \varphi - Q^t \varphi\|_{L_2} = 0.$$

Acknowledgements: This research is supported by the Russian Science Foundation grant 19-71-30002.

Bibliography

[1] L. Beghin, K. J. Hochberg and E. Orsingher. Conditional maximal distributions of processes related to higher–order heat–type equations. *Stochastic Processes and their Applications*, 85(2):209–223, 2000.

[2] T. Funaki. Probabilistic construction of the solution of some higher order parabolic differential equation. *Proceedings of the Japan Academy, Series A, Mathematical Sciences*, 55(5):176–179, 1979.

[3] I. A. Ibragimov, N. V. Smorodina and M. M. Faddeev. Probabilistic Approximation of the Evolution Operator. *Functional Analysis and Its Applications*, 52(2):101–112, 2018.

[4] J. F. C. Kingman. *Poisson Processes*. Clarendon Press, Oxford, 1992.

[5] A. Lachal. Distributions of sojourn time, maximum and minimum for pseudo-processes governed by higher-order heat-type equations. *Electronic Journal of Probability*, 8:1–53, 2003.

[6] S. Mazzucchi. Probabilistic representations for the solution of higher order differential equations. *International Journal of Partial Differential Equations*, Article ID 297857, 2013.

[7] E. Orsingher and B. Toaldo. Pseudoprocesses related to space–fractional higher–order heat–type equations. *Stochastic Analysis and Applications*, 32(4):619–641, 2014.

[8] M. V. Platonova. A probabilistic representation of the Cauchy problem solution for an evolution equation with the differential operator of the order greater than 2. *Journal of Mathematical Sciences*, 229(6):744–755, 2018.

[9] M. C. Reed and B. Simon. *Methods of Modern Mathematical Physics, vol. 2.* Academic, New York, 1975.

The limit order book model

Dragana Radojičić[1]*

[1]*Mathematical Institute of the Serbian Academy of Sciences and Arts Belgrade, Serbia*

Abstract: This paper studies a stochastic model of a limit order book in discrete time and space, driven by a simple symmetric random walk. Especially, the focus is on the key quantity, namely the avalanche length, which is defined as a period of a sequence of trade executions, such that there is no period longer than ε without trading.

Keywords: Limit Order Book, stochastic modeling, generating function
AMS subject classification: 60C05, 62P05, 62P20.

1 Introduction

The limit order book

The main object of this research, namely the Limit Order Book (LOB), records all limit orders, both the orders awaiting to be sold (placed on the ask side) and orders awaiting to be bought (placed on the bid side), see Figure 1. The LOB is defined on a discrete price grid such that each point of the price grid represents a price level. For each available price, the corresponding volume is defined as the number of orders awaiting execution at that price. The best ask price is the price of the lowest limit sell order, and the best bid price is the price of the highest limit buy order. The quoted spread is the gap between the best bid and the best ask price, and it is always positive. The mid-price is defined as the arithmetic average of the best bid and best ask price.

The model

We are given a probability space $(\Omega, \mathcal{F}, P)$. Let us define X and Y as the following two independent processes: the 'driver' process $X = (X_n)_{n\geq 0}$ which denotes the simple symmetric random walk, and the 'order times' process $Y = (Y_n)_{n\geq 0}$, where Y_n models the time of the n-th jump of the simple symmetric random walk. Note that the process $Y = (Y_n)_{n\geq 0}$ takes values in $\mathbb{Z}_+$ and it is defined as $Y_0 = 0$ and $Y_n = \sum_{i=1}^{n} \Delta Y_i$, such that the increments $\Delta Y_n = Y_n - Y_{n-1}$ are randomly distributed for every n.

*Corresponding author: gagaradojicic@gmail.com

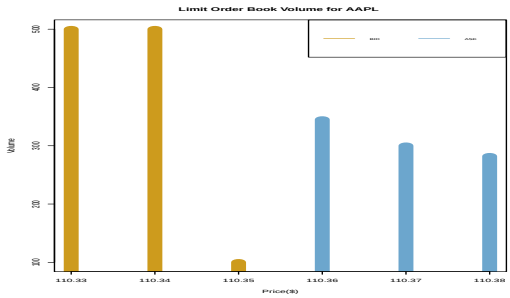


Figure 1: Snapshot of the NASDAQ limit order book for AAPL stock symbol for 3 levels at 10:02:36am (on the 1st December 2016).

Furthermore, define the order compound price process $S = (S_n)_{n\geq 0}$, which models the mid price at time $n \geq 0$, as $S_0 = 0$ and $S_n = X_{A_n-1}$, where $A = (A_n)_{n\geq 0}$ is the time-inverse process of $Y = (Y_n)_{n\geq 0}$, i.e.

$$A_n = \inf\{k \geq 0 : Y_k > n\}.$$

Let the two-parameter $\{V(n,U), n \in \mathbb{N}_0, U \in \mathbb{Z}\}$ be the order volume process, where $V(n,U)$ denotes the number of orders awaiting execution at price level U at time n. The order execution occurs when the mid-price reaches the price level at which there is some volume.

Since we ignore the size of the order, we distinguish only two cases: if there is an order at moment n at price level U, indicated by $V(n,U) > 0$ or if at that moment n there is no order at that price level, i.e. $V(n,U) = 0$. We assume that at the start point the LOB is full., i.e. for $u \in \mathbb{Z}$:

$$V(0,u) = 1_{u\geq 0}.$$

From now on, we focus on the ask-side, since similarly all the calculations can be derived for the bid-side. We fix an integer order placement parameter $\mu \geq 1$ and introduce the order book dynamics with the following equation:

$$V(Y_{n+1},U) = 1_{\{U\neq S_n\}}[V(Y_n,U) + 1_{\{U=S_n+\mu\}}].$$

This equation can be interpreted as follows: if $U = S_n$ at time Y_n and there is at least one order at that price level, indicated by $V(Y_n,U) > 0$, then the trade event occurs and all orders at that price level are removed from the LOB, the new value is consequently $V(Y_{n+1},U) = 0$. Furthermore, we assume that on the ask side at each step of the random walk a new order will be placed at the distance μ above the price S_n. Similarly, on the bid side a new order will be placed at the price level $S_n - \mu$. Note that the model introduced in this paper is enhanced version of the model established in [2].

2 The trading times and the avalanche length

We define the trading times $(\tau_i)_{i\geq 0}$ and the intertrading times $(T_i)_{i\geq 1}$:

- The trading times $\tau_0 = Y_0$, $\tau_i = \inf\{Y_n > \tau_{i-1} : V(Y_n, S_n) > 0\}$,
- The intertrading times $T_i = \tau_i - \tau_{i-1}$.

We consider two different execution mechanisms, namely Type I trades (next trade occurs after excursions to the next running maximum) and Type II trades (mid-price falls quickly by μ steps and then goes up by μ or more). Thus, the i-th trade, $i \geq 1$, is the Type I trade if $S(\tau_i) > S(\tau_{i-1})$, otherwise it is the Type II trade.

The avalanche length $A_{\mu,\varepsilon}$ is defined as:

$$A_{\mu,\varepsilon} = T_1 + \cdots + T_k,$$

where for $k \geq 1$ and $T_1 < \varepsilon, ..., T_k < \varepsilon, T_{k+1} \geq \varepsilon$, and $\varepsilon > 0$.

A path decomposition for the full avalanche length

Let us define the following four sets of random walk paths of length $n \geq 1$:

$$U_n = \{(z_0, ..., z_n) \in Z^{n+1} : z_0 = 0, |z_j - z_{j-1}| = 1 \text{ for } j = 1, ..., n\},$$

$$\mathscr{A}_{n,\mu} = \{z \in U_n : z_0 = 0, -\mu < z_k \leq 0 \text{ for } 0 \leq k \leq n-1 \text{ and } z_n = +1\},$$

$$\mathscr{B}_{n,\mu} = \{z \in U_n : z_0 = 0, -\mu < z_k \leq 0 \text{ for } 0 \leq k \leq n-1, z_{n-1} = 0 \text{ and } z_n = -1\},$$

$$\mathscr{C}_{n,\mu} = \{z \in U_n : z_0 = 0, \min(z_0, ..., z_{n-1}) = -\mu + 1 \text{ and } z_n = +1\}.$$

Moreover, define $\mathscr{A}_\mu$, $\mathscr{B}_\mu$ and $\mathscr{C}_\mu$ as:

$$\mathscr{A}_\mu = \bigcup_{n\geq 1} \mathscr{A}_{n,\mu} \qquad \mathscr{B}_\mu = \bigcup_{n\geq 1} \mathscr{B}_{n,\mu} \qquad \mathscr{C}_\mu = \bigcup_{n\geq 1} \mathscr{C}_{n,\mu},$$

and let $A_\mu(s,t)$, $B_\mu(s,t)$, $C_\mu(s,t)$ be the corresponding counting generating functions for the number of steps for the classes $\mathscr{A}_\mu, \mathscr{B}_\mu, \mathscr{C}_\mu$ respectively.

Calculating generating function for not simplified avalanche length

Motivated by the [1] [XIV.4, P.349], let $U_{z,n,m}$ be the probability that the mid-price process with initial position in $0 < z < \mu$, ends with n steps in m time units at the barrier 0. For $z > 0$ we define $U_z(s,t) = \sum_{n=0}^{\infty} \sum_{m=n}^{\infty} U_{z,n,m} s^n t^m$.

If the increments $\Delta Y_n = Y_n - Y_{n-1}$ are $(1-\theta)$ distributed for every n, where $\theta \in (0,1)$, we have:

$$U_{z,n+1,m} = \sum_{l=1}^{m-n} (\frac{1}{2} U_{z+1,n,m-l} + \frac{1}{2} U_{z-1,n,m-l}) \theta^l (1-\theta), \tag{1}$$

with boundary conditions

$$\begin{aligned} &U_{0,n,l} = U_{\mu,n,l} = 0, \qquad n \geq 1,\ l \geq 0, \\ &U_{0,0,0} = 1, \quad U_{0,0,l} = 0, \qquad l \geq 1, \\ &U_{z,0,l} = 0, \quad l \geq 0, \\ &U_{z,n,l} = 0, \qquad l < n. \end{aligned} \tag{2}$$

By shifting the summation index we obtain

$$\sum_{n=0}^{\infty}\sum_{m=n}^{\infty} U_{z,n,m}s^n t^m = \sum_{n=0}^{\infty}\sum_{m=n-1}^{\infty} U_{z,n+1,m}s^{n+1}t^m.$$

So we have:

$$\begin{aligned} U_z(s,t) &= \sum_{n=0}^{\infty}\sum_{m=n-1}^{\infty}\sum_{l=1}^{m-n}[\frac{1}{2}sU_{z+1,n,m-l}\theta^l(1-\theta)s^{n+1}t^m + \cdots] \\ &= \sum_{n=0}^{\infty}\sum_{l=1}^{\infty}\sum_{m=n}^{\infty}[\frac{1}{2}sU_{z+1,n,m}\theta^l(1-\theta)s^{n+1}t^m t^l + \cdots] \\ &= \frac{\theta t}{(1-\theta t)}(1-\theta)s(\frac{1}{2}U_{z+1}(s,t) + qU_{z-1}(s,t)) \end{aligned} \tag{3}$$

Following the procedure from [1] [XIV.4, P.349] and by subtracting $U_z(s,t) = \lambda(s,t)^z$ we get:

$$\begin{aligned} U_z(s,t) &= A(s,t)\lambda_1(s,t)^z + B(s,t)\lambda_2(s,t)^z \\ U_0(s,t) &= 1, \end{aligned} \tag{4}$$

where

$$\lambda_{1,2}(s,t) = \frac{1-\theta t \pm \sqrt{(1-\theta t)^2 - 4\theta^2(1-\theta)^2t^2s^2}}{\theta(1-\theta)ts}. \tag{5}$$

From the boundary conditions:

$$\begin{aligned} z = 0: &\qquad A(s,t) + B(s,t) = 1 \\ z = \mu: &\qquad A(s,t)\lambda_1(s,t)^\mu + B(s,t)\lambda_2(s,t)^\mu = 0, \end{aligned} \tag{6}$$

we have:

$$U_z(s,t) = \frac{\lambda_1(s,t)^\mu - \lambda_2(s,t)^\mu}{\lambda_1(s,t)^{\mu+1} - \lambda_2(s,t)^{\mu+1}}. \tag{7}$$

Lemma 1. *The probability generating functions for the classes $\mathscr{A}_\mu, \mathscr{B}_\mu, \mathscr{C}_\mu$ are given by:*

$$A_\mu(s,t) = \frac{\lambda_1(s,t)^\mu - \lambda_2(s,t)^\mu}{\lambda_1(s,t)^{\mu+1} - \lambda_2(s,t)^{\mu+1}}, \tag{8}$$

$$B_\mu(s,t) = A_\mu(s,t), \tag{9}$$

$$C_\mu(s,t) = A_\mu(s,t) - A_{\mu-1}(s,t). \tag{10}$$

Proof. Equations (8) and (9) follow from equation (7). Note that the class $\mathscr{C}_{n,\mu}$ consists of the paths of length n that hit the level $-\mu+1$, and thus the set $\mathscr{C}_{n,\mu}$ does not contain any element that belongs to the set $\mathscr{A}_{n,\mu-1}$. Further, the class $\mathscr{C}_{n,\mu}$ contains paths of length n that does not go below the level $-\mu$ and thus the class $\mathscr{C}_{n,\mu}$ contains all elements from the class $\mathscr{A}_{n,\mu}$ except the elements that belong to the class $\mathscr{A}_{n,\mu-1}$, i.e. $\mathscr{C}_{n,\mu} = \mathscr{A}_{n,\mu} \setminus \mathscr{A}_{n,\mu-1}$. Since $\mathscr{A}_{n,\mu-1} \subseteq \mathscr{A}_{n,\mu}$ and $\mathscr{C}_{n,\mu} = \mathscr{A}_{n,\mu} \setminus \mathscr{A}_{n,\mu-1}$ the equation (10) holds true.

□

Proposition 1. *The generating function of the time to the next trade is given by*

$$E[z^{T_1}] = A_\mu(z) + \frac{B_\mu(z)}{1 - B_\mu(z)} \cdot C_\mu(z). \tag{11}$$

Once the law of T_1 is found, we can compute the avalanche length distribution.

Theorem 1. *The full avalanche length for the symmetric random walk has probability generating function*

$$E[z^{L^*_{\mu\varepsilon}}] = \frac{P[T_1 > \varepsilon]}{E[1 - z^{T_1}; T_1 \le \varepsilon] + P[T_1 > \varepsilon]}, \tag{12}$$

where T_1 has the probability generating function given in (11) above.

Proof. Since $(T_i)_{i\ge1}$ are independent and identically distributed we have

$$\begin{aligned} E[z^{L^*_{\mu\varepsilon}}] &= \sum_{k\ge0} E[z^{T_1} z^{T_2} \cdots z^{T_k} : T_1 \le \varepsilon, T_2 \le \varepsilon, ..., T_k \le \varepsilon, T_{k+1} > \varepsilon] \\ &= \frac{P[T_{k+1} > \varepsilon]}{E[1 - z^{T_1} : T_1 \le \varepsilon] + P[T_1 > \varepsilon]} = \frac{P[T_1 > \varepsilon]}{E[1 - z^{T_1} : T_1 \le \varepsilon] + P[T_1 > \varepsilon]}. \end{aligned}$$

□

Acknowledgements: The author would like to thank F. Hubalek, whose work motivated the present paper.

Bibliography

[1] W. Feller. *An Introduction to Probability Theory and Its Applications.* John Wiley and Sons, New Jersey, 1968.

[2] Friedrich. Hubalek and D. Radojicic *On binomial order avalanches.* arXiv preprint arXiv:2007.07792, 2020.

Linear regression for uplift modeling

Krzysztof Rudaś,[1*] **and Szymon Jaroszewicz**[2]

[1]*Institute of Computer Science, Polish Academy of Sciences*
[2]*Faculty of Mathematics and Information Science, Warsaw University of Technology*

Abstract: Uplift modeling is an approach, which allows for predicting the effect of an action (e.g. a marketing campaign or a medical treatment) on a given individual. This contribution is a summary of [1] concerning uplift estimators for linear response (including a new method firstly mentioned in [1]) and their asymptotic properties.

Keywords: Uplift modeling, linear regression, heterogenous treatment effect
AMS subject classification: 62J05

1 Introduction

Machine learning models are frequently used to select targets for an action such as a medical treatment or a marketing campaign. A proper solution to this problem is *uplift modeling* which uses two training sets: a treatment set with objects subjected to an action, and a control set with objects left untreated. The goal is to model the *difference* in responses in both groups conditional on the predictors.

More formally, let x be a feature vector describing a customer, y^T the numerical outcome (e.g. purchase value) we would achieve after targeting the customer, and y^C the numerical outcome we would achieve had the customer not been targeted. Our purpose is to build a linear model of the form $x\beta^U$ which predicts the quantity $y^T - y^C$, called the *uplift*. The goal of this paper is to find as good an estimator of β^U as possible.

Performance of an estimator is typically measured by predictive mean squared error $\mathbb{E}\,\|X_{test}\hat{\beta^U} - X_{test}\beta^U\|^2$, where $X_{test} \in \mathbb{R}^{n_{test}\times p}$ is a matrix of predictors of observations which are not used for training. For unbiased estimators this error is determined solely by the estimator's covariance matrix [2] which should be as small as possible.

We also introduce a training matrix $X \in \mathbb{R}^{n\times p}$. Rows $x_{i\cdot}$ of X are generated independently from each other and follow the same distribution. We also define the matrices $X^T \in \mathbb{R}^{n^T\times p}$, $X^C \in \mathbb{R}^{n^C\times p}$ whose rows are rows from X assigned to treatment and control groups respectively; n^T and n^C denote, respectively, the

*Corresponding author: krzysztof.rudas@ipipan.waw.pl

number of cases in the treatment and control training sets with $n = n^T + n^C$. Denote by $q^T = \frac{n^T}{n}$ and $q^C = \frac{n^C}{n}$ the proportions of cases in both groups.

Likewise we define treatment and control response vectors $y^T \in \mathbb{R}^{n^T}$ and $y^C \in \mathbb{R}^{n^C}$ and denote by y the combined vector of all responses. Assumed relationships in the data can be written as

$$y^C = X^C\beta^C + \varepsilon^C, \tag{1}$$

$$y^T = X^T\beta^C + X^T\beta^U + \varepsilon^T = X^T\beta^T + \varepsilon^T, \tag{2}$$

where β^C are the true response coefficients in the untreated population, β^T is the respective treatment coefficient vector, and β^U is the vector of coefficients defining the strength and direction of the effect of the action on a given individual.

Random vectors ε^T and ε^C denote random components of the responses in the treatment and control groups. It is assumed that X, ε^T and ε^C are independent of each other. Moreover, we assume that $\mathbb{E}\,\varepsilon_i^T = \varepsilon_i^C = 0$, $\operatorname{Var}\varepsilon_i^T = (\sigma^T)^2$ and $\operatorname{Var}\varepsilon_i^C = (\sigma^C)^2$.

The assignment to the treatment and control group is assumed to be random, conditional on fixed values of n^C and n^T. This type of assignment is call *complete randomization* in [7].

2 The double estimator

A most obvious idea for estimating the unknown vector β^U is to estimate β^T and β^C using separate linear models and then subtract their coefficient vectors. This kind of approach will be referred to as the double estimator.

Definition 1. A vector $\hat{\beta}_d^U$ given by the formula:

$$\hat{\beta}_d^U = (X^{T\prime}X^T)^{-1}X^{T\prime}y^T - (X^{C\prime}X^C)^{-1}X^{C\prime}y^C \tag{3}$$

is called the *double estimator* of the parameter vector β^U.

Let us now give some results on the behaviour of the double estimator.

Theorem 1. *Assume that the predictor matrix X is random, $\mathbb{E}\,x'_{i\cdot} = 0$, and $\operatorname{Var}x'_{i\cdot} = \Sigma$. Assume further, that complete randomization was used. Then*

1. *$\hat{\beta}_d^U$ is unbiased, i.e. $\mathbb{E}\,\hat{\beta}_d^U = \beta^U$,*
2. *if, in addition, each row $x_{i\cdot}$ of the matrix X follows the normal distribution ${}_p(0,\Sigma)$, then $\operatorname{Var}\hat{\beta}_d^U = \left(\frac{(\sigma^T)^2}{n^T-p-1} + \frac{(\sigma^C)^2}{n^C-p-1}\right)\Sigma^{-1}$,*
3. *if $n \to \infty$ with the proportions q^T, q^C fixed, then*

$$\sqrt{n}\left(\hat{\beta}_d^U - \beta^U\right) \xrightarrow{d}_p \left(0, \left(\frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C}\right)\Sigma^{-1}\right).$$

The double estimator uses two separate models, each constructed on a different subset of data. Splitting the training data seems to make the variance of the estimator worse. The question therefore arises whether it is possible to construct a single estimator using the whole dataset.

3 The uplift estimator

The first idea for building a more direct uplift estimator is very simple. We simply reverse the sign of the response in the control group and build a regression model on concatenated data (with some weighting of cases).

Definition 2. Consider the following transformation $\tilde{y}$ of the response vector y:

$$\tilde{y_i} = \begin{cases} \frac{1}{q^T} y_i & \text{if } g_i = T, \\ -\frac{1}{q^C} y_i & \text{if } g_i = C. \end{cases} \tag{4}$$

A vector $\hat{\beta}_z^U$ given by the formula

$$\hat{\beta}_z^U = \left(X'X\right)^{-1} X'\tilde{y} \tag{5}$$

is called the *uplift estimator* of the parameter vector β^U.

Another way to look at this estimator is to rewrite it as:

$$\beta_z^U = \frac{1}{q^T}(X'X)^{-1}X^{T\prime}y^T - \frac{1}{q^C}(X'X)^{-1}X^{C\prime}y^C$$

which reveals that it is a modification of the double estimator where the matrices $(X^{T\prime}X^T)^{-1}$ and $(X^{C\prime}X^C)^{-1}$ are replaced with a single estimate $(X'X)^{-1}$ made on the full dataset. Notice that, due to randomization, the true underlying covariance matrices are identical in the treatment and control groups, so one can expect that this estimate will be better than both $(X^{T\prime}X^T)^{-1}$ and $(X^{C\prime}X^C)^{-1}$.

To verify whether the uplift estimator is really superior, let us first look at its performance in a very special case when $\beta^T = -\beta^C$.

Theorem 2. *Assume that $\beta^T = -\beta^C$, $q^T = q^C = \frac{1}{2}$, $\sigma^T = \sigma^C = \sigma$, and that complete randomization is used. Then*

1. *$\hat{\beta}_z^U$ is unbiased,*
2. *if, in addition each row $x_{i\cdot}$ of the matrix X follows the normal distribution $_p(0, \Sigma)$, then $\operatorname{Var}\hat{\beta}_z^U = 4\sigma^2\frac{\Sigma^{-1}}{n-p-1}$.*

Comparing with the variance of the double estimator:

$$\operatorname{Var}\hat{\beta}_z^U = 4\sigma^2\frac{\Sigma^{-1}}{n-p-1} < 4\sigma^2\frac{\Sigma^{-1}}{n-2p-2} = 2\sigma^2\frac{\Sigma^{-1}}{\frac{n}{2}-p-1} = \operatorname{Var}\hat{\beta}_d^U$$

we see that the uplift estimator has a lower variance and thus lower MSE. Now we introduce a theorem characterizing the behavior of the uplift estimator in the general case.

Theorem 3. *Assume that the predictor matrix X is random with $\mathbb{E}\, x'_{i.} = 0$, $\operatorname{Var} x'_{i.} = \Sigma$, and that complete randomization was used. Then*

1. *$\hat{\beta}_z^U$ is unbiased.*
2. *As $n \to \infty$ with the proportions q^T, q^C fixed*

$$\sqrt{n}(\hat{\beta}_z^U - \beta^U) \xrightarrow{d}_p \left(0, \Sigma^{-1}\left(\frac{(\sigma^T)^2}{q^T} + \frac{(\sigma^C)^2}{q^C}\right) + \Sigma^{-1} \operatorname{Var}(x'_{i.} x_{i.} b)\Sigma^{-1}\right),$$

where $b = \frac{q^C \beta^T + q^T \beta^C}{\sqrt{q^C q^T}} = \sqrt{\frac{q^C}{q^T}}\beta^T + \sqrt{\frac{q^T}{q^C}}\beta^C$.

After comparing the asymptotic variances of the double and uplift estimators, it becomes clear that the latter involves an additional nonnegative term $\operatorname{Var}(x'_{i.} x_{i.} b)$. Counterintuitively, fitting a single model gives an asymptotically worse estimator. Of course, the conclusions may not hold for smaller samples. Theorem 2 is a special case in which $b = 0$, when $\hat{\beta}_z^U$ is in fact better.

4 The corrected uplift regression estimator

In this section we introduce a new estimator (first appeared in [1]), which combines the benefits of the double estimator (good asymptotic behaviour) and the uplift estimator (better estimation of $(X'X)^{-1}$). Our new approach is based on uplift regression, but we modify the response using corrections to β^T and β^C. These corrected coefficients will reduce the additional term b in Theorem 3 to 0.

More formally, we introduce a vector β^* and use it to obtain corrected coefficients: $\beta^{T*} = \beta^T - \beta^*$ and $\beta^{C*} = \beta^C - \beta^*$. It's not difficult to see that:

$$\beta^{T*} - \beta^{C*} = \beta^T - \beta^* - \beta^C + \beta^* = \beta^U.$$

Pick $\beta^* = q^C \beta^T + q^T \beta^C$. When we replace β^T with β^{T*} and β^C with β^{C*} the vector b in Theorem 3 becomes:

$$b = \sqrt{\frac{q^C}{q^T}}\beta^{T*} - \sqrt{\frac{q^T}{q^C}}\beta^{C*} = \sqrt{\frac{q^C}{q^T}}(\beta^T - \beta^*) + \sqrt{\frac{q^T}{q^C}}(\beta^C - \beta^*) = 0,$$

so after this correction the additional term in the asymptotic variance of $\hat{\beta}_z^U$ vanishes.

Unfortunately we cannot compute β^* directly, because we don't know the exact values of β^T and β^C. To solve this problem we will estimate β^* from data. Define the vector y^* as:

$$y_i^* = \begin{cases} \frac{q^C}{q^T} y_i, & \text{if } g_i = T \\ \frac{q^T}{q^C} y_i, & \text{if } g_i = C \end{cases}$$

and apply the classical least squares estimator to it: $\hat{\beta}^* = (X'X)^{-1}X'y^*$. We cannot directly change the true coefficients β^T and β^C, so instead we will modify the response y by subtracting $X\hat{\beta}^*$ from it. As a result we obtain the following two-stage estimator:

Definition 3. A vector $\hat{\beta_c^U}$ given by the formula

$$\hat{\beta_c^U} = (X'X)^{-1}X'\widetilde{y_c},$$

where the $\widetilde{\ }$operator is given in Definition 2 and

$$y_c = y - X\hat{\beta}^*,$$

is called the *corrected uplift regression estimator* of the parameter vector β^U.

The following theorem shows that we do indeed obtain an improvement.

Theorem 4. *Assume that the predictor matrix X is random, $\mathbb{E}\, x'_{i.} = 0$, and $\operatorname{Var} x'_{i.} = \Sigma$. Assume further that complete randomization was used. Then*

1. *$\hat{\beta}^*$ is an unbiased estimator of β^*,*
2. *$\hat{\beta}_c^U$ is an asymptotically unbiased estimator of β^U,*
3. *if $n \to \infty$ with the proportions q^T, q^C fixed, then $\sqrt{n}(\hat{\beta_c^U} - \beta^U) \xrightarrow{d}_p \left(0, \left(\frac{\sigma^{T\,2}}{q^T} + \frac{\sigma^{C\,2}}{q^C}\right)\Sigma^{-1}\right)$.*

It can be seen that the corrected estimator has the same asymptotic distribution as the double regression estimator. Furthermore, both estimators $\hat{\beta}^*$ and $\hat{\beta}_C^U$ are computed based on the full dataset using better estimates of $(X'X)^{-1}$, as does the uplift regression estimator. We recommend the corrected uplift regression estimator as the right choice for uplift linear regression.

Bibliography

[1] K. Rudaś, S. Jaroszewicz, Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, 32(5):1275–1305, 2018.
[2] C. Heumann, T. Nittner, C.R. Rao, S. Scheid, H. Toutenburg, *Linear Models: Least Squares and Alternatives*. Springer, New York, 2013.
[3] J. Pearl. *Causality*. Cambridge University Press 2009.
[4] P. Rzepakowski, S. Jaroszewicz, Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 2011.
[5] Ł. Zaniewicz, S. Jaroszewicz, Support Vector Machines for Uplift Modeling. *The First IEEE ICDM Workshop on Causal Discovery (CD 2013)*, 2013.
[6] N. J. Radcliffe, P. D. Surry, *Real-World Uplift Modelling with Significance-Based Uplift Trees*. Portrait Technical Report TR-2011-1, Stochastic Solutions, 2011.
[7] G. W. Imbens, D. B. Rubin, *Causal Inference for Statistics, Social and Biomedical Sciences*. Cambridge University Press, 2015.

Generalized Gaussian model for EEG data

Z. Salinger,[1*] N.N. Leonenko[1], A. Sikorskii[2], N. Šuvak[3] and M.J. Boivin[4]

[1] *School of Mathematics, Cardiff University, Cardiff, United Kingdom*
[2] *Departments of Psychiatry and Statistics and Probability, Michigan State University, East Lansing, Michigan, USA*
[3] *Department of Mathematics, J.J. Strossmayer University of Osijek, Osijek, Croatia*
[4] *Departments of Psychiatry and Neurology and Ophtalmology, Michigan State University, East Lansing, Michigan, USA*

Abstract: Previous analysis of electroencephalogram (EEG) data from [9] showed that stochastic modelling of EEG features can improve the explanation of variation in neurodevelopmental and cognitive outcomes of children who were affected by cerebral malaria. In our analysis, EEG increments are viewed as discrete time observations from a diffusion process with marginal generalized Gaussian distribution (GGD). The GGD parametrization comprises both light-tailed and heavy-tailed distributions. Two versions of this model were fit to the data. In the first model, marginal distributions were from the light-tailed GGD subfamily. In the second model, marginal distributions were heavy-tailed (Student), and the tail index was estimated using the empirical scaling function. The estimated parameters from models across EEG channels were explored as potential predictors of neurocognitive outcomes of these children 6 months after recovering from illness. Some of these EEG parameters were shown to be important predictors of neurodevelopment and cognition.

Keywords: Generalized Gaussian distribution, Tail index, EEG modelling, Elastic net regression.

AMS subject classification: 62G07, 62J20, 62P10.

1 Introduction

Cerebral malaria is the most severe neurological complication of infection with a parasite *Plasmodium falciparum* where 90% of all cases occur in sub-Saharan Africa. Survivors sustain brain injury, which may affect subsequent neurodevelopment and cognitive functioning. Identification of factors that can predict the extent of neurocogntive impairment and other outcomes following cerebral malaria illness is an

*Corresponding author: salingerz@cardiff.ac.uk

important problem. Electroencephalogram (EEG) is used for monitoring electrical neural activity of the brain and signals are captured by multiple electrodes called channels located over the scalp. In this analysis we build upon the past work done in [9] and we investigate the underlying stochastic processes of EEG data. The analysis shows that the distributions of EEG increments are symmetrical and display both heavy and lighter tails. We provide a unifying stochastic model that captures different types of tails across the range of model parameters.

2 Distributional properties of EEG increments

Our approach in dealing with EEG signal is based on its transformation into EEG increments, which have a symmetric distribution with the maximum at zero. To reflect the diversity of the empirically observed distribution candidates, we parametrize the probability density function (PDF) of the distribution of increments as follows:

$$f_{s,b}(x) = \begin{cases} \dfrac{1}{2(s\sigma^2)^{1/s}\Gamma\left(1+\frac{1}{s}\right)} e^{-\frac{|x|^s}{s\sigma^2}} & , \ b = 0 \\ \dfrac{bs}{2\sigma^2}\left(\dfrac{s\sigma^2}{b}\right)^{-1/s} \dfrac{\Gamma\left(1+\frac{1}{s}+\frac{\sigma^2}{b}\right)}{\Gamma\left(\frac{1}{s}\right)\Gamma\left(\frac{\sigma^2}{b}\right)} \left(1+\dfrac{b}{s\sigma^2}|x|^s\right)^{-\frac{\sigma^2}{b}-\frac{1}{s}-1} & , \ b > 0, \end{cases} \tag{1}$$

where the value of parameter b is used as an indicator for making a distinction between light-tailed ($b = 0$) and heavy-tailed ($b > 0$) distributions within this family. We will refer to the distributions from this family as generalized Gaussian distribution (GGD). The subfamily characterized by $b = 0$ resembles the usual GGD parametrization including, for $s = 2$, the zero-mean normal distribution with variance σ^2. For more details on light-tailed GGD subfamily we refer to [8] and [4]. For $b > 0$, distributions in the GGD subfamily admit heavy tails, e.g., for $s = 2$ this distribution is of the Student type. For more information on Student distribution and related processes we refer to [6], whereas for parametrization similar to (1) we refer to [7].

In order to characterize important probabilistic properties of increments including their dependence structure, we view the EEG increments as discrete-time observations from the diffusion process $(X_t, t \geq 0)$ with the stationary PDF (1). Since the PDF (1) is continuous, bounded, and strictly positive on the whole $\mathbb{R}$, according to the Theorem 2.1, page 193 in [1] the stochastic differential equation (SDE)

$$dX_t = -\theta X_t\, dt + v(t)\, dB_t, \quad \theta > 0, \quad t \geq 0, \tag{2}$$

driven by the standard Brownian motion $(B_t, t \geq 0)$, admits the unique weak ergodic solution and defines the diffusion with stationary distribution (1), which we call the generalized Gaussian diffusion (GGDiff).

3 Parameter estimation

Estimation of light-tailed GGD parameters

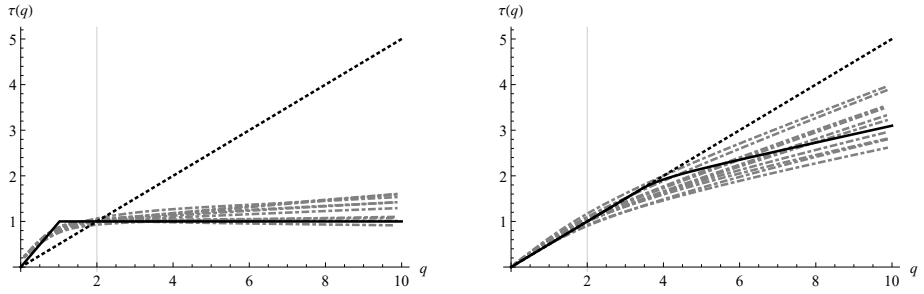
In the light-tailed case ($b = 0$), the two-dimensional parameter $\zeta = (s, \sigma^2)$ of the stationary distribution of the GGDiff $X = (X_t,\, t \geq 0)$ is estimated by the quasi-likelihood method. For the purpose of estimation of parameter ζ we disregard the existing exponentially decaying autocorrelation structure of the diffusion and define the quasi log-likelihood function as

$$l(\zeta) = \sum_{i}^{n} \ln \left(\frac{1}{2(s\sigma^2)^{1/s}\Gamma\left(1+\frac{1}{s}\right)} e^{-\frac{|X_i|^s}{s\sigma^2}} \right), \tag{3}$$

The estimate $\hat{\zeta} = (\hat{s}, \hat{\sigma^2})$ of the parameter $\zeta = (s, \sigma^2)$ is then obtained by maximising (3), which can be performed using existing non-linear optimization methods. For more details on maximum likelihood estimation for diffusion processes we refer to [2] and [3].

Tail index estimation

Recall that for $b > 0$ the distribution (1) is heavy-tailed. The tails of this distribution decay as $|x|^{-1-s\left(\frac{\sigma^2}{b}+1\right)}$, so the tail index is of the form $\alpha = s\left(\frac{\sigma^2}{b}+1\right)$. To estimate the tail index, we use the approach introduced in [5] based on the empirical scaling function. The shape of the scaling function is strongly influenced by the tail index and graphical inspection is used for estimation of the tail index of the corresponding distribution. An example of this can be seen in Figure 1.



(a) $\hat{\alpha} = 1.10$ (b) $\hat{\alpha} = 3.49$

Figure 1: Tail index estimates of EEG increments

4 Prediction of neurodevelopmental and cognitive scores

Measures included in the study

Data were collected during the observational study of the pathogenesis of severe malaria in surviving children and EEG signals were recorded during coma for the children who were diagnosed with cerebral malaria. Age-independent single measure of neurodevelopment and cognition (z-score) was obtained for children enrolled in the study. Other non-EEG data that were collected during the study included demographic, anthropometric and socioeconomic characteristics, along with biomarker panels from plasma and cerebrospinal fluid taken at the point of hospitalization.

Methods and models used

To identify important predictors of neurodevelopment and cognition 6 months after the coma from cerebral malaria, elastic net regression was used. The method was introduced in [10] as a way of controlling for correlations among predictors and dealing with the case where the number of predictors is much bigger than the number of observations, which was the case in our analysis. Additionally, elastic net regression was chosen in accordance with the previous analysis of the same data in [9]. The response variable was the standardized neurodevelopment or cognitive score taken 6 months after the discharge from the hospital. Predictor variables were taken from three sets of features. First feature matrix included just the non-EEG features. Second feature matrix included a combination of non-EEG features and EEG features obtained from fitting GGD to EEG increments (estimates of s and σ^2). Third feature matrix was a combination of non-EEG features and median values of tail index estimates α. To reduce the noise of this variable within the model, additional feature matrix was created by classifying tail index values based on distributional tertiles. In all three feature matrices, we included socio-demographic characteristics and the neurodevelopmental or cognitive score immediately after discharge from the hospital (baseline ND).

5 Results and discussion

Our results show that the baseline neurodevelopmental score (taken right after coma) was the most important predictor of neurodevelopment at point 6 months after coma which was expected as it is a direct measure of the outcome variable taken at a different time point. Other non-EEG features retained in our model generally overlap with the non-EEG features found to be important predictors in the analysis of [9] and mostly contain biomarker panels from cerebrospinal fluid and/or plasma. The addition of EEG features from fitting of GGD and estimation of tail index resulted in an improved RMSE for both light-tailed and heavy-tailed stochastic models, which can be seen in Table 1.

Table 1: Model comparison based on elastic net regression results

Model features included (number of features)	RMSE	Number of non-zero coefficients	Number of non-zero coefficients from EEG features subset
Non-EEG features (54)	0.5670	12	N/A
Non-EEG (54) and GGD (38) features	0.5655	13	1
Non-EEG (54) and continuous tail index features (19)	0.5670	12	0
Non-EEG (54) and categorical tail index features (38 dummy variables)	**0.5499**	**10**	1

In summary, the addition of stochastic EEG modelling improved the prediction of children's brain function 6 months following coma. Further improvement can be made by investigating other marginal distributions appropriate for modelling of EEG signal increments and extending the analysis to other infectious diseases that could affect the brain.

Acknowledgements: The studentship for Z. Salinger is funded through the UK Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Partnership (project reference 2275322).

Bibliography

[1] B. M. Bibby, I. M. Skovgaard, and M. Sørensen. Diffusion-type models with given marginal distribution and autocorrelation function. *Bernoulli*, 11(2):191–220, 2005.

[2] J. P. N. Bishwal. *Parameter Estimation in Stochastic Differential Equations*. Springer-Verlag, Berlin, Heidelberg, 2007.

[3] L. Chenxu. Maximum-likelihood estimation for diffusion processes via closed-form density expansions. *Ann. Stat.*, 41(3):1350–1380, 2013.

[4] A. Dytso, R. Bustin, H. V. Poor and S. Shamai. Analytical properties of generalized Gaussian distributions. *J. Stat. Distrib. Appl.*, 5(1), 2018.

[5] D. Grahovac, M. Jia, N. Leonenko, and E. Taufer. Asymptotic properties of the partition function and applications in tail index inference of heavy-tailed data. *Statistics (Ber).*, 49(6):1221–1242, 2015.

[6] B. Grigelionis. *Student's t-Distribution and Related Stochastic Processes*. Springer-Verlag, Berlin, Heidelberg, 2013.

[7] E. Lutwak, D. Yang, and G. Zhang. Moment-entropy inequalities. *Ann. Probab.*, 32(1B):757–774, 2004.

[8] E. Nadarajah. A generalized normal distribution. *J. Appl. Stat.*, 32(7):685–694, 2005.

[9] M. A. Veretennikova, A. Sikorskii, and M. J. Boivin. Parameters of stochastic

models for electroencephalogram data as biomarkers for child's neurodevelopment after cerebral malaria. *J. Stat. Distrib. Appl.*, 5(1), 2018.

[10] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, 67(2):301–320, 2005.

Study of neural networks to predict survival in oncology

Mathilde Sautreuil,[1*] **Sarah Lemler**[1] **and Paul-Henry Cournède**[1]

[1]*Laboratory of Mathematics and Informatics (MICS), CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France*

Abstract: We are interested in studying neural networks' potential to predict survival in oncology. In a clinical study in oncology, the number of variables to characterize patients can be huge with, for example, clinical, genomic, and radiology image data. In contrast, the number of patients in cohorts remains relatively small. We are in what is called a high-dimensional framework, which is when the number of variables is much larger than the sample size. A classical model used to deal with survival data is the Cox model. Some regularization procedures have been proposed to deal with this model in high-dimension. However, they prove limited when the dimension becomes too large. Moreover, the Cox model's hypothesis is not always satisfied. Neural networks have provided state-of-the-art models in a lot of research domains. We have explored their potential in survival analysis, especially in high-dimension, and we considered two strategies of neural networks. The first one is based on the Cox model: the neural network replaces the linear dependency in the covariates to determine the Cox hazard function. A second and less studied approach based on a discrete-time model predicts the discretized hazard function directly. We have adapted this method to the high-dimensional setting. This paper focuses on the last neural network and compares it to the neural network based on the Cox model. We also consider a Lasso procedure applied to the Cox partial log-likelihood as the benchmark. We applied them to one real dataset in oncology, and the neural network based on a discrete-time model gets the best performance.

Keywords: Neural networks, survival analysis, high-dimension, Cox model
AMS subject classification: 62N02

1 Introduction

Survival analysis consists of studying the elapsed time until an event of interest, such as the death or recovery of a patient in medical studies. This paper aims to compare methods to predict a patient's survival from clinical and gene expression data.

*Corresponding author: mathilde.sautreuil@gmail.com

The Cox model [4] is the reference model in the field of survival analysis. It relates the survival duration of an individual to the set of explanatory covariates. With high-throughput sequencing techniques, transcriptomics data are more and more often used as covariates in survival analysis. Adding these covariates raise issues of high-dimensional statistics, when we have more covariates than individuals in the sample. Methods based on regularization or screening [6, 13] have been developed and used to solve this issue. Another particularity of the Cox model is that it relies on the proportional hazard hypothesis; and in its classical version, it does not account for nonlinear effects or interactions, which is limited in some real situations. Therefore, in this paper, we focus on another type of methods: neural networks. Deep learning methods are more and more popular, notably due to their flexibility and their ability to handle interactions and nonlinear effects, including in the biomedical field. In this paper, we focus on a neural network based on a discrete-time survival model, as introduced by Biganzoli *et al.* [2]. Biganzoli *et al.* [2] have studied this neural network only in low-dimension. Our objective is to study and adapt this model to the high-dimensional cases and compare its performances to two other methods: the two-step procedure to estimate the entire risk in a Cox model, with the classical estimation of the regression parameters with a Lasso penalty and a kernel estimator of the baseline function (as in [8]), noted Cox-L1 in the paper, and the Cox-nnet neural network [3] based on the partial likelihood of the Cox model. Section 2 presents the different models compared to predict patients' survival in high-dimension. Section 3 shows the results of the studied methods on a real dataset. We also did a simulation study that we do not present in this document, but we detailed its design and results in Sautreuil *et al.* [12]. Finally, we underline the results to conclude with the potential of neural networks in survival analysis.

2 Models

Cox-nnet

In 1995, Faraggi *et al.* [7] developed a neural network based on the proportional hazards model. The idea of Faraggi *et al.* [7] was to replace the linear prediction of the Cox regression with the neural network's hidden layer's output. Faraggi *et al.* [7] only applied their neural network to survival analysis from clinical data, in low dimension. More recently, some authors revisited this method [3, 9, 10]. We will use Cox-nnet [3], which was already applied in a high-dimensional setting, as a benchmark in our study.

Discrete-time neural network

Biganzoli *et al.* [2] have proposed a neural network based on a discrete-time model. They introduced L time intervals $A_l = (t_{l-1}, t_l]$, and build a model predicting in which interval, the failure event occurs. We write the discrete hazard as the conditional probability of survival:

$$h_{il} = P(Y_i \in A_l | Y_i > t_{l-1}), \tag{1}$$

with Y_i the survival time of individual i. Biganzoli *et al.* [2] duplicates the individuals as input of the neural network. Indeed, the Biganzoli *et al.* [2]'s neural network takes as input the set of variables of the individual and an additional variable corresponding to the mid-point of each interval. Due to the addition of this variable, the p variables of each individual are repeated for each time interval. The output is thus the estimated hazard $h_{il} = h_l(X_i, a_l)$ for the individual i at time a_l. The duplication of individuals gives to that neuron network a more original structure than that of a classical multi-layer perceptron. Biganzoli *et al.* [2] initially used a 3-layer neural network with a logistic function as the activation function for both the hidden and output layers. The output of the neural network with H neurons in the hidden layer and $p+1$ input variables is given by:

$$h_{il} = h(x_i, t_l) = f_2\left(a + \beta^T f_1\left(b + W^T X_{i.}\right)\right),$$

where $W = (w_{dh})_{1\leq d\leq p+1, 1\leq h\leq H}$, and $\beta = (\beta_1, \ldots, \beta_H)^T$ are the weights of the neural network, a and b are the biases of the neural network to be estimated, and f_1 and f_2 the sigmoid activation functions. The target of this neural network is the death indicator d_{il}, which indicates if the individual i dies in the interval A_l. We introduce $l_i \leq L$ the number of intervals in which individual i is observed, $d_{i0}, \ldots, d_{i(l_i-1)} = 0$ whatever the status of the individual i and d_{il_i} is equal to 0 if the individual i is censored and 1 otherwise. The cost function used by Biganzoli *et al.* [2] is the cross-entropy function and the weights of the neural network can be estimated by minimizing it. Biganzoli *et al.* [2] added a ridge penalty to their cross-entropy function:

$$\mathcal{L}(V) = -\sum_{i=1}^{n}\sum_{l=1}^{l_i} d_{il}\log(h_{il}) + (1-d_{il})\log(1-h_{il}) + \lambda\|V\|_2, \tag{2}$$

with λ the hyperparameter and $V = (W^T, \beta^T, a^T, b^T)$ the weights and biases of the neural network. In Biganzoli *et al.* [2], λ was chosen by deriving an Information Criterion. We choose instead to use cross-validation since it improves model the predictive capacity.

After estimating the parameters of the neural network by minimizing the loss function (2), the output obtained is the estimate of the discrete risk $\widehat{h}_{il}$ for each individual i and the survival function of individual i is estimated using: $\widehat{S}(T_{l_i}) = \prod_{i=1}^{l_i}(1-\widehat{h}_{il})$. This model was only applied for low-dimensional inputs, and this paper investigates its performance and capacity to adapt to high-dimensional settings. We denote this network NNsurv. We noticed an improvement of the performance when using a ReLU activation function for the hidden layers and thus used it instead of the original sigmoid functions. Moreover, the original neural network only has one hidden layer. We propose to add one supplementary hidden layer to study if a deeper structure could improve the neural network prediction capacity. We call it the deeper version NNsurv-deep.

3 Results

We compare the performances of the four models (Cox-L1, Cox-nnet, NNsurv, NNsurv-deep) on a real dataset.

Concordance index

To compare the performance, we use the C-index as proposed by Antolini *et al.* [1], noted C_{td}. The index measures whether the prediction of the model under study matches the rank of the survival data. A value of the C-index close to one indicates the discrimination performance of the model is good.

Application on real breast cancer dataset

The METABRIC dataset (for *Molecular Taxonomy of Breast Cancer International Consortiulm*) [5, 11] consisted of six clinical variables (age, tumor size, hormone therapy, chemotherapy, tumor grades, number of invaded lymph nodes), and 863 genes for 1981 patients. The percentage of censored individuals is equal to 55%.

The results of the METABRIC dataset are summarized in TABLE 1. We can see that NNsurv-deep manages to get the highest value of C_{td}. The C_{td} of NNsurv is equivalent to that of Cox, but Cox-nnet has a lower value. Due to the low difference of C-index between the different methods, we also used another metric, the Integrated Brier Score. The detailed results are in Sautreuil *et al.* [12].

		Cox-L1	Cox-nnet	NNsurv-deep	NNsurv
METABRIC	C_{td}	0.6757	0.6676	**0.6853**	0.6728

Table 1: Results of different methods on the breast dataset (METABRIC)

4 Discussion

This work is a study of neural networks' potential for the prediction of survival in high-dimension. We studied two approaches of neural networks: a first one based on the Cox model, called Cox-nnet [3] and a second one based on a discrete-time model [2] and its adaptation to the high-dimensional setting was the main contribution of our work. We compared these two approaches: Cox-nnet and this based on a discrete-time model adapted to the high dimension (NNsurv, and NNsurv-deep) with the standard Cox model coupled with Lasso penalty. On the METABRIC data, NNsurv-deep performs the best, but only marginally better than the Cox partial log-likelihood-based Lasso estimation procedure (Cox-L1), suggesting slight non-linearity and interactions. We also compared these methods on simulation datasets. This comparison study is detailed in Sautreuil *et al.* [12]. We concluded from the simulation study that the best neural network in most situations is Cox-nnet. It can handle nonlinear effects as well as interactions. However, the neural

network based on discrete-time modeling, which directly predicts the hazard risk, with several hidden layers (NNsurv-deep), has shown its superiority in the most complex situations, especially in the presence of non-proportional risks and intersecting survival curves. The neural networks seem to be interesting methods to predict survival in high-dimension and, in particular, in the presence of complex data. But, the Cox model stays privileged by the domain's users nowadays thanks to the ease of use and interpretation.

Bibliography

[1] L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, 2005.

[2] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*, 17(10):1169–1186, 1998.

[3] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational Biology*, 14(4):e1006076, 2018.

[4] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[5] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, METABRIC Group, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[6] J. Fan, Y. Feng, and Y. Wu. High-dimensional variable selection for Cox's proportional hazards model. Institute of Mathematical Statistics, 2010.

[7] D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995.

[8] A. Guilloux, S. Lemler, and M.-L. Taupin. Adaptive kernel estimation of the baseline function in the cox model with high-dimensional covariates. *Journal of Multivariate Analysis*, 148:141–159, 2016.

[9] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

[10] H. Kvamme, Ørnulf Borgan, and I. Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

[11] L. Mangravite. METABRIC Data for Use in Independent Research. 2013. Publisher: Synapse.

[12] M. Sautreuil, S. Lemler, and P.-H. Cournède. Neural networks to predict survival from RNA-seq data in oncology. working paper or preprint, 2021.

[13] R. Tibshirani. The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, 16(4):385–395, 1997.

Testing for the Markov property in sequential decision making

Chengchun Shi[1*]

[1]*Department of Statistics, London School of Economics and Political Science*

Abstract: The Markov assumption (MA) is fundamental to the empirical validity of reinforcement learning. In this paper, we review the Forward-Backward Learning procedure developed by [6] to test MA in sequential decision making. The test does not assume any parametric form on the joint distribution of the observed data and plays an important role for identifying the optimal policy in high-order Markov decision processes (MDPs) and partially observable MDPs.

Keywords: Reinforcement Learning; Markov Decision Process; Markov Property; Forward-Backward Learning

AMS subject classification: 68T05.

1 Introduction

Reinforcement learning (RL) is a general technique that allows an agent to interact with an environment. In RL, the state-action-reward triplet is typically modelled by the Markov decision process (MDP, see e.g. [5]). Central to the empirical validity of various RL algorithms is the Markov assumption (MA). Under MA, there exists an optimal stationary policy that is no worse than any non-stationary or history-dependent policies [5]. When this assumption is violated, however, the optimal policy might depend on lagged variables and any stationary policy can be sub-optimal. Thus, MA forms the basis for us to select the set of state variables to implement RL algorithms.

Shi et al. [6] recently developed a Forward-Backward Learning procedure to test MA in sequential decision making. In the following, we summarize the advances of this test. First, the test is useful in identifying the optimal policy in high-order MDPs (HMDPs). Under HMDPs, the optimal policy at time t depends not only on the current state variables S_t, but also the past state-action pairs (S_{t-1}, A_{t-1}), $\cdots$, $(S_{t-\kappa_0+1}, A_{t-\kappa_0+1})$ for some $\kappa_0 > 1$. In real-world applications, it remains challenging to properly select the look-back period κ_0. On one hand, κ_0 shall be sufficiently large to guarantee MA holds. On the other hand, including too many lagged variables will result in a very noisy policy. To determine κ_0, we can construct the state by concatenating measurements taken at time points $t, \cdots, t-k+1$ and

*Corresponding author: c.shi7@lse.ac.uk

sequentially apply the test for $k = 1, 2, \cdots$, until the null hypothesis MA is not rejected. Then we use existing RL algorithms based on the constructed state to estimate the optimal policy. It was shown in Section 5 of [6] that the estimated policy based on the constructed states achieves the largest value in a number of cases.

Second, the test is useful in detecting partially observable MDPs (POMDPs). Suppose we concatenate measurements over sufficiently many decision points and the test still rejects MA. Then we shall consider modelling the system dynamics by POMDPs or other non-Markovian problems. Applying RL algorithms designed for these settings has been shown to outperform those for standard MDPs [see e.g. 3].

Third, major challenges for testing MA arise when the state vector is moderate or high-dimensional. This is certainly the case as we convert the process into an MDP by concatenating data over multiple decision points. Modern machine learning (ML) algorithms are well-suited for prediction tasks in high dimensions. Yet, the large bias of the resulting estimates makes statistical inference (e.g., hypothesis testing) extremely difficult. The key ingredient of the test lies in constructing a doubly robust estimating equation to alleviate the biases. This ensures it has a tractable limiting distribution even in high dimensions. Consequently, it well controls the type-I error rate (see Theorem 3 of [6]).

Lastly, the test is valid as either the number of trajectories n or the number of decision points T in each trajectory diverges to infinity. It can thus be applied to a variety of sequential decision making problems ranging from the Framingham heart study [7] with over two thousand trajectories to the OhioT1DM dataset [4] that contains eight weeks' worth of data for six trajectories. It can also be applied to applications from video games where both n and T approach infinity.

2 Model Setup

MDP and Existence of the Optimal Stationary Policy

The objective of RL is to learn an optimal policy that maximizes discounted cumulative reward under this policy. In MDPs, it is typically assumed that the Markov assumption holds such that future state and current reward are conditionally independent of the past observations given the current state-action pair. Under MA, there exists an optimal stationary policy whose value is no worse than any history-dependent policies (see e.g., Lemma 1 in [6]). This observation forms the basis of most existing RL algorithms. Under MA, it suffices to restrict attention to stationary policies. It greatly simplifies the estimating procedure of the optimal policy. When MA is violated however, we need to focus on history-dependent policies as they may yield larger value functions. In the following, we introduce two variants of MDPs, including HMDPs and POMDPs. These models are illustrated in Figure 1.

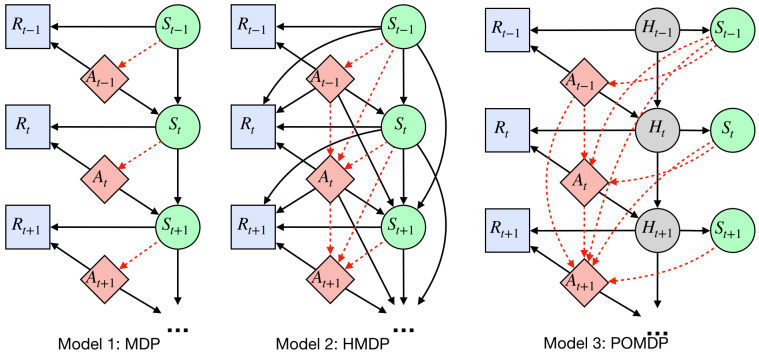


Figure 1: Causal diagrams for MDPs, HMDPs (second order) and POMDPs. (S_t, A_t, R_t) denotes the state-action-reward triplet at time t. The solid lines represent the causal relationships and the dashed lines indicate the information needed to implement the optimal policy.

HMDP

It can be seen from Figure 1 that HMDPs are very similar to MDPs. The difference lies in that in HMDPs, S_{t+1} and R_t depend not only on (S_t, A_t), but $(S_{t-1}, A_{t-1}), \cdots, (S_{t-\kappa_0+1}, A_{t-\kappa_0+1})$ for some integer $\kappa_0 > 1$ as well. For any integer $k > 0$, define a new state variable

$$S_t(k) = (S_t^\top, A_t, S_{t+1}^\top, A_{t+1}, \cdots, S_{t+k-1}^\top)^\top.$$

Let $A_t(k) = A_{t+k-1}$ and $R_t(k) = R_{t+k-1}$ for any t, k. It follows that the new process formed by the triplets $(S_t(\kappa_0), A_t(\kappa_0), R_t(\kappa_0))_{t\geq 0}$ satisfies MA. As such, there exists an optimal stationary policy that depends only on $S_t(\kappa_0)$. This suggests that in HMDPs, identification of the optimal policy relies on correct specification of the look-back period κ_0. To determine κ_0, we can sequentially test whether the triplets $\{(S_t(k), A_t(k), R_t(k))\}_{t\geq 0}$ satisfy MA for $k = 1, 2, \cdots$, until the null of MA is not rejected.

POMDP

The POMDP model can be described as follows. At time $t-1$, suppose the environment is in some hidden state H_{t-1}. The hidden variables $\{H_t\}_{t\geq 0}$ are unobserved. Suppose the agent chooses an action A_{t-1}. Similar to MDPs, this will cause the environment to transition to a new state H_t at time t. At the same time, the agent receives an observation S_t and a reward R_t that depend on H_t and A_{t-1}. The observations in POMDPs do not satisfy the Markov property. As a result, MA will not hold no matter how many past measurements the state variable includes. This suggests in POMDPs, the optimal policy could be history dependent.

3 Testing the Markov Assumption

We introduce the Forward-Backward Learning procedure in this section. We focus on testing the following pair of hypotheses:

$$\mathcal{H}_0\text{: The system is a MDP, i.e, MA holds v.s}$$
$$\mathcal{H}_1\text{: The system is a HMDP or POMDP.}$$

Assume the data generating process is stationary in time. Theorem 1 in [6] characterizes MA based on the notion of Conditional Characteristic Function (CCF). It shows that under $\mathcal{H}_0$,

$$[\{\exp(i\mu^\top S_{t+q+1}) - \varphi^*(\mu|X_{t+q})\}\exp(i\nu^\top X_{t-1})] = 0, \tag{1}$$

for any t, q, μ, ν where $X_t = (S_t^\top, A_t)^\top$, the state-action pair at time t and φ^* denotes the CCF of S_{t+1} given X_t, i.e., $\varphi^*(\mu|\bullet) = \{\exp(i\mu^\top S_{t+1})|X_t = \bullet\}$.

To construct the test statistic based on (1), the CCF φ^* needs to be estimated from the observed data. Modern machine learning algorithms are well-suited to estimating φ^* in moderate or high-dimensional cases. However, naively plugging ML estimators will cause a heavy bias in the estimating equation. Because of that, the resulting test does not have a tractable limiting distribution. Kernel smoothers or local polynomial regression can be employed to reduce the estimation bias by properly choosing the bandwidth parameter. However, these methods suffer from the curse of dimensionality and will perform poorly in cases where concatenate data over multiple decision points to detect HMDP.

The Forward-Backward Learning method addresses these concerns by presenting a doubly-robust estimating equation to alleviate the estimation bias. When observations are time independent, the method shares similar spirits with the double machine learning method proposed by [1] for statistical inference of the average treatment effects in causal inference.

Specifically, define another CCF of X_{t-1} given X_t by

$$\psi(\nu|x) = \{\exp(i\nu^\top X_{t-1})|X_t = x\}.$$

The procedure is motivated by the following identity,

$$[\{\exp(i\mu^\top S_{t+q+1}) - \varphi^*(\mu|X_{t+q})\}\{\exp(i\nu^\top X_{t-1}) - \psi^*(\nu|X_t)\}] = 0,$$

for any $t > 0$, $q \geq 0$, $\mu \in \mathbb{R}^p$, $\nu \in \mathbb{R}^{p+1}$. This equation is doubly-robust. That is, it holds as long as either φ^* or ψ^* is correctly specified.

Forward-Backward Learning estimates both φ^* and ψ^* using ML methods without specifying their parametric forms. Let $\widehat{\varphi}$ and $\widehat{\psi}$ denote the corresponding estimators. Note that computing φ^* is essentially estimating the characteristic function of S_t given X_{t-1}. This corresponds to a forward prediction task. Similarly, estimating ψ^* is a backward prediction task. Thus, $\widehat{\varphi}$ and $\widehat{\psi}$ are referred to as **forward** and **backward learners**, respectively. The procedure constructs a maximum-type statistic based on these learners and applies the multiplier bootstrap [2] to simulate the critical value. Please refer to Algorithm 1 of [6] for details.

Based on this test, we can choose which RL model to use to model the system dynamics. Given a large integer K, one can sequentially test the null hypothesis MA based on the concatenated data as described in Section 2 for $k = 1, \cdots, K$. Once the null is not rejected, we can conclude the system is a k-th order MDP and terminate our procedure. Otherwise, we conclude the system is most likely a POMDP. Please refer to Algorithm 2 of [6] for details.

4 Discussion

In this paper, we briefly review the Forward and Backward Learning procedure [6] for testing the goodness of fit of a MDP model. The test can be naturally coupled with existing state-of-the-art RL algorithms to improve their performance. It has extensive potential values in many real-world applications, including robotics, bidding, ridesharing, mobile health, among others. The validity of the test relies on a stationarity assumption that requires the observed data process to be stationarity over time. In theory, under the stationartiy assumption as well as other mild conditions imposed in [6], it would be impossible for the test to reject the null hypothesis at a small value of k but then rejects the null for a large value of k. However, if such a phenomenon occurs in practice, then some of the imposed assumptions are likely to be violated. In particular, the stationarity assumption shall be further investigated.

Bibliography

[1] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *Econom. J.*, 21(1):C1–C68, 2018.

[2] V. Chernozhukov, D. Chetverikov, K. Kato, et al. Gaussian approximation of suprema of empirical processes. *Annals of Statistics*, 42(4):1564–1597, 2014.

[3] M. Hausknecht and P. Stone. Deep recurrent q-learning for partially observable mdps. In *2015 AAAI Fall Symposium Series*, 2015.

[4] C. Marling and R.C. Bunescu. The ohiot1dm dataset for blood glucose level prediction. In KHD@ IJCAI, pages 60–63, 2018.

[5] M.L. Puterman. Markov decision processes: discrete stochastic dynamic programming. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley Sons, Inc., New York, 1994. A Wiley-Interscience Publication.

[6] C. Shi, R. Wan, R. Song, W. Lu, and L. Leng. Does the markov decision process fit the data: testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pages 8807–8817. PMLR, 2020.

[7] C.W. Tsao and R.S. Vasan. Cohort profile: The framingham heart study (fhs): overview of milestones in cardiovascular epidemiology. *International journal of epidemiology*, 44(6):1800–1813, 2015.

Mixed moment estimator for space-time heteroscedastic extremes: semi-parametric inference on extreme rainfall

Jessica Silva Lomba,[1*] **Maria Isabel Fraga Alves**[1] **and Cláudia Neves**[2]

[1]*CEAUL, Faculdade de Ciências, Universidade de Lisboa, Portugal*
[2]*Department of Mathematics, Faculty of Natural, Mathematical & Engineering Sciences, King's College London, United Kingdom*

Abstract: Extreme meteorological phenomena such as heavy precipitation seem to be growing more severe and frequent as a result of accelerating climate change. Estimation of this evolution falls within the scope of extreme value statistics, to which end the estimation of the so-called extreme value index is key. We look at how one can use series of data collected at isolated locations to model extremes of the whole space-time process by employing the mixed moment estimator of the extreme value index. We show the asymptotic normality of this estimator, seamlessly incorporating space-time non-stationarity and dependence. For illustration, we apply the estimator to precipitation data from a homogeneous region in the North Sea.

Keywords: Extreme value statistics, extreme rainfall, non-identical distributions, semi-parametric inference.

AMS subject classification: 60G70, 62G05, 62G20, 62G30, 62G32, 62P12.

1 Introduction

When discussing the apparent acceleration of climate change, it is vital to analyze the frequency and severity of extreme weather events, beyond what is observable with average events. For example, studying extreme precipitation accounting for potential trend and dependence across time and/or space may lead to more robust inference, in turn allowing for better preparation against potentially catastrophic

*Corresponding author: jslomba@fc.ul.pt

rainfall. However, estimation of this evolution in extreme weather events remains subject to large uncertainty.

Extreme Value Analysis provides a complete and flexible framework for the statistical study of rare and extreme events that may never yet been observed [1]. In this work, we follow a semi-parametric approach, making assumptions regarding the asymptotic behavior of the tail of the distribution underlying the data, without demanding a strict parametric fit.

In the univariate case, with $\{X_i\}_{i\geq 1}$ a sequence of i.i.d. random variables with distribution function (d.f.) F, if there exist constant sequences $a_n > 0$, $b_n \in \mathbb{R}$ and a non-degenerate d.f. G such that

$$\lim_{n\to\infty} P\left\{(a_n)^{-1}(\max(X_1,\ldots,X_n) - b_n) \leq x\right\} = G(x)$$

for all x, then G is of the type of the Generalized Extreme Value distribution $G_\gamma(x) = \exp\left(-(1+\gamma x)^{-1/\gamma}\right)$, for $1+\gamma x > 0$. We then write $F \in \mathcal{D}(G_\gamma)$, i.e., F belongs to a max-domain of attraction with extreme value index (EVI) γ. This index controls the tail weight; with $x^* := \sup\{x : F(x) < 1\}$ the right endpoint of F, we have: $\gamma > 0$, heavy tails with infinite x^*; $\gamma < 0$, light tails with finite x^*; $\gamma = 0$, exponential tails with finite or infinite x^*. Effective estimation of the EVI is thus paramount to all extreme value statistical endeavors. This is the focus of the present work.

We consider independent random vectors $(X_{i,1},\ldots,X_{i,m})$ with time points $i = 1,\ldots,n$, at spatially dependent locations $j = 1,\ldots,m$. The interest is in an equivalence class within the max-domain of attraction, allowing the presence of non-stationarity in the form of a non-monotonic trend. We assume there exists a baseline d.f. $F_0 \in \mathcal{D}(G_\gamma)$ for some $\gamma \in \mathbb{R}$ such that, for $i = 1,\ldots,n$, $j = 1,\ldots,m$

$$\lim_{x\uparrow x^*} \frac{1-F_{i,j}(x)}{1-F_0(x)} = c\left(\frac{i}{n}, j\right) \in (0,\infty) \tag{1}$$

where $F_{i,j}$ is the d.f. of $X_{i,j}$, x^* is the right endpoint of F_0, and the *scedasis* $c(\cdot, j)$, at each location j, is a positive continuous function on $[0,1]$. The scedasis embodies the so-called space-time trend in extremes [2]. As a result of (1), we have $F_{i,j} \in \mathcal{D}(G_\gamma)$ with common γ, for all $i = 1,\ldots,n$, and $j = 1,\ldots,m$. The scedasis is uniquely determined through

$$\sum_{j=1}^{m} C_j(1) = 1\,, \quad \text{where} \quad C_j(t) := \frac{1}{m}\int_0^t c(u,j)\,du\,, \quad 0 \leq t \leq 1.$$

We further assume that the standardized joint d.f. $\tilde{F}(x_1,\ldots,x_m) := F_i(U_{i,1}(x_1),\ldots,U_{i,m}(x_m))$ is independent of i and in a multivariate max-domain of attraction (see [1]), where $U_{i,j}(t) := (\frac{1}{1-F_{i,j}})^{\leftarrow}(t)$ is the left-continuous inverse of $1/(1-F_{i,j})$. This spatial dependence structure is captured through the tail copula of (X_{i,j_1}, X_{i,j_2}), given by

$$R_{j_1,j_2}(x_1,x_2)=\lim_{t\downarrow 0}\frac{1}{t}P\left(1-F_{i,j_1}(X_{i,j_1})\le tx_1\,,1-F_{i,j_2}(X_{i,j_2})\le tx_2\right)\,,$$

for $(x_1,x_2)\in[0,\infty]^2\setminus\{(\infty,\infty)\}$ and $1\le j_1\,,j_2\le m$.

Under this setup, we propose estimation of the common EVI by the mixed moment estimator (MMe). In Section 2, we present the MMe and show it is an appealing estimator for heteroscedastic extremes (asymptotically normal with null first order bias). Section 3 briefly describes the application of the MMe to precipitation data from the North Sea. A sketch of proof for the main theorem is given in Section 4.

2 Mixed Moment Estimator

The MMe was introduced in [3] for the univariate, independent setting and seen to have convenient properties when compared to other common EVI estimators. We investigate its performance when pooling information from the full set of $N:=n\times m$ observations collected across time and space and considering the exceedances of an overall, random threshold.

Definition 1. Let $X_{1:N}\le\ldots\le X_{N:N}$ be the order statistics of the pooled sample $\{X_{i,j}\}_{i=1}^{n}{}_{j=1}^{m}$ consisting of N space-time random variables. The **mixed moment estimator** of the EVI $\gamma\in\mathbb{R}$ is defined as (cf. [3])

$$\widehat{\gamma}_N^{MM}(k):=\frac{\widehat{\varphi}_N(k)-1}{1+2\min\{\widehat{\varphi}_N(k)-1,0\}}$$

where, with $X_{N-k:N}$ the common threshold, $M_N(k):=\frac{1}{k}\sum_{i=1}^{k}\log\left(\frac{X_{N-i+1:N}}{X_{N-k:N}}\right)$ and $L_N(k):=\frac{1}{k}\sum_{i=1}^{k}\left(1-\frac{X_{N-k:N}}{X_{N-i+1:N}}\right)$,

$$\widehat{\varphi}_N(k):=\frac{M_N(k)-L_N(k)}{\left(L_N(k)\right)^2}\tag{2}$$

is the estimator of the function $\varphi(\gamma):=1+\gamma$ if $\gamma>0$, $\varphi(\gamma):=\frac{1-\gamma}{1-2\gamma}$ if $\gamma\le 0$.

Compared with the Maximum Likelihood estimator (MLe), the MMe has the advantage of an explicit and simple form, making it very computationally effective; also, the MMe is widely applicable, suitable for any $\gamma\in\mathbb{R}$. Its asymptotic properties have been derived for the i.i.d. univariate case in [3].

To conclude on the asymptotic behavior of the MMe in the present framework, we make use of the weighted approximation to the tail empirical quantile process $X_{N-[ks]:N}$ presented in Theorem 2.1 b) of [2]. We must, thus, impose a second order condition controlling the speed of convergence of the baseline quantile function: with $\gamma\in\mathbb{R}$, $\rho<0$, assume there exist functions $\tilde{a}_0$, positive and A_0, eventually not changing sign, satisfying $\lim_{t\to\infty}A_0(t)=0$ such that

$$\lim_{t\to\infty}\frac{\frac{U_0(tx)-U_0(t)}{\tilde{a}_0(t)}-\frac{x^\gamma-1}{\gamma}}{A_0(t)}=\Psi_{\gamma,\rho}(x):=\begin{cases}\dfrac{x^{\gamma+\rho}-1}{\gamma+\rho} & \text{if } \gamma+\rho\neq 0\\ \log(x) & \text{if } \gamma+\rho=0\end{cases},$$

for all $x>0$ (cf. Corollary 2.3.5 in [1]). A similar condition holds replacing $\tilde{a}_0$, $U_0(t)$ and Ψ by suitable functions a_0, b_0 and $\bar{\Psi}$ (cf. Corollary 2.3.7 in [1]). Other

necessary conditions on the scedasis function are detailed in [2]. We now state the main result regarding the MMe's asymptotic normality.

Theorem 1. *Under the conditions of Theorem 2.1 of [2], for an appropriate intermediate sequence* $k \equiv k_n \to \infty$, $k/n \to 0$ *as* $n \to \infty$ *such that* $\sqrt{k} A_0(N/k) \to 0$,

$$\sqrt{k}\left(\widehat{\gamma}_N^{MM}(k) - \gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_{MM}^2\right)$$

with $\sigma_{MM}^2 :=$

$$\begin{cases} (1+\gamma)^2 + \frac{(1+\gamma)^4}{\gamma^2} \iint\limits_{[0,1]^2} f_\gamma(s) f_\gamma(t) \sum\limits_{j_1 \neq j_2} \mathbb{E}\left[\alpha_{j_1}(s)\alpha_{j_2}(t)\right] ds\, dt, & \gamma > 0 \\ \frac{(1-\gamma)^2(1-2\gamma)(6\gamma^2-\gamma+1)}{(1-3\gamma)(1-4\gamma)} + \frac{(1-\gamma)^4(1-2\gamma)^2}{\gamma^2} \iint\limits_{[0,1]^2} g_\gamma(s) g_\gamma(t) \sum\limits_{j_1 \neq j_2} \mathbb{E}\left[\beta_{j_1}(s)\beta_{j_2}(t)\right] ds\, dt, & \gamma \leq 0 \end{cases} \tag{3}$$

where

$$f_\gamma(s) := 1 - (1+2\gamma)s^\gamma \;\; \textit{and} \;\; g_\gamma(s) := (1-2\gamma)s^{-\gamma} - 1, \tag{4}$$

$$\alpha_j(s) := s^{-1} W_j(s, C_j(1)) - W_j(1, C_j(1)) \;\; \textit{and} \tag{5}$$

$$\beta_j(s) := s^{-\gamma-1} W_j(s, C_j(1)) - W_j(1, C_j(1)), \tag{6}$$

for $j = 1, \ldots, m$ *and where* W_j *is a bivariate Wiener process, for each* j, *as given in Theorem 2.1 of [2].*

Remark 4. The first term appearing in each branch of (3) corresponds to the *intra-station* variability, where observations are independent, and thus coincides with the asymptotic variance in the i.i.d. setting.

3 Application: Extreme Precipitation

Our data consists of daily rainfall totals collected at $m = 5$ stations within a homogeneous region in the North Sea, off the coast of the UK. Records span the Summer and Winter seasons between January 1979 and December 2010, totaling $n = 1358/n = 1341$ days, respectively.

Applying the tests given in [2] for each season, we found there is evidence of spatial dependence, as well as of a trend in the rainfall extremes (stronger in Winter). There was no evidence that assuming a common EVI is inappropriate.

Figure 1 shows clear differences between the MMe and MLe sample paths, to be studied. The MMe-based approximately 95% confidence bands for γ show that disregarding the dependence structure results in narrower intervals, potentially leading to underestimated risk of very heavy rainfall. With $k = 400$ largest observations, the common threshold is set at 13.2mm/13.9mm, which determines $\widehat{\gamma}_{N,S}^{MM}(400) = 0.118$ and $\widehat{\gamma}_{N,W}^{MM}(400) = 0.060$.

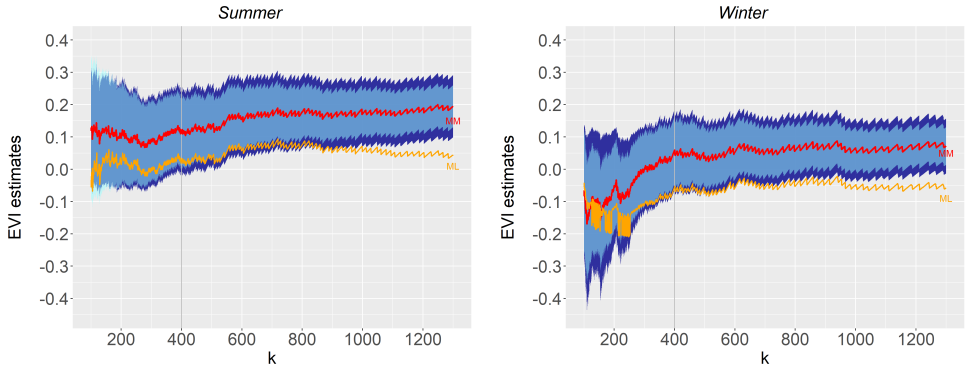


Figure 1: MMe sample path (red) for the North Sea precipitation data in the Summer (left) and Winter (right) seasons, with approximately 95% confidence bands considering spatial dependence (dark blue) or independence (light blue). Sample path for the MLe (orange) for comparison.

4 Proof of Theorem 1

Proof. For the sake of brevity, we present only a sketch of the proof, which relies strongly on Theorem 2.1 b) of [2]. Define $M_N(k) := \int_0^1 \log \frac{X_{N-[ks]:N}}{X_{N-k:N}} ds$ and $L_N(k) := \int_0^1 1 - \frac{X_{N-k:N}}{X_{N-[ks]:N}} ds$. The asymptotic behavior of $\widehat{\varphi}_N(k)$ follows from

$$\sqrt{k}\left\{\widehat{\varphi}_N(k) - \varphi(\gamma)\right\} = \sqrt{k}\left\{M_N(k) - L_N(k) - \varphi(\gamma)\left(L_N(k)\right)^2\right\}\left(L_N(k)\right)^{-2}. \tag{7}$$

For $\gamma > 0$, Theorem 2.1 b) of [2] suggests an analogous result to that of Theorem 2.4.8 of [1], allowing for the following asymptotic representations

$$\begin{aligned} M_N(k) &= \gamma + \frac{\gamma}{\sqrt{k}} \int_{\frac{1}{2k}}^1 \mathcal{D}_m(s)\, ds + o_p\left(\frac{1}{\sqrt{k}}\right), \\ L_N(k) &= \frac{\gamma}{1+\gamma} + \frac{\gamma}{\sqrt{k}} \int_{\frac{1}{2k}}^1 s^\gamma \mathcal{D}_m(s)\, ds + o_p\left(\frac{1}{\sqrt{k}}\right) \text{ and} \\ \left(L_N(k)\right)^2 &= \left(\frac{\gamma}{1+\gamma}\right)^2 + \frac{2\gamma^2}{(1+\gamma)\sqrt{k}} \int_{\frac{1}{2k}}^1 s^\gamma \mathcal{D}_m(s)\, ds + o_p\left(\frac{1}{\sqrt{k}}\right), \end{aligned} \tag{8}$$

as $n \to \infty$, with $\mathcal{D}_m(s) := \sum_{j=1}^m \alpha_j(s)$ and $\alpha_j(s)$ as in (5). Then, combining (7) and (8) we get

$$\sqrt{k}\left\{\widehat{\varphi}_N(k) - \varphi(\gamma)\right\} \xrightarrow{d} \frac{(1+\gamma)^2}{\gamma} \int_0^1 f_\gamma(s) \mathcal{D}_m(s)\, ds\,,$$

with $f_\gamma(s)$ as defined in (4); the corresponding variance is

$$\sigma_\varphi^2 = \frac{(1+\gamma)^4}{\gamma^2} \iint\limits_{[0,1]^2} f_\gamma(s) f_\gamma(t) \sum_{j_1=1}^m \sum_{j_2=1}^m \mathbb{E}\left[\alpha_{j_1}(s)\, \alpha_{j_2}(t)\right] ds\, dt\,.$$

For $\gamma \leq 0$, again by Theorem 2.1 b) of [2] we get the asymptotic representation

$$\left(L_N(k)\right)^2 = \left\{\frac{a_0(N/k)}{U_0(N/k)}\right\}^2 \left\{\frac{1}{1-\gamma}\right\}^2 \left\{1 + \frac{2(1-\gamma)}{\sqrt{k}} \int_{\frac{1}{2k}}^1 \mathcal{Y}_m(s) ds + o_p\left(\frac{1}{\sqrt{k}}\right)\right\} \tag{9}$$

as $n \to \infty$, with $\mathcal{Y}_m(s) := \sum_{j=1}^{m} \beta_j(s)$ and $\beta_j(s)$ as in (6). Defining $\mathcal{X}_N(s) := 1 - \frac{X_{N-k:N}}{X_{N-[ks]:N}}$ and noting that $\frac{x^2}{2} < -\log(1-x) - x < \frac{x^2}{2} + \frac{x^3}{3(1-x)}$ for $x \in (0,1)$, we bound the numerator of (2) by

$$\mathcal{LB} := \frac{1}{2} \left\{ \frac{U_0(N/k)}{a_0(N/k)} \right\}^2 \int_0^1 \{\mathcal{X}_N(s)\}^2 \, ds \leq \left\{ \frac{U_0(N/k)}{a_0(N/k)} \right\}^2 (M_N(k) - L_N(k))$$
$$\leq \mathcal{LB} + \frac{1}{3} \left\{ \frac{U_0(N/k)}{a_0(N/k)} \right\}^2 \int_0^1 \{\mathcal{X}_N(s)\}^3 \{1 + \mathcal{X}_N(s)\} (1 + o_p(1)) \, ds \, .$$

The last term in the above can be shown to be negligible, in probability, leading to

$$\left\{ \frac{U_0(N/k)}{a_0(N/k)} \right\}^2 (M_N(k) - L_N(k)) = \frac{1}{(1-2\gamma)(1-\gamma)} + \frac{1}{\gamma \sqrt{k}} \int_{\frac{1}{2k}}^1 (s^{-\gamma} - 1) \mathcal{Y}_m(s) ds + o_p\left(\frac{1}{\sqrt{k}}\right). \tag{10}$$

Combining (7), (9) and (10) we have that

$$\sqrt{k} \{\widehat{\varphi}_N(k) - \varphi(\gamma)\} \xrightarrow{d} \frac{(1-\gamma)^2}{\gamma(1-2\gamma)} \int_0^1 g_\gamma(s) \, \mathcal{Y}_m(s) ds \, ,$$

with $g_\gamma(s)$ as defined in (4). The asymptotic variance follows as before.
Finally, the asymptotics of $\widehat{\gamma}_N^{MM}(k)$ follow by application of Cramér's delta method. □

Acknowledgements: The authors gratefully acknowledge support from: FCT, I.P. PhD grant SFRH/BD/130764/2017 (JSL) and project UIDB/00006/2020 (MIFA); EPSRC-UKRI Innovation Fellowship grant EP/S001263/1 (CN).

Bibliography

[1] L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer New York, 2006.

[2] J. H. J. Einmahl, A. Ferreira, L. de Haan, C. Neves, and C. Zhou. Spatial dependence and space-time trend in extreme events. *Ann Stat*, 2021. To appear. arXiv:2003.04265

[3] M. I. Fraga Alves, M. I. Gomes, L. de Haan, and C. Neves. Mixed moment estimator and location invariant alternatives. *Extremes*, 12(2):149–185, 2009.

Combinatorial regression in abstract simplicial complexes

Andrej Srakar,[1*] **and Miroslav Verbic**[2]

[1]*Institute for Economic Research (IER), Ljubljana*
[2]*School of Economics and Business, University of Ljubljana and Institute for Economic Research (IER), Ljubljana*

Abstract: In regression analysis of the market share data several main parametric and nonparametric type models are prevalent. We extend this arsenal of possibilities with a new type labelled combinatorial regression, based on combining n-tuplets of sampling units into groups and treating them in abstract simplicial complex framework. The novel perspective is estimated in two stages, using two different initial stage perspectives: multivariate distance matrix regression and Bradley-Terry based maximum likelihood approach, and a recently developed simplicial complex network estimation approach on simplicial complexes in the second, final stage. This allows plethora of future research perspectives and allows applications also to very small datasets as the number of units in the new model can be expressed in terms of generalized factorial products of units of original sample. We provide the analysis of new approach for different n-tuple combinations using generalized Jensen-Shannon divergence measures and provide short analysis of new estimator properties. In conclusion, extensions and open questions raised by the new perspective are presented.

Keywords: regression models, abstract simplicial complexes, algebraic topology, simplicial complex networks, Jensen-Shannon divergence

AMS subject classification: 62J99.

1 Introduction - regressions on a simplex

In geometry, a simplex (plural: simplexes or simplices) is a generalization of the notion of a triangle or tetrahedron to arbitrary dimensions. The simplex is so-named because it represents the simplest possible polytope in any given space (for example, a 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, a 4-simplex is a 5-cell).

A *composition* is defined as a vector of D positive components $x = (x_1, x_2, \ldots, x_D)$ summing up to a given constant κ. It is generally - although not universally - agreed that the appropriate sample space for compositional data is the standard simplex (also called the "unit simplex"). It is defined as

$$S^D = \{x = [x_1, x_2, \ldots, x_D] | x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa\}.$$

*Corresponding author: andrej.srakar@ier.si

In regression analysis with compositional data (CoDa) four main (parametric) type models are prevalent: multinomial logistic regression, attraction models of various types, Dirichlet covariance models, and compositional regression. Nonparametric regression with CoDa at present consists of three models: local polynomial (Di Marzio et al., 2015); simplicial splines (Machalova, Hron and Talska, 2019); and simplicial wavelets (Srakar and Fry, 2019).

We extend this arsenal of possibilities with a novel regression perspective (applicable to simplicial complexes), named *combinatorial regression*, based on combining n-tuplets of sampling units into groups. This perspective is based on a broad generalization of the Full-Factorial Attraction model from marketing (Howie & Kleczyk, 2007). We extend the Howie and Kleczyk perspective by considering instead of pair of "brands" (regions, etc.) triplets, quadruplets, indeed, any *combinatorial variation* of units as the basis for constructing new regression units.

In our article we develop a large extension of a decade and half ago developed transformation of the MCI model, called Full-Factorial Attraction Model, as developed in Howie and Kleczyk (2007). The approach is based on a reconceptualization of any market share variable for each brand as a series of two-product markets (in this way, the number of units grows to $\frac{I!}{2!}$ (see Howie and Kleczyk, 2007 - I is the number of units/brands) which gains quite a lot of degrees of freedom for the analysis).

The final equation for Full-Factorial Attraction Model is provided below:

$$m_{ijt} = \alpha_i + \beta X_{ijt} + \epsilon_{it},$$

where

- $m_{ijt} = \frac{M_{it}}{(M_{it}+M_{jt})}$ where $i = 1, \ldots, I-1$; $j = 1, \ldots, I-1$ and $i \neq j$, $t = 1, \ldots, T$;
- M_{it} is the market share of brand i at time t;
- $X_{ijt} = x_{it} - x_{jt}$ where $i = 1, \ldots, I$; $j = 1, \ldots, I$ and $i \neq j$, $t = 1, \ldots, T$;
- t is a time variable and T is the maximal time;
- α_i is a parameter for the constant influence of brand i;
- ϵ_i is a random error term.

2 Abstract simplicial complexes and construction of the estimator

Topological data analysis (TDA) is a data analysis method that provides information about the shape of data. It has been developed within the last twenty years and is rooted in the mathematical field of algebraic topology (" *Topology is the branch of mathematics that studies shape, and algebraic topology is the application of tools from abstract algebra to quantify shape.*")

A simplicial complex K consists of:

- A set of objects, $V(K)$, i.e. vertices
- A set, $S(K)$, of finite non-empty subsets of $V(K)$, i.e. simplices such that simplices satisfy the following conditions:
  1. If $\sigma \in V(K)$ is a simplex and $\tau \in \sigma, \tau \neq 0$, then τ is also a simplex;

2. Every singleton $\{v\}, v \in V(K)$, is a simplex.

We say τ is a face of σ. If $\sigma \in S(K)$ has $p+1$ elements it is said to be a p-simplex. The set of p -simplices of K is denoted by K_p. The dimension of K is the largest p such that K_p is non-empty. A map of simplicial complexes $K \to L$ is a function $f : V(K) \to V(L)$ such that whenever $\sigma \in V(K)$ belongs to $S(K)$, the image $f(\sigma)$ belongs to $S(L)$.

We define *geometric simplicial complex* as a finite collection of simplices K called the faces of K such that:

- $\forall \sigma \in K, \sigma$ is a simplex
- $\sigma \in K, \tau \in \sigma \Rightarrow \tau \in K$
- $\forall \sigma, \tau \in K$, either $\sigma \cap \tau = \emptyset$ or $\sigma \cap \tau$ is a common face of both

Given a finite set of elements P, an *abstract simplicial complex* K with vertex set P is a set of subsets of P such that:

- $\forall p \in P, p \in K$
- if $\sigma \in K$ and $\tau \subseteq \sigma$, then $\tau \in K$

McArdle and Anderson (2001) proposed a nonparametric regression approach, based on pairwise distances between vectors of scores on the outcome variables. Multivariate Distance Matrix Regression (MDMR) quantifies structure in the data based on similarities between subjects rather than similarities between variables. Distance between two vectors of scores on a multivariate outcome is defined as the result of a function $d(Y_i^{'}, Y_j^{'})$ that quantifies the dissimilarity of the response profiles of subjects i and j, i.e. distance between their $q \times 1$ vectors of scores on the variables comprising Y.

Second possibility of individual equation level estimation is to use generalized Bradley-Terry perspectives, such as Placket Luce model.

Our final estimation of the combinatorial regression coefficients in simplical complex framework uses a simplicial complex network (SCN) perspective (Firouzi et al., 2019).

It reformulizes the regression problem using simplicial complex network:

- In case that a SCN has no hidden node (no subdivisioning process), it can be viewed as linear regression reformulation. A real valued linear function $f : \Delta^d \to R$ from a d-dimensional simplex $\Delta^d = [v_o, \ldots, v_d]$, can be specified by the values of f at each v_i. These values are represented by $f(v_i)$.
- Assume a data matrix $X \in R^{N \times d}$ of N samples within Δ^d, and their corresponding output in a vector y. We formulate the linear regression problem with training a weight w that minimizes $\|Xw - y\|_2^2$.
- Coefficients of the representation samples in X as represented as a convex combination of $v_0, \ldots, v_d$ in a matrix $C \in R^{N \times (d+1)}$ with a rank of at most d, where i-th row indicates the corresponding coefficients for i-th sample. Then the linear regression problem can be reformulated as $\|Cf - y\|_2^2$ where f is a $(d+1)$ dimensional vector representing the function value at v_i as its i-th element. With a straightforward computation, one can verify that the optimal w or f can be computed from the optimal value of the other one.

3 Properties of the estimation approach for Jensen-Shannon divergence measure

We present results of simulation study using Jensen-Shannon divergence measure as the basis for the construction of the estimators.

The Jensen-Shannon divergence (JSD) $M_+^1(A) \times M_+^1(A) \to [0, \infty)$ is a symmetrized and smoothed version of the Kullback-Leibler divergence $D(P \in Q)$. It is defined by: $JSD(P \in Q) = \frac{1}{2}D(P \in M) + \frac{1}{2}D(Q \in M)$ where $M = \frac{1}{2}(P + Q)$.

Figure 1 presents the results of two simulations, where the data generating process has been constructed for a full combinatorial set of 4-tuplets (left) and 8-tuplets (right). Interestingly, for the used data generating processes, generalized Bradley-Terry based estimation approach seems as the most consistent and best performing.

Figure 1: Results of Monte Carlo simulation for two different data generating processes

			DGP_1						DGP_2		
Data	**OLS**	**LL**	**MRMD-JS**	**MRMD-GJS**	**GBT**	**Data**	**OLS**	**LL**	**MRMD-JS**	**MRMD-GJS**	**GBT**
Gaussian	0.8703	0.8790	0.8730	0.8785	0.9196	**Gaussian**	0.8355	0.8439	0.8555	0.8345	0.9012
10	0.1765	0.1747	0.1808	0.1783	0.1808	**10**	0.1747	0.1677	0.1754	0.1765	0.1736
	0.9066	0.9157	0.8818	0.8964	0.9289		0.8885	0.8699	0.8377	0.8605	0.9103
20	0.1234	0.1172	0.1271	0.1281	0.1126	**20**	0.1222	0.1137	0.1220	0.1243	0.1092
	0.9215	0.8849	0.9204	0.9252	0.9368		0.8754	0.8407	0.8928	0.9067	0.9087
50	0.0979	0.0989	0.0876	0.0909	0.0723	**50**	0.0940	0.0979	0.0858	0.0891	0.0701
	0.9338	0.9244	0.9279	0.9307	0.9438		0.8964	0.8966	0.9001	0.9028	0.8966
100	0.0823	0.0782	0.0850	0.0874	0.0693	**100**	0.0807	0.0743	0.0808	0.0857	0.0672
Log normal	0.8687	0.8921	0.8376	0.8797	0.9029	**Log normal**	0.8600	0.8654	0.8209	0.8622	0.8668
10	0.1708	0.1725	0.1746	0.1725	0.1648	**10**	0.1623	0.1656	0.1711	0.1691	0.1582
	0.9049	0.9001	0.8817	0.8977	0.9213		0.8868	0.8911	0.8376	0.8708	0.8937
20	0.1212	0.1176	0.1247	0.1254	0.1190	**20**	0.1188	0.1164	0.1222	0.1204	0.1166
	0.9296	0.9026	0.9200	0.9321	0.9382		0.8924	0.8575	0.8832	0.9041	0.9007
50	0.0979	0.0999	0.0962	0.1022	0.0787	**50**	0.0960	0.0969	0.0952	0.1002	0.0748
	0.9305	0.9238	0.9314	0.9446	0.9451		0.9119	0.9146	0.9035	0.8974	0.9262
100	0.0823	0.0807	0.0859	0.0881	0.0614	**100**	0.0782	0.0790	0.0842	0.0872	0.0602

Source: Own calculations.

4 Conclusion

We present a new and unexplored regression perspective, to our knowledge second one on simplicial complexes, opening up vast area for future research with most of the options the approach provides still unexplored, for example:

- Statistical criteria for the selection of combinations to be included in the combinatorial regression analysis and model fit criteria
- Parametric, semi- and nonparametric perspectives
- Combinations with other approaches in mathematical statistics and econometrics, for example Bayesian approaches, causal inference, additional combinations with machine learning methods
- Probabilistic perspectives: stochastic processes on simplicial complexes (lattice models, e.g. Ising)
- Extension of the perspectives from algebraic topology and algebraic statistics - regression models on other topological objects (Vietoris-Rips and Cech complexes, matroids, greedoids)

Bibliography

[1] M. Di Marzio, A. Panzera, and K. Venieri. Non-parametric regression for compositional data. *Statistical Modelling*, 15(2):113–133, 2015.

[2] M. Firouzi, S. Boreiri, and H. Firouzi. Simplicial Complex Networks. *ICLR 2020 Conference Blind Submission*, 2020.

[3] P. Howie, and E. Kleczyk. New Developments in Panel Data Estimation: Full-Factorial Panel Data Model. *AAEA Meeting. Portland, Oregon*, 2007.

[4] B. Korte, L. Lovasz, and R. Schrader. *Greedoids*. Springer Verlag, Berlin, 1991.

[5] J.M. Lee. *Introduction to Topological Manifolds*. Springer, New York and London, 2011.

[6] J. Machalova, K. Hron, and R. Talska. *Simplicial splines: application and possible extensions. Proceedings of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019)*, 2019.

[7] B.H. McArdle, and M.J. Anderson. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, 82:290–297, 2001.

[8] A. Srakar, and T.R.L.. Fry. *Wavelet regressions for compositional data. Proceedings of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019)*, 2019.

Application of survival techniques to establish environmental and operational controls on road bridge deterioration

Nicola-Ann Stevens,[1*] Myra Lydon[1], Adele H. Marshall[2] and Su E. Taylor[1]

[1]*FacultySchool of Natural and Built Environment, Queen's University Belfast*
[2]*School of Mathematics and Physics, Queen's University Belfast*

Abstract: One of the key components of a bridge management system (BMS) is the deterioration model, whose accuracy will determine the quality of future maintenance, rehabilitation and replacement (MR&R) decisions. The current state of the art for deterioration models in commercial BMS is the application of Markov chains to determine the probability of transitioning between condition states (CS). However, in recent years research has moved to looking at alternative approaches due to the assumption of a constant bridge population and stationary transition probabilities. This can compromise the efficacy of the deterioration model for predicting deterioration of complex aging bridge structures. This research focuses on the application of survival analysis where the survival time is the time spent in each CS and the 'failure' is the transition of the bridge condition to a worse state. Bridge condition states are measured on an ordinal or numerical scale, common examples range from 1-4 or 0-100 for the UK standardised bridge condition index (BCI). This paper presents the application of survival analysis to establish environmental controls and explores the comprehensive assessment of the utilisation of the Cox Proportional Hazards (PH) model. Expanding on the authors previous research which applied survival techniques to identify bridge performance indicators, this research uses live bridge condition data from Northern Ireland, based on over 6000 bridges which form the strategic and regional road network.

Keywords: survival analysis, bridge management systems, Markov chains, deterioration modelling

AMS subject classification: 62N03

1 Introduction

The public road network in Northern Ireland (NI) has an estimated value of £26billion which makes it NI's most valuable capital asset. This road network contains 6978 bridges of which approximately 6000 bridges meet the criteria to require regular inspections. Bridges are subject to regular inspections to ensure functionality and safety to users. Since the early 1980's, a score on an ordinal scale between 1 and 4 was given at the time of inspection where 1 represented minimal damage and 4 indicated immediate intervention was required.

*Corresponding author: nstevens01@qub.ac.uk

Based on the scores for the individual elements, an overall score was awarded. This score was then used to allocate budgets and spending. The undefined boundaries and broad range of categories led to problems in identifying structures which were most in need of immediate action. As a result, the introduction of a new inspection scoring mechanism was needed. In 2015, the first inspection using the Bridge Condition Index (BCI) in NI was conducted. With this change of inspection method resulting in a large portion of the inspection data being in a different format, research was carried out to investigate converting this inspection data to the new format. The process undertaken was detailed in Stevens et al [6]. The investigation included how missing data and anomalies found in the inspection data were handled. Converting over 17 years of inspection data into BCI means that approximately 20 years of inspection data can be used to build deterioration models.

This paper details an extension to work carried out in [5] where survival analysis techniques were applied to bridge inspection data. This study included quantifying the effect of four bridge characteristics on the survival in each CS. An additional three characteristics will be investigated here. Section 2 will provide a brief literature review of bridge deterioration modelling. An introduction to the data and an in-depth look at the attributes investigated in this study will be given in Section 3 before results are presented in Section 4. Conclusions from this study will be given in Section 5.

2 Literature Review

A variety of models have been used for deterioration modelling. The current state-of-the-art is the application of Markov models where the future condition of the bridge is predicted by calculating the transition probabilities between states. A full review of the literature is available in [5].

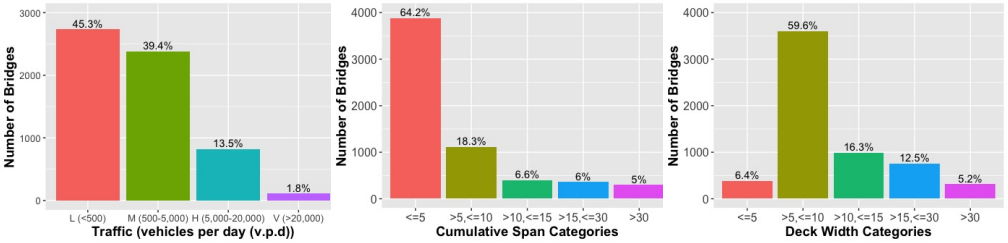
Over the years, research has advanced from Markov models due to the assumptions made in utilising the Markov models such as constant bridge population and stationary transition probabilities. The semi-Markov model was introduced in [3] as some of the restrictions are relaxed. Survival analysis was first used in the calculation of transition probabilities for a semi-Markov model [4]. In addition to this, several studies have used survival analysis to model the time spent in each CS using the Kaplan-Meier method [1], Cox PH [2] and the Weibull and Hypertabastic distributions [8].

3 Introduction to the Data

Preliminary analysis on the NI bridge stock was undertaken and can be seen in [7]. Key points from that investigation were that approximately 53% of bridges have masonry arch construction and approximately 83% have the bridge function of road over river. There are 6978 bridges in NI with approximately 6000 bridges meeting the conditions to require regular inspections. The results presented in [5] were based on all inspection data, therefore including the bridges that aren't regularly inspected. In this paper, only the inspected bridges are included since removing the bridges which aren't inspected at regular intervals will lead to the removal of anomalies of larger time in CSs due to irregular times between inspections. There are four condition states which are defined based on the BCI Average score at the time of inspection, as shown in [5]. Each of these condition states define the overall condition of the bridge where state 1 represents the as-new condition and state 4 represents failure of the bridge. Survival analysis was performed on the entire bridge stock

where the survival time is the time spent in each CS and the 'failure' is the transition of the bridge condition to a worse state.

In [5], several key bridge attributes including construction type, bridge function, single span or not and road class are investigated for their impact on survival in states 1,2 and 3. In Section 4, results are presented for other characteristics such as traffic, cumulative span and deck width (see Figure 1). Each of the graphs in Figure 1 show a right skew. For the traffic variable, this distribution is similar to that of the different road types. For example, motorway bridges make up 1.3% of the bridge stock and as expected the Very High traffic category is similar with 1.8%. Since approximately 78% of bridges are single span, the graphs shown in Figure 1(b)&(c) are as expected since the smaller the bridge the smaller the cumulative span and deck width.

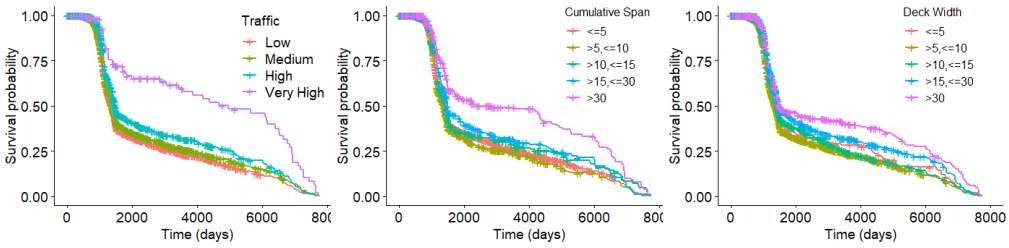


(a) Traffic (b) Cumulative Span (c) Deck Width

Figure 1: Bar charts of the bridge characteristics investigated in this paper: (a) Traffic; (b) Cumulative Span; (c) Deck Width

4 Results

Figure 2 shows the Kaplan-Meier (KM) curves for the time spent in CS 1 stratified by 3 characteristics which indicate a possible difference in survival.



(a) Traffic (b) Cumulative Span (c) Deck Width

Figure 2: KM survival curves for the time spent in CS 1 stratified by: (a) Traffic; (b) Cumulative Span; (c) Deck Width

In particular for the bridges where traffic is very high the survival probability is much higher than that of the bridges with lower traffic levels. Likewise, it can be seen that

as cumulative span and deck width increases the higher the probability of survival. The Wald's tests was carried out on these characteristics for state 1,2 and 3 (results are shown in Table 1). It is apparent that there is a significant difference in the traffic, span and deck widths for states 1 and 2. These variables are included in a Cox PH model and hazard ratios are shown in Table 2. If the hazard ratio is equal to 1 it represents no effect on the hazard of the event, a hazard ratio over 1 would indicate an increase in the hazard and below 1 would indicate a decrease in the hazard. A baseline bridge has the following characteristics: not masonry arch, not road over river, multi-span, unclassified road, low traffic levels, cumulative span and deck width less than 5m. It is important to note that on no occasion does the CIs for the hazard ratios include the value of 1.

	State 1	State 2	State 3
Traffic	9e-14	7e-04	0.5
Cumulative Span Categories	8e-10	0.06	0.001
Deck Width Categories	7e-13	0.01	0.6

Table 1: Wald's test p-values for each of the characteristics in CS 1,2 and 3.

	Hazard Ratio	CI for Hazard Ratio
Masonry Arch	1.29	[1.19,1.39]
Road Over River	1.16	[1.05, 1.28]
Road Class - Motorway	0.606	[0.378,0.973]

Table 2: A table showing a few key results from the Cox PH model.

5 Discussion and Conclusion

This paper shows the application of KM curves to determine the effect of bridge characteristics on the time spent in each CS before deteriorating to a worse CS. From Figure 2(a), a difference between the very high traffic levels and the other traffic levels is significant. This difference may be due to the very high traffic levels being on motorway bridges. These bridges are under a maintenance contract and their maintenance is very regular to ensure safety to the users who rely on these bridges for their daily commute. Table 2 shows the hazard ratios along with the confidence intervals. If this confidence interval contains 1 it would mean that there is no significant difference between the variable level and the baseline. Since the confidence intervals in Table 2 do not contain the value of 1 it would suggest that all of these results are significantly different than the baseline level. The hazard ratio for Road Class Motorway (Table 2) is 0.606 suggesting that there is a decrease in the hazard of deteriorating to a worse state from condition 1 to any worse state for a motorway bridge compared to a bridge on an unclassified road. From the results of this paper and [5], it can be concluded that BCI Average by itself is not sufficient to inform decisions regarding maintenance actions. The BCI Average score is based on the overall condition of all elements. At the time of inspection BCI Critical is also calculated, this value describes the condition of the elements which are of very high importance to the bridge. Therefore a point of further work would be to consider of combination the significant variables (shown in Tables 1 and 2) along with the BCI Average and BCI Critical in order to obtain a more better metric for informing decisions.

Acknowledgements: The authors would like to thank the Department for Infrastructure (DfI), Royal Academy of Engineering for the financial support and The Queen's University Belfast Leverage Studentship Scheme.

Bibliography

[1] Beng, S. S., & Matsumoto, T. (2012). Survival analysis on bridges for modeling bridge replacement and evaluating bridge performance. Structure and infrastructure engineering, 8(3), 251-268.

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

[3] Ng, S. K., & Moses, F. (1998). Bridge deterioration modeling using semi-Markov theory. A. A. Balkema Uitgevers B. V, Structural Safety and Reliability., 1, 113-120.

[4] Sobanjo, J. O. (2011). State transition probabilities in bridge deterioration based on Weibull sojourn times. Structure and Infrastructure Engineering, 7(10), 747-764.

[5] Stevens, N. A. et al. (2020) Identification of Bridge Key Performance Indicators Using Survival Analysis for Future Network-Wide Structural Health Monitoring. Sensors, 20(23), 6894.

[6] Stevens, N. A. et al. (2020). Conversion Of Legacy Inspection Data To Bridge Condition Index (BCI) To Establish Baseline Deterioration Condition History For Predictive Maintenance Models.

[7] Stevens, N. A. et al. (2021). Analysis of data for 6,978 bridges to inform a data strategy for predictive maintenance. IABMAS 2020.

[8] Tabatabai, H., Lee, C. W., & Tabatabai, M. A. (2015). Reliability of bridge decks in the United States. Bridge Structures, 11(3), 75-85.

Estimation of in vitro bactericidal potency based on colony counting method

Máté Szalai,[1*] **and Péter Kevei**[1]

[1] *Bolyai Institute, University of Szeged, Aradi vértanúk tere 1, 6720 Szeged, Hungary*

Abstract: To model the growth of a bacterial population in the presence of antibiotics we use the stochastic model from Bogdanov et al. [2]. We assume that bacterial cells either die or duplicate, with probabilities $p_0(c)$ and $p_2(c)$, where $p_2(c) = 1/(1 + \alpha c^\beta)$ for some α, β, where c stands for the antibiotic concentration. Using measurements based on colony counting method we obtain weakly consistent, asymptotically normal estimator both for (α, β) and for the minimal inhibitory concentration (MIC), a relevant parameter in pharmacology.

Keywords: Galton–Watson process, extinction probability, asymptotically normal estimator, MIC

AMS subject classification: 92C40, 60J80

1 Introduction

The correct estimation of bactericidal potency is a critical issue for the safe and proper use of antibiotics. In Bogdanov et al. [2] we worked out a Bienaymé–Galton–Watson branching model for the growth of the bacterial population, and we obtained weakly consistent asymptotically normal estimators for the relevant parameters when for the biological measurements quantitative PCR (qPCR) method is used. In [2] we found that the 2-parameter model fits very well to real biological data. In the present note we provide an estimator under the same model assumptions but for different biological data: we assume that the experimental data was obtained using colony counting method. The qPCR method measures the total bacterial genom, which is the total number of *dead and alive* bacterial cells multiplied by a constant. On the other hand, colony counting gives an estimator for the extinction probability. The basic experiment is the following. Originally, x_0 bacterial cells (e.g. *Escherichia coli*) are inoculated onto agar plates containing a series of antibiotic concentration, and after the incubation period all the viable colonies are enumerated, see e.g. Liu et al. [1].

As in [2] we assume that the bacterial population is homogeneous, in particular, there is no resistant type. Long-term evolution of bacterial populations with both *resistant* and *susceptible* types was investigated in several papers using deterministic models, see Svara and Rankin [4], Paterson et al. [3], and the references therein. Closest to our model is the deterministic model given by Liu et al. [1], where the biological measurements were obtained by colony counting. In [1] a deterministic expression for the number of colony forming units was obtained in terms of the antibiotic concentration.

*Corresponding author: szalaim@math.u-szeged.hu

Next we describe the mathematical model. We consider a simple Galton–Watson branching process where each bacterium either dies (leaves no offspring) or divides (leaves 2 offsprings) with respective concentration dependent probabilities $p_0 = p_0(c)$ and $p_2 = p_2(c) = 1 - p_0(c)$. Let $f(s) = f_c(s) = p_0 + p_2 s^2$ denote the offspring generating function and $m = m(c) = 2p_2(c)$ the offspring mean if the antibiotic concentration is c. The process starts with a single ancestor $X_{0;c} = 1$, and

$$X_{n+1;c} = \sum_{i=1}^{X_{n;c}} \xi_{i;c}^{(n)},$$

where $\{\xi_c, \xi_{i;c}^{(n)} : i \geq 1, n \geq 1\}$ are independent and identically distributed (iid) random variables with generating function f_c. We further assume that the offspring distribution is given by

$$p_2(c) = \frac{1}{1 + \alpha c^\beta}, \tag{1}$$

where $\alpha > 0$, $\beta > 0$ are unknown parameters. Note that as $m = 2p_2$ this is the same assumption as in [2]. Under this model the minimal inhibitory concentration (MIC), the smallest antibiotic concentration preventing bacterial growth, is the smallest c for which $m(c) = 1$, that is $\alpha^{-1/\beta}$.

If $m \leq 1$ then the process dies out almost surely, while if the process is supercritical, i.e. $m > 1$ then the probability of extinction is the smaller root of $f_c(q) = q$, which is in our setup

$$q(c) = \begin{cases} \frac{1-p_2(c)}{p_2(c)}, & \text{if } p_2(c) > 1/2, \\ 1, & \text{if } p_2(c) \leq 1/2. \end{cases} \tag{2}$$

2 Estimation of the parameters

Assume that the initial number of bacterial cells is x_0, that is we observe x_0 independent copies of the Galton–Watson process $(X_{n;c})$. Then the number Y_c of living colonies has binomial distribution with parameters x_0 and $1 - q(c)$. Therefore, the natural estimator for $q(c)$ is $\widehat{q}(c) = 1 - \frac{Y_c}{x_0}$. The law of large numbers and the central limit theorem implies that $\widehat{q}(c)$ is a weakly consistent estimator and as $x_0 \to \infty$

$$\frac{\sqrt{x_0}}{\sqrt{q(c)(1-q(c))}} \left(\widehat{q}(c) - q(c)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{3}$$

where $\xrightarrow{\mathcal{D}}$ stands for convergence in distribution.

From (2) we see that we can estimate $p_2(c)$ only if $q(c) < 1$, or equivalently $m(c) > 1$, in which case

$$\widehat{p_2}(c) = \frac{1}{1 + \widehat{q}(c)}. \tag{4}$$

We assume that the offspring mean as a function of c satisfies (1) for some unknown parameters $\alpha > 0$, $\beta > 0$. Rewriting (1)

$$\log \alpha + \beta \log c = \log\left(\frac{1}{p_2(c)} - 1\right).$$

Assume that we have measurements for $K \geq 2$ different concentrations $c_1 < c_2 < \ldots < c_K$, such that $m(c_K) > 1$. As in (4), we obtain the estimator $\widehat{p}_2(c_i)$ at different concentrations, from which, using simple least squares estimator we obtain the estimator

$$\widehat{\beta} = \frac{K\sum_{i=1}^K f_i\ell_i - \sum_{i=1}^K f_i L_1}{KL_2 - L_1^2},$$
$$\widehat{\alpha} = \exp\left\{\frac{\sum_{i=1}^K f_i - \widehat{\beta}L_1}{K}\right\},$$

where to ease notation we write

$$f_i = \log\left(\frac{1}{\widehat{p}_2(c_i)} - 1\right), \quad \ell_i = \log c_i,$$

and $L_1 = \sum_{i=1}^K \ell_i$, $L_2 = \sum_{i=1}^K \ell_i^2$. By the Cauchy–Schwarz inequality the denominator of $\widehat{\beta}$ is strictly positive for $K \geq 2$.

Under the assumption (1) the MIC equals $\vartheta = \alpha^{-1/\beta}$, therefore its natural estimator is

$$\widehat{\vartheta} = \widehat{\alpha}^{-1/\widehat{\beta}}.$$

Using (3), as in [2] we can prove that these estimators are asymptotically normal. Introduce the notation

$$k_i = \frac{p_2(c_i)}{1 - p_2(c_i)}\sqrt{q(c_i)(1 - q(c_i))}, \quad i = 1, 2, \ldots, K.$$

Proposition 1. *Assume that $c_1 < \ldots < c_K$ are given concentrations such that $m(c_K) > 1$. Then as $x_0 \to \infty$, $\widehat{\alpha}, \widehat{\beta}$, and $\widehat{\vartheta}$ are weakly consistent estimators of the corresponding quantities. Furthermore, as $x_0 \to \infty$*

$$\sqrt{x_0}(\widehat{\alpha} - \alpha, \widehat{\beta} - \beta) \xrightarrow{\mathcal{D}} (U, V),$$

where (U, V) is a two-dimensional normal random vector with mean 0 and covariance matrix $\begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{pmatrix}$, where

$$\sigma_\alpha^2 = \frac{\alpha^2}{(KL_2 - L_1^2)^2}\sum_{i=1}^K k_i^2(L_2 - L_1\ell_i)^2,$$
$$\sigma_{\alpha\beta} = \frac{\alpha}{(KL_2 - L_1^2)^2}\sum_{i=1}^K k_i^2(K\ell_i - L_1)(L_2 - L_1\ell_i),$$
$$\sigma_\beta^2 = \frac{1}{(KL_2 - L_1^2)^2}\sum_{i=1}^K k_i^2\,(K\ell_i - L_1)^2,$$

and $\sqrt{x_0}(\widehat{\vartheta} - \vartheta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_\vartheta^2)$ as $x_0 \to \infty$, with

$$\sigma_\vartheta^2 = \frac{\vartheta^2\,(\log\alpha)^2}{\beta^2(KL_2 - L_1^2)^2}\sum_{i=1}^K k_i^2\left(\frac{L_2 - L_1\ell_i}{\log\alpha} - \frac{K\ell_i - L_1}{\beta}\right)^2.$$

3 Simulation study

If $m(c_K) > 1$, then regardless of the fixed values $\mathbf{c} = (c_1, \ldots, c_K)$ the estimate $(\widehat{\alpha}, \widehat{\beta})$ is weakly consistent and asymptotically normal as $x_0 \to \infty$. However, the asymptotic variances in Proposition 1 do depend on the specific choice of $K \geq 2$ and the values $c_1 < \ldots < c_K$. Intuitively, it is clear that we should choose values for the concentrations where the derivative of m is large, that is m is close to 1, see Figure 1.

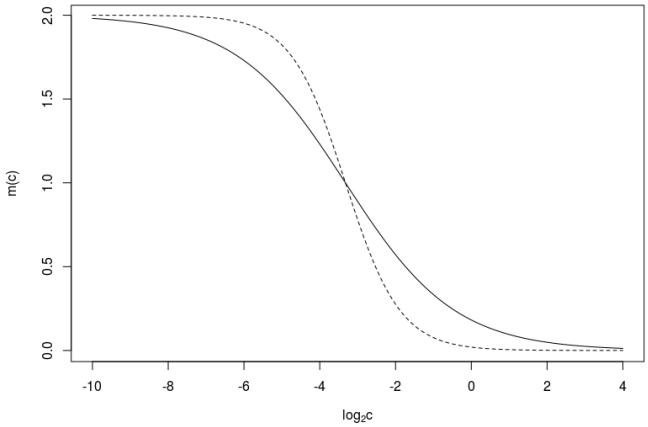


Figure 1: $m(c)$ in a logarithmic scale (solid $(\alpha, \beta) = (10, 1)$, dashed $(\alpha, \beta) = (100, 2)$)

As in [2] we compare two rather different biologically relevant scenarios: $(\alpha, \beta) = (10, 1)$ and $(\alpha, \beta) = (100, 2)$. In Figure 1 we see the mean function for these two cases. Note that in both cases $\vartheta = 0.1$. Table 1 contains the theoretical variances given in Proposition 1 for different choices of the concentrations. For the steeper function ($(\alpha, \beta) = (100, 2)$) the variances of α and β are significantly larger, however the variance of the MIC is of the same order. We also see that a wrong choice of the concentrations might result much larger variations. For $\mathbf{c}_3$ all the concentrations are small, the antibiotic does not have any effect, so we cannot make a good estimate from observations at these concentrations.

concentrations	σ^2_{10}	σ^2_1	$\sigma^2_{0.1}$	σ^2_{100}	σ^2_2	$\sigma^2_{0.1}$
$\mathbf{c}_1 = (2^{-7}, 2^{-4})$	2424	2.87	0.015	$2.98 \cdot 10^6$	38	0.0027
$\mathbf{c}_2 = (2^{-5}, 2^{-4.5}, 2^{-3.4})$	875	1.36	0.0016	$3.54 \cdot 10^5$	5.5	0.0014
$\mathbf{c}_3 = (2^{-9}, 2^{-8}, 2^{-7})$	$8.99 \cdot 10^4$	32	2.89	$3.84 \cdot 10^8$	1448	29

Table 1: Asymptotic variances for $(\alpha, \beta) = (10, 1)$ and $(\alpha, \beta) = (100, 2)$.

Choosing the right antibiotic concentration is important to get a good estimate. The larger variances above are not surprising, because in the present setup the estimator for the mean $m(c)$ works only for supercritical processes, that is for those c, for which $m(c) > 1$. That is we can sample only from the upper part of the mean function $m(c)$ in Figure 1. This is in sharp contrast to the situation treated in [2], where the total number of dead and alive bacteria was counted, and the estimator for the mean works for any c.

x_0	$\overline{\alpha}$	$\overline{\beta}$	$\overline{\vartheta}$	$\widehat{\sigma}^2_\alpha$	$\widehat{\sigma}_{\alpha,\beta}$	$\widehat{\sigma}^2_\beta$	$\widehat{\sigma}^2_\vartheta$
50	11.25	1.01	0.101	1464	41	1.3	0.002
100	10.79	1.01	0.1004	1349	43	1.48	0.0019
300	10.23	1.003	0.1	981	36.2	1.44	0.0018
500	10.17	1.003	0.1	931	34.9	1.34	0.0016
∞	10	1	0.1	875	34	1.36	0.0017

Table 2: Empirical mean and variances for $(\alpha, \beta) = (10, 1)$.

With $\alpha = 10$, $\beta = 1$ and concentration vector $\mathbf{c}_2$ we simulate the process as follows. For a given concentration c_k, $k = 1, \ldots, K$, we calculate $p_2(c_k)$ from (1). From each measurement we calculate the estimation $(\widehat{\alpha}, \widehat{\beta})$ as described in (2). We simulated the measurements 1000 times. The resulting means and empirical variances of $\sqrt{x_0}(\widehat{\alpha}-\alpha, \widehat{\beta}-\beta)$ and $\sqrt{x_0}(\widehat{\vartheta}-\vartheta)$ are given in Table 2. We see that even for small initial number of bacteria the empirical variances are close to the theoretical counterparts.

Bibliography

[1] Y. Q. Liu, Y. Z. Zhang, and P. J. Gao. *Novel concentration-killing curve method for estimation of bactericidal potency of antibiotics in an in vitro dynamic model.* Antimicrobial. Agents and chemotherapy 48(10):3884-3891, 2004.

[2] A. Bogdanov, P. Kevei, M. Szalai, and D. Virok. *Stochastic modeling of in vitro bactericidal potency.* https://arxiv.org/abs/2104.11525, 2021.

[3] I.K. Paterson, A. Hoyle, G. Ochoa, C. Baker-Austin, and N.G.H. Taylor. *Optimising Antibiotic Usage to Treat Bacterial Infections.* Sci. Rep. 6(37853), 2016.

[4] Fabian Svara, and Daniel J. Rankin. *The evolution of plasmid-carried antibiotic resistance.* BMC Evolutionary Biology, 11(130), 2011.

Modelling block maxima with the blended generalised extreme value distribution

Silius M. Vandeskog,[1*] **Sara Martino**[1] **and Daniela Castro-Camilo**[2]

[1]*Norwegian University of Science and Technology*
[2]*University of Glasgow*

Abstract: The blended generalised extreme value (bGEV) distribution has been proposed as an alternative to the generalised extreme value (GEV) distribution for modelling block maxima. Simulation studies are performed to analyse the performance of the bGEV distribution when the data follow the GEV distribution. We propose a new hierarchical two-step model for block maxima that borrows strength from the peaks over threshold method for less wasteful inference. Simulation studies are implemented to evaluate the performance of the two-step model. We find that the bGEV distribution is a promising alternative to the GEV distribution when modelling block maxima, and that the two-step model is able to improve inference by using more information from the available data.

Keywords: Extreme value theory, bGEV, INLA
AMS subject classification: 62P12, 62G32

1 Introduction

A well-known method to accurately estimate high quantiles is based on the so-called block maxima approach. The generalised extreme value (GEV) distribution is the only non-degenerate limit distribution of a standardised block maximum, meaning that it can be used as a model for the maximum of a large block of random variables [3]. Unfortunately, inference with the GEV distribution is somewhat wasteful, as only one observation per block is kept for analysis. Moreover, the GEV support depends on its parameters, which makes inference unstable and introduces artificial boundary restrictions on the data. This is particularly problematic in a covariate-dependent setting, which is common when dealing with spatially distributed data. In such frameworks, flexible modelling is usually introduced by allowing the model's parameters to vary according to the values of covariates [4]. The restrictive amount of data usually available using the block maxima approach makes it challenging to introduce complex models on multiple parameters. To cope with these issues, we propose a two-step approach that models the scale parameter of the GEV distribution using the peaks over threshold method, which, in some cases, uses extreme observations more efficiently than block-maxima. It has been suggested that the peaks over threshold technique is preferable over block maxima when the interest is in quantile estimation [1]. The opposite holds if the interest is in estimating return levels. By borrowing strength between both methods, we take advantage of the merits and improve

*Corresponding author: silius.m.vandeskog@ntnu.no

the pitfalls of both methods, thus allowing for less wasteful and more stable inference. For inference over block maxima data, we use the blended GEV (bGEV) distribution, recently proposed as an alternative to the GEV when the tail parameter is positive [2]. The bGEV distribution has infinite support and therefore avoids problems arising from parameter-dependent supports.

The main aims of this paper are to 1) test the performance of the bGEV in a simulation study to understand its properties, and 2) show the advantages of borrowing strength between the block maxima and the peaks over threshold approaches. Our proposed model belongs to the class of latent Gaussian models and therefore can be implemented using the `R-INLA` package [6], which allows fast and accurate inference. The remainder of the article is organised as follows. In Section 2 we provide details of the bGEV distribution, while in Section 3 we derive our two-step modelling approach. Section 4 presents the results of our simulation study. Conclusions are given in Section 5.

2 The bGEV distribution

The GEV distribution function is

$$F(x;\mu,\sigma,\xi) = \exp\{-(1+\xi(x-\mu)/\sigma)_+^{-1/\xi}\},$$

with $(a)_+ = \max(0,a)$ [3], which means that the support of F depends on its parameters. The bGEV distribution has been proposed as an alternative to the GEV when $\xi \geq 0$ [2]. It is defined as

$$H(x;\mu,\sigma,\xi,a,b) = F(x;\mu,\sigma,\xi)^{p(x;a,b)} G(x;\tilde{\mu},\tilde{\sigma})^{1-p(x;a,b)}, \tag{1}$$

where F is a GEV distribution with $\xi \geq 0$, G is a Gumbel distribution and the weight function $p(x;a,b)$ is the distribution function of a beta distribution that is zero for $x \leq a$ and one for $x \geq b$. The parameters $(\tilde{\mu},\tilde{\sigma})$ are injective functions of (μ,σ,ξ) that guarantee continuity in (1). The main advantage of the bGEV distribution is that it has infinite support, thus simplifying inference by not introducing artificial boundary restrictions. See [2] for more details about the distribution.

3 The two-step model

Let $y_i(\boldsymbol{s})$ denote the ith block maximum at location $\boldsymbol{s} \in \mathcal{S}$, where $\mathcal{S}$ is the area of interest. Assume that $y_i(\boldsymbol{s})$ follows a bGEV distribution with parameters $(\mu(\boldsymbol{s}),\sigma(\boldsymbol{s}),\xi)$. It is a common approach to let μ and σ vary in space while ξ is left constant to simplify inference [4]. In data-sparse situations, a large observation at a single location can be explained by a large tail parameter or a large scale parameter. In practice, if the model on $\sigma(\boldsymbol{s})$ is complex, this might cause identifiability issues between $\sigma(\boldsymbol{s})$ and ξ even though the parameters are identifiable in theory. To put a flexible model on the scale while avoiding any identifiability issues, we propose a two-step procedure that consists of first modelling $\sigma(\boldsymbol{s})$ using peaks over threshold data. Then the estimated $\sigma(\boldsymbol{s})$ is used to standardise the block maxima and fit a bGEV distribution where both scale and shape parameters are constant in space.

It is known that, for some large enough threshold $u(\boldsymbol{s})$, the distribution of an observation larger than $u(\boldsymbol{s})$ is given by a generalised Pareto distribution with tail parameter ξ and scale parameter $\tilde{\sigma}(\boldsymbol{s}) = \sigma(\boldsymbol{s}) + \xi(u(\boldsymbol{s}) - \mu(\boldsymbol{s}))$ [3]. We assume that the difference

$u(\boldsymbol{s}) - \mu(\boldsymbol{s})$ is proportional to the scale parameter $\sigma(\boldsymbol{s})$, since the tail parameter is constant. This assumption leads to a proportionality between the scale $\sigma(\boldsymbol{s})$ and the standard deviation of all observations larger than the threshold $u(\boldsymbol{s})$. Consequently, it is possible to model the spatial structure of the scale parameter independently of the location and tail parameter. Denote $\sigma(\boldsymbol{s}) = \sigma_0^* \cdot \sigma^*(\boldsymbol{s})$, with σ_0^* a standardising constant and $\sigma^*(\boldsymbol{s})$ the standard deviation of all observations larger than $u(\boldsymbol{s})$ at location $\boldsymbol{s}$. Dividing the block maxima by $\sigma^*(\boldsymbol{s})$ gives

$$y_i^*(\boldsymbol{s}) = y_i(\boldsymbol{s})/\sigma^*(\boldsymbol{s}) \sim \text{bGEV}(\mu^*(\boldsymbol{s}), \sigma_0^*, \xi), \tag{2}$$

where $\mu^*(\boldsymbol{s}) = \mu(\boldsymbol{s})/\sigma^*(\boldsymbol{s})$. Inference is considerably easier when modelling $(\mu^*(\boldsymbol{s}), \sigma_0^*, \xi)$ instead of $(\mu(\boldsymbol{s}), \sigma(\boldsymbol{s}), \xi)$, since both σ_0^* and ξ are constants.

4 Simulation studies

The bGEV distribution

The performance of the bGEV distribution is evaluated when the true distribution of the data is a GEV distribution. We draw $n \in \{10, 50, 100, 500, \ldots, 10000\}$ samples from a GEV distribution with parameters (μ_i, σ_i, ξ_i), $i = 1, 2, \ldots, 500$. Inference is performed using `R-INLA`. Point and interval estimates are provided for the GEV parameters and for different return levels. Point estimates are evaluated using the mean squared error (MSE). Coverage probabilities for the 95% credible intervals are displayed in Table 1. For small values of n, the interval estimates of the bGEV distribution correspond well with the true values from the GEV distribution. However, as n increases, the coverage probabilities deviates away from 95%. We find that the bGEV distribution tends to overestimate the tail parameter and underestimate the scale parameter of the GEV distribution. This is not surprising as we expect a misspecification error from using an incorrect likelihood. However, in a real-world setting, it is extremely rare to observe 1000 or more block maxima. The block maxima are also not perfectly GEV distributed, as the blocks must be of finite size. Consequently, misspecification error will also be present when using the GEV distribution for modelling block maxima.

The logarithm of the MSE is displayed in Table 2. It seems to decrease almost linearly with the logarithm of n. This means that the point estimates are getting closer to the truth even though the interval estimates are not entirely correct. The results indicate that the bGEV distribution can be a viable alternative to the GEV for modelling block maxima.

Table 1: Estimated probability of covering the true GEV parameters and return levels inside the 95% credible intervals of the bGEV distribution. The T block return level is displayed as "RT".

n	μ	σ	ξ	R10	R50	R100	R500
50	92.4%	92.4%	85.0%	88.9%	86.8%	85.0%	86.0%
100	93.4%	92.2%	92.2%	91.8%	91.2%	92.0%	92.0%
1000	96.4%	89.2%	94.8%	94.0%	95.6%	96.0%	96.0%
10000	99.8%	47.3%	65.5%	95.4%	93.8%	90.4%	85.8%

Table 2: Log-MSE between the true GEV parameters and return levels and those estimated with the bGEV distribution. The T block return level is displayed as "RT".

n	μ	σ	ξ	R10	R50	R100	R500
100	0.84	0.49	−4.59	3.23	5.27	6.87	7.84
1000	−1.53	−1.72	−6.59	1.05	3.32	5.08	6.16
10000	−3.27	−2.97	−7.91	−1.17	1.36	3.49	4.73

The two-step model

We simulate 250 locations $\boldsymbol{s}_i,\ i = 1, \ldots, 250$, together with k spatially varying covariates $\mathbf{x}(\boldsymbol{s})$. At each location $\boldsymbol{s}$, we draw $24 \cdot 365 \cdot n(\boldsymbol{s})$ observations from a Fréchet distribution with parameters μ, ξ and $\sigma(\boldsymbol{s}) = \exp(\mathbf{x}(\boldsymbol{s})^T\boldsymbol{\beta})$, which represents $n(\boldsymbol{s})$ years of hourly observations. The numbers $n(\boldsymbol{s})$ are drawn randomly so the total number of block maxima is $\sum_{i=1}^{250} n(\boldsymbol{s}_i) = 1500$. Block maxima are then computed at all locations. We use 200 locations for parameter estimation and the remaining 50 as test data. This simulation procedure is repeated 300 times. Each time, the number of covariates k is drawn randomly between 1 and 4, and the values of $\mathbf{x}(\boldsymbol{s})$, $\boldsymbol{\beta}$, μ and ξ are changed. For estimation of $\sigma^*(\boldsymbol{s})$, the threshold $u(\boldsymbol{s})$ is set equal to the 80% empirical quantile of all observations at location $\boldsymbol{s}$. We place a linear Gaussian model on $\log(\sigma^*(\boldsymbol{s}))$ for estimation at locations with no available observations. Uncertainty is propagated by drawing 100 samples from the posterior of $\log(\sigma^*(\boldsymbol{s}))$, and estimating $(\mu^*(\boldsymbol{s}), \sigma_0^*, \sigma)$ for each of the 100 samples. The two-step model is compared with a model where all bGEV parameters are estimated jointly (the *joint model* henceforth), using `R-INLA`. Comparison is performed with the expected value of the threshold weighted continuous ranked probability score (twCRPS, [5]) with a quantile weight function $w(p) = I(p \geq 0.9)$, meaning that we only focus on the performance in the right tail.

The mean expected twCRPS over all 300 trials for the joint model is 0.9572 while the two-step model achieves a mean score of 0.9570. This difference might not seem considerable, but non-parametric bootstrapping shows that the difference between the two-step twCRPS and the joint twCRPS is significantly different from zero at a 5% level. This is an impressive result because the data in this simulation study is drawn directly from the joint model, and one would therefore expect it to perform well. The fact that the two-step model is able to perform better shows that the peaks over threshold data can be used for improving inference for the block maxima method. Comparing the two models in a more complex setting is difficult, as `R-INLA` is unable to place more flexible models on the scale parameter of the joint model.

5 Conclusion

The bGEV distribution allows for faster and simpler inference than the GEV distribution, and performs well in estimating GEV parameters and return levels when little data is available. Interval estimates are not correct because of misspecification error, but this will also be present for the GEV distribution in a real-world setting with finite block sizes.

The two-step model allows for less wasteful and more stable inference in situations where little data is available, and one aims to place a complex model on the scale parameter. Even in a simple setting where the log-scale is a linear combination of covariates, the two-

step model is able to improve the prediction skill over the joint model. We expect the performance improvement to be even more considerable as the complexity of the scale parameter increases.

Bibliography

[1] A. Buecher and C. Zhou. A horse racing between the block maxima method and the peak-over-threshold approach. *Statistical Science*, 2020.

[2] D. Castro-Camilo. *Blended GEV: a tutorial using R-INLA*, 2020. R vignette "bGEV-tutorial" from the R package INLA.

[3] S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001.

[4] D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840, 2007.

[5] T. Gneiting and R. Ranjan. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011.

[6] H. Rue, A. Riebler, S.H. Sørbye, J.B. Illian, D.P. Simpson, and F.K. Lindgren. Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4(1):395–421, 2017.

Modeling the superposition of dependent binary signals with hidden Markov models

Laura Jula Vanegas[1*]

[1]*Institute for Mathematical Stochastic, University of Goettingen, Germany*

Abstract: We give a methodological framework for the case of dependent superimposed binary Markov chains. The superposition may not be directly observable, in a Hidden Markov model (HMM) sense. For that purpose, we introduce a class of multidimensional Markov chains where full recovery of the dependency structure is possible. One of the properties this class will have is that the "sum" process is again Markov. This allow us to use standard tools for HMM estimation with some modifications.

Keywords: Hidden Markov models, vector norm dependency, permutation invariance, lumping property, aggregated data.

AMS subject classification: 62M05, 62P10, 62H12

1 Introduction

HMMs are a widely used tool for the modeling of non directly observable Markov chains, for example to model current recordings of ion channels in the cell [3]. Recently, there is a growing interest in the multivariate case [2]. Note that if the entries of X_k are independent, the problem can be solved with standard tools [5]. However, the situation becomes more complex once independence can not be assumed and the complete multidimensional vector is not observable. This becomes even harder when only the superposition of signals can be detected. In an application analyzed in [7], we find the need to model such a situation, where only the total measure of a piece of membrane with multiple dependent ion channels can be captured.

2 Setting

Let $X^{(1)}, \dots, X^{(\ell)}$ be ℓ homogeneous binary Markov chains (i.e., the state space is $\{0, 1\}$) that are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define an ℓ-dimensional homogeneous Markov chain $(X_k)_{k\in\mathbb{N}}$ with $X_k = (X_k^{(1)}, \dots, X_k^{(\ell)})^T$. Note that the ℓ binary Markov chains may not be independent, and, therefore, the structure of the transition matrix $M \in \mathbb{R}^{2^\ell \times 2^\ell}$ of the multidimensional chain encodes the dependency structure between the binary chains. This matrix will be an object of interest in this paper.

*Corresponding author: ljulava@mathematik.uni-goettingen.de

Moreover, we assume that only the superposition of the individual Markov chains can be observed. Therefore, our main object of interest is the "sum" process $(S_k)_{k\in\mathbb{N}}$ on the finite state space $[\ell\,] := \{0,\dots,\ell\}$ given by

$$S_k := \sum_{j=1}^{\ell} X_k^{(j)}.$$

This process may be interesting for some applications due to the fact that it can be interpreted as counting how many ones are in a given time point $k \in \mathbb{N}$ in a multidimensional chain.

Note that the only by observing a path of $(S_k)_{k\in\mathbb{N}}$ it is generally not possible to recover the individual paths of . Even worse, due to the dependency between the chains, the process $(S_k)_{k\in\mathbb{N}}$ may not necessarily be a Markov chain itself. This will be explored in Section 3.

Moreover, assume that the sum process $(S_k)_{k\in\mathbb{N}}$ is itself not observable and can be modeled as a HMM with observable sequence $(Y_k)_{k\in\mathbb{N}}$. For example through Gaussian noise with mean and variance depending on the sum,

$$Y_k = \mu(S_k) + \sigma(S_k)\xi, \quad k \in \mathbb{N},$$

for a standard normally distributed real-valued random variable ξ, $\mu(S_k) \in \mathbb{R}$, and $\sigma(S_k) \in (0,\infty)$ for $k \in \mathbb{N}$.

The aim of this paper is to provide a methodological framework to estimate the transition matrix M of the multidimensional process (and therefore the dependency structure of the binary chains) given observations $(Y_k)_{k\in\mathbb{N}}$ as described before. To do this, we first introduce a set of properties to ensure that the sum process is a Markov chain and that M can be identified from the transition matrix of $(S_k)_{k\in\mathbb{N}}$ (Section 3). Then, we describe how to estimate M from observations $(Y_k)_{k\in\mathbb{N}}$ by modifying standard tools for HMMs (Section 3).

3 Methodological framework

As we mentioned before, due to the dependency between the binary chains, the process $(S_k)_{k\in\mathbb{N}}$ may not be a Markov chain. However, *the lumping property* is a sufficient condition on the multidimensional process to ensure that the sum is again Markovian (see [6]).

Definition 1 (Lumping property)**.** We say that satisfies the lumping property if for any $k \in \mathbb{N}$, $j \in \{1,\dots,\ell\}$ and $x, y \in \{0,1\}^\ell$ such that $\sum_{i=1}^{\ell} x_i = \sum_{i=1}^{\ell} y_i$, it holds

$$\mathbb{P}(S_{k+1} = j \mid X_k = x) = \mathbb{P}(S_{k+1} = j \mid X_k = y),$$

whenever $\mathbb{P}(X_k = y) \cdot \mathbb{P}(X_k = x) > 0$.

If has this property, then we can modify standard HMM tools, such as the Baum-Welch algorithm, to estimate the transition matrix of $(S_k)_{k\in\mathbb{N}}$ from observations $(Y_k)_{k\in\mathbb{N}}$.

VND Markov Chains

We start by defining a class of Markov chains that fulfill the lumping property. This class may seem abstract at first, however, we provide an easy equivalent characterization that show its intuitiveness and practical applications. For $x \in \{0,1\}^\ell$, let $x = (x^{(1)},\dots,x^{(\ell)})^T$ and define the 1-norm by $\|x\|_1 := \sum_{i=1}^{\ell} |x^{(i)}|$.

Definition 2 (Vector Norm Dependency)**.** A multidimensional Markov chain on $\{0,1\}^\ell$ with transition matrix $M^{(\mathrm{VND})} = (m_{x,y}^{(\mathrm{VND})})_{x,y\in\{0,1\}^\ell}$ is called vector norm dependent (VND) if for all $i \in \{1,\dots,\ell\}$, $r \in [\ell]$, $k \in \mathbb{N}_0$ and $b \in \{0,1\}$, the expression

$$\mathbb{P}\left(X_{k+1}^{(i)} = b \mid X_k^{(i)} = b, \|X_k\|_1 = r\right) \tag{1}$$

is independent of i and k and

$$m_{x,y}^{(\mathrm{VND})} = \prod_{i=1}^{\ell} \mathbb{P}\left(X_{k+1}^{(i)} = y^{(i)} \mid X_k^{(i)} = x^{(i)}, \|X_k\|_1 = \|x\|_1\right), \tag{2}$$

for any $x = (x^{(1)},\dots,x^{(\ell)})^T$ and $y = (y^{(1)},\dots,y^{(\ell)})^T$ with $x, y \in \{0,1\}^\ell$.

From Definition 2 we can easily observe that the transition matrix $M^{(\mathrm{VND})}$ is determined by 2ℓ parameters, which is a significant reduction from the $2^{2\ell}$ of the general case. To understand the logic behind this class of Markov chains, we introduce two properties.

Definition 3 (Permutation invariance)**.** We call a vector Markov chain *permutation invariant* if for any $k \in \mathbb{N}$, $x, y \in \{0,1\}^\ell$, and any permutation matrix $P \in \{0,1\}^{\ell\times\ell}$, it holds

$$\mathbb{P}(X_{k+1} = y \mid X_k = x) = \mathbb{P}(X_{k+1} = Py \mid X_k = Px).$$

Note that a Markov chain that has the permutation invariance property, has also the lumping property. In practical terms, this property says that we can relabel the binary chains without changing the permutation matrix M.

Definition 4 (Conditional independence)**.** We call conditionally independent (w.r.t. the past), if

$$\mathbb{P}(X_{k+1} = y \mid X_k = x) = \prod_{i=1}^{\ell} \mathbb{P}(X_{k+1}^{(i)} = y^{(i)} \mid X_k = x),$$

for any $k \in \mathbb{N}_0$ and for all $x, y \in$ where $y = (y^{(1)},\dots,y^{(\ell)})^T$.

This property says that the dependency between channels is not instantaneous but is only possible through the state at the previous time point. With this two properties, we can state the characterization theorem for VND Markov chains.

Theorem 1 (Characterization of VND Markov chains)**.** *For a vector Markov chain assume that the initial distribution is permutation invariant. Then, the following statements are equivalent:*

1. *The Markov chain is vector norm dependent;*
2. *The Markov chain is permutation invariant and conditional independent.*

This characterization provides a more intuitive and practical way of thinking about VND Markov chains and when to apply this framework in real data applications. Moreover, in [7], we show that for a VND Markov chain model with corresponding transition matrix $M^{(\mathrm{VND})}$, we can essentially uniquely recover $M^{(\mathrm{VND})}$ from the transition matrix of $(S_k)_{k\in\mathbb{N}}$.

VND in a HMM setting

Let the measurable space $(\Omega, \mathcal{F})$ be equipped with a family of probability measures $(\mathbb{P}_\theta)_{\theta\in\Theta}$, where $\Theta \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$. Let $\ell \in \mathbb{N}$ and $(S_k, Y_k)_{k\in\mathbb{N}}$ be a bivariate stochastic process as described in Section 2, such that $(Y_k)_{k\in\mathbb{N}}$, conditioned on $(S_k)_{k\in\mathbb{N}}$, is a real-valued, independent sequence of random variables. We say that $(S_k, Y_k)_{k\in\mathbb{N}}$ is a (homogeneous) VND-HMM if addtionaly $(S_k)_{k\in\mathbb{N}}$ comes from a VND Markov chain.

Since VND Markov chains are permutation invariant, and therefore, fulfill the lumping property, $(S_k)_{k\in\mathbb{N}}$ is a Markov chain. Let Q be the transition matrix of the sum process $(S_k)_{k\in\mathbb{N}}$. Then, we can use the Baum-Welch algorithm [1] to estimate Q. Moreover, this algorithm can be modified so we include the parametrization of Q with respect to the 2ℓ parameters as in (1).

We provide a brief discussion of the algorithm. Let $s_{1:K} \in \{1, \dots, \ell\}^K$ and $y_{1:K} \in \mathbb{R}^K$ be the paths of $(S_k)_{k=1,\dots,K}$ and $(Y_k)_{k=1,\dots,K}$ respectively. We assume that the emission distribution belongs to a parametric class determined by densities $\{g_\theta\}_{\Theta_E}$. The complete log-likelihood function is given by

$$\ell(\theta, s_{1:K}, y_{1:K}) = \log \pi^{(s_1)} + \sum_{k=1}^{K-1} \log q^{(\mathrm{VND})}_{s_k, s_{k+1}}(\theta_H) + \sum_{k=1}^{K} \log g_{\theta_E}(y_k, s_k). \tag{3}$$

for parameters $\theta = (\theta_H, \theta_E) \in \Theta$ that parametrize the transition matrix Q and the emission distribution g_{θ_E} respectively. Note that only in the second term of (3) the parametrized components of the transition matrix $Q^{(\mathrm{VND})}(\theta_H)$ appear, in contrast to the general case. This log-likelihood function can be used to implement a Expectation-Maximization kind of algorithm. However, in contrast to classical cases, the maximization step does not have a closed form, and we use instead a least squares approximation. This is due to the convoluted form of the parametrization of Q(see [7]).

The implementation of an R package for simulation and estimation in the VND model can be found at https://github.com/ljvanegas/VND.

4 Conclusion and Application

This framework can be applied to multiple applications where we have super-imposed binary chains, meaning that we only get a chance to see the number of "ones". As long as this Markov chains fulfill the permutation invariance and conditional independence properties, we can recover the original dependency structure. Moreover, if there is noise on top in a HMM way (for example by summing Gaussian noise), a implemented algorithm gives us a good estimate of the structural properties of the Markov chains. In [7], we apply this framework to real data obtained from ion channels in cardiac cells.

Acknowledgements: Your acknowledgements. This text is based in joint work with Benjamin Eltzner, Daniel Rudolf, and Axel Munk that can be found in [7]. The author acknowledge support of the DFG SFB 803 project Z02 and the DFG Cluster of Excellence 2067 MBExC.

Bibliography

[1] Baum, L, E., Petrie, T., Soules, G., Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Ann. Math. Stat* 41(1), 164-171.

[2] Bielecki T., Jakubowski J., Nieweglowski, M. (2013). Intricacies of dependence between components of multivariate Markov chains: weak Markov consistency and weak Markov copulae. *Electron. J. Probab.* 18, 1-21.

[3] Chung S.H., Anderson, O. S., Krishnamurthy, V, V. (2007). *Biological Membrane Ion Channels: Dynamics, Structure, and Applications*. Springer-Verlag.

[4] Chung S.H., Kennedy R. A. (1996). Coupled Markov chain model: Characterization of membrane channel currents with multiple conductance sublevels as partially coupled elementary pores. *Math. Biosci.* 133(2), 111–137.

[5] Dabrowski, A. R., McDonald, D. (1992). Statistical Analysis of Multiple Ion Channel Data. *Ann. Stat.* 20(3), 1180-1202.

[6] Kemeny, J. G., Snell, J. L. (1976). *Finite Markov chains: With a new appendix "generalization of a fundamental matrix"*. Springer-Verlag, Undergraduate Texts in Mathematics.

[7] Vanegas L.J., Eltzner B., Rudolf D., Dura M., Lehnart S. E., Munk, A. (2021). Analyzing cross-talk between superimposed signals: Vector norm dependent hidden Markov models and applications *Submitted*. https://arxiv.org/abs/2103.06071

Clustering based on multivariate mixed type longitudinal data with an application to the EU-SILC database

Jan Vávra,[1*] **and Arnošt Komárek**[1]

[1]*Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic*

Abstract: We present a statistical model for joint modelling of several mixed-type longitudinal outcomes, while performing unsupervised clustering with respect to different patterns. Method is demonstrated on the EU-SILC dataset consisting of Czech households followed in a time span 2005 – 2018.

Keywords: Multivariate longitudinal data, Mixed type outcome, Model based clustering, EU-SILC

AMS subject classification: 62H30

1 Introduction

The European Union Statistics on Income and Living Conditions project (EU-SILC, [3]) is an instrument with the goal to collect timely and comparable cross-sectional and *longitudinal multidimensional* data on income, poverty, social exclusion and living conditions. Data are obtained via questionnaires leading to outcome variables of diverse nature: *numeric* (e.g., income), *binary* (e.g., affordability of paying unexpected expenses) and *ordinal* (e.g., level of ability to make ends meet). It is our primary aim to use such longitudinally gathered outcomes towards segmentation of households according to typical patterns of their temporal evolution.

To this end, we propose a statistical model capable of joint modelling of longitudinal outcomes of diverse nature (*numeric, binary, ordinal*) while taking potential dependencies among different outcomes obtained at each occasion into account. Consequently, we use the model within a Bayesian model based clustering (MBC) procedure to perform unsupervised clustering of study units (households) into groups whose characteristics are not known in advance. Finally, we present some of the results of use of this methodology on the Czech subset of the EU-SILC dataset.

2 Joint model for mixed type longitudinal data

Let all the outcomes associated with unit (household) $i = 1, \ldots, n$ denote by $\mathbf{Y}_i$ consisting of $Y_{ij}^r, j = 1, \ldots, n_i, r = 1, \ldots, R$, that stands for jth observation of outcome r.

*Corresponding author: vavraj@karlin.mff.cuni.cz

Throughout the whole paper we will model the distribution of outcomes given a set of covariates (time, level of urbanisation, ...), which accompanies each individual observation. Specifically, we suppose that the distribution of Y_{ij}^r depends on the predictor $\eta_{ij}^r = \mathbf{X}_{ijr}^\top \boldsymbol{\beta}_r + \mathbf{Z}_{ijr}^\top \mathbf{b}_i^r$ consisting of a fixed part $\mathbf{X}_{ijr}^\top \boldsymbol{\beta}_r$ and a random part $\mathbf{Z}_{ijr}^\top \mathbf{b}_i^r$, where $\mathbf{X}_{ijr}$ and $\mathbf{Z}_{ijr}$ are formed from available covariates and their structure may differ across all $r = 1, \ldots, R$. Parameter $\boldsymbol{\beta}_r$ then denotes fixed effects for outcome r while $\mathbf{b}_i^r, i = 1, \ldots, n$ are (random) coefficients specific for each unit.

If the rth longitudinal outcome is of either ordinal (of L_r categories) or binary ($L_r = 2$) nature, we take a natural thresholding approach (see [2]) and assume each Y_{ij}^r to be determined by corresponding latent continuous variable $Y_{ij}^{\star,r}$. Falling into one of the intervals given by the set of ordered thresholds $\boldsymbol{\gamma}^r : -\infty = \gamma_0^r < \gamma_1^r < \cdots < \gamma_{L_r}^r = \infty$, that is, $\gamma_{l-1}^r < Y_{ij}^{\star,r} \leq \gamma_l^r$, is equivalent to attaining category $l = 1, \ldots, L^r$ by the observed Y_{ij}^r. We fix the first γ_1^r for identifiability purposes, in particular, binary outcomes do not require estimation of any thresholds.

In case of numeric outcome r we also introduce the notation $Y_{ij}^{\star,r} = Y_{ij}^r$, so that $Y_{ij}^{\star,r}$ now represents numeric (but possibly latent) value for any $r = 1, \ldots, R$. For each $Y_{ij}^{\star,r}$ we then suppose classical linear mixed model (LMM, Laird and Ware [5]), that is, $Y_{ij}^{\star,r} \big| \mathbf{b}_i^r \sim N\left(\eta_{ij}^r, \tau_r^{-1}\right)$ independently for all r, i, j, where τ_r is the precision parameter (reciprocal of variance) of normal distribution for rth outcome, which had to be fixed in case of categorical outcomes for identifiability purposes.

We stand by the classical assumption of centered normal distribution for the random effects, however, we first create a collection $\mathbf{b}_i$ of *all* random effects, i.e. $\mathbf{b}_i = \{\mathbf{b}_i^r, r = 1, \ldots, R\}$. By assumption of centered multivariate normal distribution with a completely general covariance matrix $\boldsymbol{\Sigma}$ for $\mathbf{b}_i$ independently for all $i = 1, \ldots, n$ we incorporate possible dependencies among the outcomes into our model since the distribution of $\mathbf{Y}_i^\star$ unconditioned by $\mathbf{b}_i$ loses its independence structure. Nevertheless, we still can take advantage of the independence structure while conditioning by $\mathbf{b}_i$ in the estimation process.

3 Model Based Clustering

Banfield and Raftery [1] were first to introduce the method of Model Based Clustering (MBC) that in general searches for G mutually distinguishable groups members of which share the same characteristics of observed outcomes. In particular, assume a probability distribution for outcomes $\mathbf{Y}_i$ within gth cluster described by a probability density function $h(\mathbf{Y}_i; \boldsymbol{\psi}^g)$ depending on an unknown set of parameters $\boldsymbol{\psi}^g$. Then the overall distribution of $\mathbf{Y}_i$ is given by a mixture $f(\mathbf{Y}_i; \boldsymbol{\theta}) = \sum_{g=1}^G w_g \cdot h(\mathbf{Y}_i; \boldsymbol{\psi}^g)$, where $\boldsymbol{\theta}$ stands for the set of all unknown parameters $\boldsymbol{\psi}^g$ and marginal cluster probabilities $0 < w_g < 1, g = 1, \ldots, G, w_1 + \cdots + w_G = 1$.

The model for $\mathbf{Y}_i$ from previous section is described by a collection of unknown parameters $\boldsymbol{\psi} = \{\boldsymbol{\beta}_r, \tau_r, \boldsymbol{\gamma}_r, \boldsymbol{\Sigma};\ r = 1, \ldots, R\}$. The corresponding probability distribution function originates from integrating all latent random variables ($\mathbf{Y}_i^\star$ and $\mathbf{b}_i$) out of their joint distribution:

$$h(\mathbf{Y}_i; \boldsymbol{\psi}) = \int\int \underbrace{t(\mathbf{Y}_i|\mathbf{Y}_i^\star)}_{\text{thresholding}} \cdot \underbrace{p(\mathbf{Y}_i^\star|\mathbf{b}_i)}_{\text{LMM}} \cdot \underbrace{p(\mathbf{b}_i)}_{N(\mathbf{0},\boldsymbol{\Sigma})} \, \mathrm{d}\mathbf{b}_i \mathrm{d}\mathbf{Y}_i^\star. \tag{1}$$

Each cluster then falls into the same family of probability distributions and is distinguished by a different set of $\boldsymbol{\psi}^g$ parameters. If all parameters are cluster-specific,

$\boldsymbol{\theta} = \{w_g, \boldsymbol{\beta}_r^g, \tau_r^g, \boldsymbol{\gamma}_r^g, \boldsymbol{\Sigma}^g;\ g = 1, \ldots, G,\ r = 1, \ldots, R\}$ denotes the final set of all unknown parameters in our model.

We can decompose the mixture distribution f into the conditional distribution given cluster $h(\mathbf{Y}_i; \boldsymbol{\psi}^g)$ and the marginal distribution of belonging to one of the groups, which is described by $P(U_i = g) = w_g$, where U_i are cluster allocation indicators that take value g in $\{1, \ldots, G\}$ whenever the ith unit falls into cluster g. Using Bayes theorem we obtain the probability of falling into a cluster given the data we have at our disposal:

$$P\left(U_i = g | \mathbf{Y}_i; \boldsymbol{\theta}\right) \propto w_g \cdot h(\mathbf{Y}_i; \boldsymbol{\psi}^g), \tag{2}$$

which in our case involves the double integration (1). Random effects $\mathbf{b}_i$ can be integrated out relatively easily due to the assumption of normal distribution over both latent outcomes and random effects. Then the part corresponding to observed numeric outcomes survives the other integration, while the rest remains and leads to the truncated multivariate normal probabilities. Their computation for higher dimension, which in our case is a product of n_i and the number of considered categorical outcomes, becomes problematic and is solved numerically using algorithm by Genz [4].

Due to the complexity of our model, traditional methods like maximum likelihood would be problematic to perform. Fortunately, our model can be easily translated to a Bayesian setting by assigning some uninformative prior over each element of $\boldsymbol{\theta}$. The inference about $\boldsymbol{\theta}$ is then based on its posterior distribution where we enrich the prior by the information provided by the observed data. Despite the natural choice of distributions we are not able to express the posterior distribution in a closed form, hence we need to resort to Markov Chain Monte Carlo (MCMC) methods.

By suitable choice of the prior distributions we ensure that the full-conditional distributions of each of the parameters or latent elements fall into well-known families, and hence are easily to be sampled from. This allows us to follow Gibbs sampling scheme where we sample each parameter from its full-conditional distribution given the last known values of other parameters. Generated values of $\boldsymbol{\theta}$ are used for estimation of the posterior distribution of individual parameters as well as the clustering probabilities (2).

4 Application

EU-SILC dataset is gathered annually by a 4-year rotational panel - each year a quarter of households is replaced by a set of new ones. Between the years 2005 and 2018 this study followed $n = 23\,360$ Czech households for exactly $n_i = 4$ consecutive years. The "Equivalised total disposable income" and the "Lowest income to make ends meet" were chosen as the numeric outcomes in a log-scale. The "Financial burden of total housing cost" and the "Ability to make ends meet" are outcomes of 3 and 6 ordered categories, respectively. Considered binary outcomes are Yes or No questions: "Can you afford annually a week holiday away from home?" and "Do you have a capacity to face unexpected financial expenses?".

The primary covariate - time corresponding to the year in which the household was interviewed - was parametrized by quadratic splines with one inner knot to allow for a change in evolution after the financial crisis. Fixed part of the predictor was extended by the equivalised household size (number of household members weighted by age), the highest education level attained by a member of the household (divided into 5 categories from Primary to Tertiary), the urbanisation level of the locality (we separate the capital city Prague from other 3 categories divided by population density) and other indicators like the presence of a baby or a student in the household. The random part of the model was left solely to the intercept term for each outcome.

Selected results are depicted in Figure 1, which shows cluster-specific mean evolutions of four chosen (latent) numeric outcomes. Here the blue cluster represents households requiring high income to pay for their usual expenses. One way they can achieve it better than the other two clusters is through higher levels of education. They are also more likely to afford a week holiday abroad. However, these households struggle with housing cost the more people live there. On the other hand, the red cluster represents the less fortunate households probably effected by consequences of the financial crisis since estimated curves in time begin to decline. Almost a half of the households fall into the green group of fairly average behaviour.

5 Conclusions

In this paper, we proposed a statistical model for joint modelling of several mixed-type longitudinal outcomes, while clustering with respect to different patterns. Proposed method was subjected to an extensive simulation study that covered different number of latent clusters G, structures of random effects and differences among clusters. In some circumstances, our procedure did not quite reach the true values of underlying parameters for low $n = 100$. Nevertheless, the performance of our estimators remarkably improved with an increased number of units n, already for $n = 1\,000$ it correctly classified more than 80% of units in the vast majority of scenarios.

However, the proposed Gibbs sampling was inefficient for threshold parameters $\boldsymbol{\gamma}_r$,

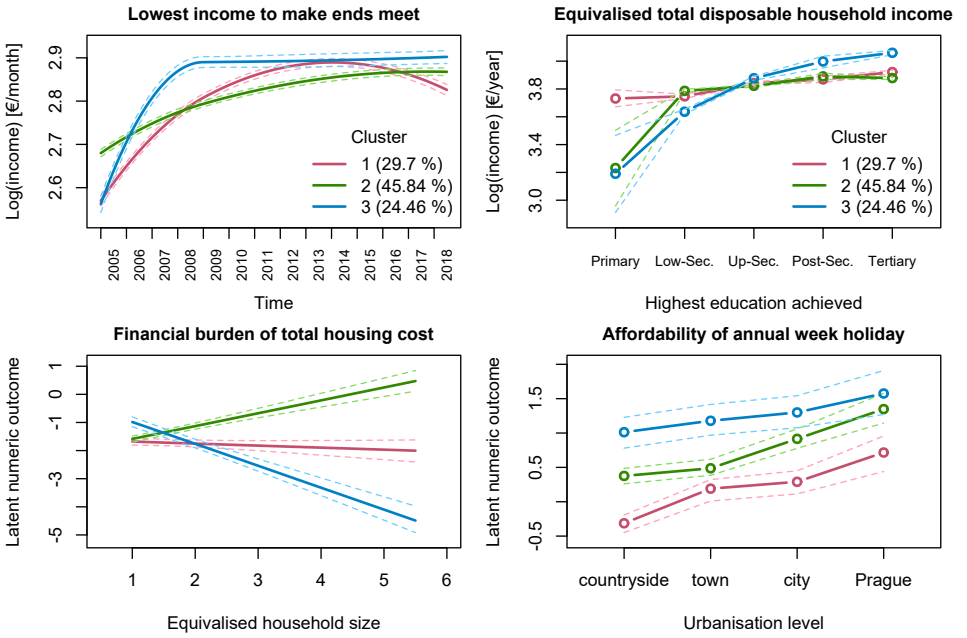


Figure 1: Cluster-specific dependence of several (latent) numeric outcomes on chosen covariates.

resulting in a very slow convergence to posterior distribution. Therefore, we started to develop analogous model that uses Generalized LMM for categorical outcomes which removes the necessity of latent numeric outcomes and allows to model even general (non-ordered) categorical outcomes. However, we cannot make use of the conjugacy of distributions for fixed and random effects like we did here. Therefore, in our current research we work on overcoming this issue.

Acknowledgements: This research was supported by the Czech Science Foundation (GAČR) grant 19-00015S.

Bibliography

[1] J. D. Banfield and A. E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics* 49(3), 803–821, (1993).

[2] D. B. Dunson. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B*, **62**, 355 – 366, (2000).

[3] European Union Statistics on Income and Living Conditions. URL https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions

[4] A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* 1(2), 141–149, (1992).

[5] N. M. Laird and J. H. Ware. Random-Effects Models for Longitudinal Data. *Biometrics* 38(4), 963–974, (1982).

Author index

Šuvak, N., 101

Arab, Idir, 3
Aydoğdu, Halil, 76

Babilua, Petre, 8
Barbu, Vlad Stefan, 46
Beatriz, Santos, 3
Boivin, M. J., 101

Castro-Camilo, Daniela, 138
Catana, Luigi-Ionut, 13
Cournède, Paul-Henry, 107

Dvořák, Jiří, 56

Fraga Alves, Maria Isabel, 117
Frommlet, Florian, 35

Golovkine, Steven, 18
Gudan, Jovita, 24
Gurova, Silvi-Maria, 30

Hadjikyriakou, Milto, 3
Heuchenne, Cédric, 41
Hubin, Aliaksandr, 35

Jacquemain, Alexandre, 41
Jaroszewicz, Szymon, 96

Kalligeris, Emmanouil-Nektarios, 46
Karagrigoriou, Alex, 46
Karjalainen, Joona, 51
Kevei, Péter, 133
Klutchnikoff, Nicolas, 18
Koňasová, Kateřina, 56
Komárek, Arnošt, 148

Larsson, Johan, 61
Lemler, Sarah, 107
Leonenko, N. N., 101
Leordeanu, Marius, 81
Lydon, Myra, 128

Makrides, Andreas, 46
Marshall, Adele H., 128
Martino, Sara, 138

Naveiro, Roi, 66
Neves, Cláudia, 117

Oliveira, Paulo Eduardo, 3

Padyšák, Matúš, 71
Parpoula, Christina, 46
Patilea, Valentin, 18
Pekalp, Mustafa Hilmi, 76
Petrica, Marian, 81
Pircalabelu, Eugen, 41
Platonova, Mariia, 86
Popescu, Ionel, 81

Radojičić, Dragana, 91
Rudaś, Krzysztof, 96

Salinger, Zeljka, 101
Sautreuil, Mathilde, 107
Shi, Chengchun, 112
Sikorskii, A., 101
Silva Lomba, Jessica, 117
Srakar, Andrej, 123
Stevens, Nicola-Ann, 128
Stochitoiu, Radu D., 81
Storvik, Geir, 35
Szalai, Máté, 133

Taylor, Su E., 128

Vandeskog, Silius M., 138
Vanegas, Laura Jula, 143
Verbic, Miroslav, 123
Vávra, Jan, 148

Sponsors

Auspices